A Companion to Philosophical Logic

Dale Jacquette

BLACKWELL PUBLISHERS

A Companion to Philosophical Logic

Blackwell Companions to Philosophy

This outstanding student reference series offers a comprehensive and authoritative survey of philosophy as a whole. Written by today's leading philosophers, each volume provides lucid and engaging coverage of the key figures, terms, topics, and problems of the field. Taken together, the volumes provide the ideal basis for course use, representing an unparalleled work of reference for students and specialists alike.

Already published in the series

1. The Blackwell Companion to Philosophy
   Edited by Nicholas Bunnin and Eric Tsui-James
2. A Companion to Ethics
   Edited by Peter Singer
3. A Companion to Aesthetics
   Edited by David Cooper
4. A Companion to Epistemology
   Edited by Jonathan Dancy and Ernest Sosa
5. A Companion to Contemporary Political Philosophy
   Edited by Robert E. Goodin and Philip Pettit
6. A Companion to Philosophy of Mind
   Edited by Samuel Guttenplan
7. A Companion to Metaphysics
   Edited by Jaegwon Kim and Ernest Sosa
8. A Companion to Philosophy of Law and Legal Theory
   Edited by Dennis Patterson
9. A Companion to Philosophy of Religion
   Edited by Philip L. Quinn and Charles Taliaferro
10. A Companion to the Philosophy of Language
    Edited by Bob Hale and Crispin Wright
11. A Companion to World Philosophies
    Edited by Eliot Deutsch and Ron Bontekoe
12. A Companion to Continental Philosophy
    Edited by Simon Critchley and William Schroeder
13. A Companion to Feminist Philosophy
    Edited by Alison M. Jaggar and Iris Marion Young
14. A Companion to Cognitive Science
    Edited by William Bechtel and George Graham
15. A Companion to Bioethics
    Edited by Helga Kuhse and Peter Singer
16. A Companion to the Philosophers
    Edited by Robert L. Arrington
17. A Companion to Business Ethics
    Edited by Robert E. Frederick
18. A Companion to the Philosophy of Science
    Edited by W. H. Newton-Smith
19. A Companion to Environmental Philosophy
    Edited by Dale Jamieson
20. A Companion to Analytic Philosophy
    Edited by A. P. Martinich and David Sosa
21. A Companion to Genethics
    Edited by Justine Burley and John Harris
22. A Companion to Philosophical Logic
    Edited by Dale Jacquette

Forthcoming

A Companion to African American Philosophy
Edited by Tommy Lott and John Pittman

A Companion to African Philosophy
Edited by Kwasi Wiredu

A Companion to Ancient Philosophy
Edited by Mary Louise Gill

A Companion to Early Modern Philosophy
Edited by Steven Nadler

A Companion to Medieval Philosophy
Edited by Jorge J. E. Gracia, Greg Reichberg, and Timothy Noone

Blackwell
Companions to
Philosophy

A Companion to Philosophical Logic

Edited by

DALE JACQUETTE

Blackwell
Publishers

First published 2002

2 4 6 8 10 9 7 5 3 1

Blackwell Publishers Inc.
350 Main Street
Malden, Massachusetts 02148
USA

Blackwell Publishers Ltd
108 Cowley Road
Oxford OX4 1JF
UK

Library of Congress Cataloging-in-Publication Data

A companion to philosophical logic / edited by Dale Jacquette.
p. cm. (Blackwell companions to philosophy ; 22)
Includes bibliographical references and index.
ISBN 0-631-21671-5 (alk. paper)
1. Logic. I. Jacquette, Dale. II. Series.
BC71 .C65 2002
160—dc21
2001043236

British Library Cataloguing in Publication Data
A CIP catalogue record for this book is available from the British Library.

Typeset in 10 on 12½ pt Photina
by Best-set Typesetter Ltd., Hong Kong
Printed in Great Britain by T. J. International, Padstow, Cornwall

This book is printed on acid-free paper.

Contents

Contributors

Kent Bach, Professor of Philosophy, San Francisco State University, California.

José A. Benardete, Professor of Philosophy, University of Syracuse, New York.

Ermanno Bencivenga, Professor of Philosophy, University of California, Irvine, California.

Emma Borg, Lecturer in Philosophy, University of Reading, Pennsylvania.

E. P. Bos, University Lecturer in Ancient and Medieval Philosophy, Leiden University, The Netherlands.

Bryson Brown, Professor of Philosophy, University of Lethbridge, Lethbridge, Alberta, Canada.

James Cargile, Professor of Philosophy, University of Virginia, Virginia.

Robert Demolombe, Director of Research, ONERA, Department of Information Processing and Modeling, Toulouse Centre, Toulouse, France.

Maarten de Rijke, Professor, Institute for Logic, Language and Computation (ILLC), Department of Mathematics, Computer Science, Physics and Astronomy, University of Amsterdam, The Netherlands.

Branden Fitelson, Acting Assistant Professor in the Department of Philosophy, Stanford University, Stanford, California, and Scientific Associate, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois.

Melvin Fitting, Professor of Philosophy, Lehman College and the Graduate Center, City University of New York, New York.

Peter A. Flach, Reader in Machine Learning, Department of Computer Science, University of Bristol, UK.

Peter Forrest, Professor of Philosophy, School of Social Science, University of New England, Armidale, New South Wales, Australia.

D. M. Gabbay, FRSC, Augustus De Morgan Professor of Logic, Group of Logic and Computation, Department of Computer Science, King's College, London, UK.

Rolf George, Professor of Philosophy, University of Waterloo, Ontario, Canada.

Stephen Glaister, Lecturer, Department of Philosophy, University of Washington, Seattle, Washington.

Paul Gochet, Professor Emeritus, Department of Philosophy, and Ministère de l'Education et de la Recherche scientifique de la Comunauté Française, University of Liège, Belgium.

Richard Grandy, McManis Professor of Philosophy, Rice University, Houston, Texas.

Petr Hájek, Professor and Head of the Department of Theoretical Computer Science, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic.

Risto Hilpinen, Professor of Philosophy, University of Miami, Florida.

Herbert Hochberg, Professor of Philosophy, University of Texas, Austin, Texas.

Andrew W. Hodges, Wadham College, Oxford University, UK.

Dale Jacquette, Professor of Philosophy, Pennsylvania State University, University Park, Pennsylvania.

Andrew J. I. Jones, Professor, Department of Philosophy and Norwegian Research Centre for Computers and Law, University of Oslo, Norway.

Michael Jubien, Professor of Philosophy, University of California, Davis, California.

Matt Kaufmann, Senior Member of the Technical Staff, Advanced Micro Devices, Inc., Austin, Texas.

Gregory Landini, Associate Professor of Philosophy, University of Iowa, Ames, Iowa.

Ernest Lepore, Professor and Director, Center for Cognitive Science, Rutgers University, New Jersey.

Grzegorz Malinowski, Professor and Head of the Department of Logic, University of Łodz, Poland.

Edwin D. Mares, Senior Lecturer and Department Head, Department of Philosophy, Victoria University of Wellington, New Zealand.

J. Strother Moore, Admiral B. R. Inman Centennial Chair in Computing Theory, Department of Computer Sciences, University of Texas, Austin, Texas.

Gary Ostertag, Visiting Scholar, Department of Philosophy, New York University, New York City, New York.

Graham Priest, Boyce Gibson Professor of Philosophy, University of Melbourne, Australia, and Arché Professorial Fellow, Department of Logic and Metaphysics, University of St. Andrews, Scotland.

Greg Ray, Associate Professor of Philosophy, University of Florida, Gainesville, Florida.

Nicholas Rescher, University Professor of Philosophy, University of Pittsburgh, Pennsylvania.

Nathan Salmon, Professor of Philosophy, University of California, Santa Barbara, California.

Gerhard Schurz, Professor of Philosophy, Chair of Philosophy of Science, University of Erfurt, Germany, and Special Research Program, Institute for Philosophy, University of Salzburg, Austria.

Stewart Shapiro, Professor of Philosophy, Ohio State University at Newark, and Professorial Fellow, Department of Logic and Metaphysics, University of St. Andrews, Scotland.

Gila Sher, Professor of Philosophy, University of California, San Diego, California.

Keith Simmons, Professor of Philosophy, University of North Carolina at Chapel Hill, North Carolina.

Robin Smith, Professor of Philosophy, Texas A&M University, College Station, Texas.

Roy A. Sorensen, Professor of Philosophy, Dartmouth College, Hanover, New Hampshire.

B. G. Sundholm, Professor of Philosophy and History of Logic, Leiden University, The Netherlands.

Mary Tiles, Professor of Philosophy, University of Hawaii at Manoa, Hawaii.

Alasdair Urquhart, Professor of Philosophy, University of Toronto, Ontario, Canada.

Mark van Atten, Institute of Philosophy, Catholic University, Louvian, Belgium.

Johan van Benthem, Professor of Mathematical Logic and its Applications, University of Amsterdam, The Netherlands, and Bonsall Chair of Philosophy, Stanford University, California.

Dirk van Dalen, Professor, History of Logic and Philosophy of Mathematics, University of Utrecht, The Netherlands.

James Van Evra, Associate Professor of Philosophy, University of Waterloo, Ontario, Canada.

Heinrich Wansing, Professor of Logic and Philosophy of Science, Dresden University of Technology, Germany.

Scott Weinstein, Professor of Philosophy, University of Pennsylvania, Philadelphia, Pennsylvania.

Jan Woleński, Professor of Philosophy, Institute of Philosophy, Department of Epistemology, Jagiellonian University, Cracow, Poland.

Larry Wos, Senior Mathematician, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois.

Preface

The essays collected in this volume are previously unpublished contributions to philosophical logic from some of the most respected researchers in the field. In inviting these specialists to write on their specialities, I have sought to combine a representative breadth of coverage with an accessible depth of philosophical and mathematical sophistication that offers a clear picture of the historical development and current state of the art in philosophical logic. To whatever extent the book succeeds in meeting its objective, credit is due to the superb work of the logicians and philosophers who agreed to be part of this immoderate editorial undertaking.

My strategy has been to identify what I consider to be the most important topic areas in philosophical logic from the standpoint of students as well as professional scholars, and then in each case to recruit three or more of the best experts I could find who I thought were likely to disagree in interesting ways, encouraging each to address the questions they believe most important in their own way and in their own voice, without concern for what any of their co-contributors have to say. The result is a remarkable testimony to a thriving industry in contemporary philosophical logic, and, despite some detractors' premature eulogies of its imminent demise, the vitality of contemporary analytic philosophy.

With the exception of my introductory essay, the papers are clustered thematically, although the order is not always obvious. The first invisible division in the text proceeds from milestones in the history of logic to the relation of symbolic logic to ordinary language. Logical paradoxes and their philosophical implications are then introduced as essential for understanding Tarski's truth semantics and responses especially to the liar paradox which have been so fundamental in shaping the theory of meaning in modern philosophical logic. A discussion of selected paradoxes is accordingly followed by a choice of topics involving Tarski's concept of truth and Russell's theory of definite description in classical semantics that continue to play an essential role in current discussions in philosophical logic. The stage is thereby set for investigations of more recent trends in logic, emphasizing alternative concepts of logical consequence, and questions of existence presuppositions and ontology in logic. Metatheoretical considerations about the scope and limits of logic come next, advances that are naturally complemented by a suite of papers on the logical foundations of set theory and mathematics. Here another invisible threshold is attained, after which nonclassical logics begin to

appear, starting with modal logics in several categories, a larger section than most, because of the importance of modal logics in the development of set theoretical semantics and their many applications, followed by intuitionistic, free and many-valued logics, inductive, fuzzy and quantum logics, relevance, and paraconsistent logics. In the final grouping of papers, two sections complete the book's discussion of the implications for and practical applications of philosophical logic in machine theory and cognitive science, and the mechanization of logical inference and automated theorem and proof discovery.

Although some of the papers are more technical than others, all are intended for an introductory audience, and can be read with good understanding by beginning students in philosophy who have completed a first course in symbolic logic. This is especially true if the essays are read sequentially as they are presented within each section and from section to section. Inevitably, a full understanding of some topics treated at earlier stages of the book may require familiarity with principles and methods of logic that are considered in detail only in later sections, for which some looking ahead may occasionally be required. Additional background materials related to the study of philosophical logic can also be found in my simultaneously published Blackwell collections, *Philosophy of Logic: An Anthology* and *Philosophy of Mathematics: An Anthology*. The present volume will serve its purpose if it helps provide readers at all levels with a sufficient sense of interest in its subject to pursue advanced study of the concepts, methods, and problems of philosophical logic.

Acknowledgments

I wish to thank Steve Smith at Blackwell Publishers for inviting me to edit this collection of new essays on philosophical logic. I am grateful to the Alexander von Humboldt-Stiftung for supporting this project during my tenure as Research Fellow at the Bayerische-Julius-Maximilians-Universität-Würzburg in Germany. I thank the Franz Brentano Forschung and its Director, my friend Wilhelm Baumgartner, for graciously hosting my research visit during my sabbatical leave from the Pennsylvania State University in the academic term 2000–2001, when this edition was in preparation. I am also indebted to L. Jeffry Pelletier, for sage advice, and to Brian Armstrong for his professional scholarly assistance, particularly in compiling the list of Resources for Further Study. The book is dedicated to my wife, Tina, for all her tender mercies and moral inspiration.

Dale Jacquette

This page intentionally left blank

Introduction: Logic, Philosophy, and Philosophical Logic

DALE JACQUETTE

1 Philosophy as Logic

It has been many years since Bertrand Russell provocatively identified philosophy with logic. Although some logicians and philosophers continue to accept Russell's thesis, not least because it bears the stamp of Russell's authority in both fields, most commentators today prefer to describe the relationship between logic and philosophy as more complex. If logic remains important to philosophy, and philosophy to logic, it is undoubtedly because of what each can offer the other as an autonomous discipline.

Logic is no longer the monolithic edifice to which Russell could point in 1914, when in *Our Knowledge of the External World*, he made his famous observation that: "[E]very philosophical problem, when it is subjected to the necessary analysis and purification, is found either to be not really philosophical at all, or else to be, in the sense in which we are using the word, logical" (1914: 42). When contemporary philosophers speak of logic, they generally mean to refer to any of a variety of alternative formal symbolisms that can be used to formulate particular aspects of the formal inferential structures of language, including but not limited to languages in which philosophical ideas are conveyed. While logic is a useful tool in clarifying and perspicuously representing philosophical reasoning, many philosophers believe that there are areas, indeed, most parts, of legitimate philosophical inquiry, that have nothing directly to do with the specialized study of formal symbolic logic. Such a conclusion is especially plausible when philosophy is viewed broadly to include literary as well as scientific projects, particularly those that do not use or take any special notice of logic and mathematics, and that may even disclaim efforts to arrive at the truth about any philosophical subject, as in certain outgrowths of postmodern philosophy. Russell also feels the need to qualify the identification of philosophy with logic, adding immediately after his statement quoted above: "But as the word 'logic' is never used in the same sense by two different philosophers, some explanation of what I mean by the word is indispensable at the outset" (1914: 42).

The fact, as Russell observes, that philosophers have many different ideas of logic constitutes one of the most fundamental problems for philosophical logic and the philosophy of logic. To define the concept of logic, to understand the diverse kinds of systems that have been considered logics, and to arrive at a satisfactory definition of

the concept of logic that applies alike to Aristotelian syllogisms, Boolean algebras, Frege's *Begriffsschrift*, Whitehead and Russell's *Principia Mathematica*, and unlimitedly many nonstandard formal systems, and informal logic in several traditions, grading off into rhetoric, argumentation theory, and discourse analysis, is a formidable task. What makes all of these projects logical, a part or different forms of logic, or distinct logics? A working definition that may be correct if somewhat uninformative as far as it goes is to say that logic in any of its manifestations is the systematic study of principles of correct reasoning. The principles of logic can then be explored formally or informally, and by any of a number of different styles of exposition, some of which may be highly specialized in dealing with very particular areas of reasoning.

Logic is both a symbolism for the expression of the formal structures of thought and an inference mechanism for calculating and drawing conclusions from assumptions in reasoning. The dual nature of logic has figured prominently in the range of issues that have come to be associated with the problems of philosophical logic.

2 Logic and Philosophy of Language

A primary source of problems in philosophical logic is the analysis of language. Philosophers are interested in language and semantics or theory of meaning for a number of reasons. The problems and methods of applied logic in studying the philosophy of language are directly associated with the traditional domain of philosophical logic.

Language facility distinguishes human beings from other animals we know of, higher primates who have been taught by humans to make limited use of sign-language and computer push-button languages notwithstanding. Philosophers interested in human nature and what makes our species unique in the animal kingdom as a result are attracted to problems of understanding language as a way of gaining insight into the human condition. The complexity of language and the difficulty of formulating an adequate theory of meaning for ordinary and scientific language by itself is a sufficient invitation for many philosophers to answer the challenge of articulating a philosophical semantics. More importantly, logicians and philosophers in the analytic tradition have considered unclarities in the expression of philosophical ideas to be the foundation of philosophical puzzles and paradoxes, and have accordingly sought to solve, avoid, or at least gain a better perspective on the problems by way of the theory of meaning.

This is undoubtedly part of what Russell means in pronouncing all of philosophy properly so-called identical with logic. Symbolic logic has been the tool of choice for philosophers investigating the properties of language in philosophical logic, because it is itself a language whose syntax and semantics are at the disposal and under the control of the logician where they can be better studied in more ideal abstract terms. A formal system of logic considered as a language has definite advantages over colloquial discourse as a model of how language works, where its factors are more readily discerned and rigorously formulated independently of the ambiguities and etymological confusions that are endemic to natural language, which, as Ludwig Wittgenstein aptly remarks in the *Tractatus Logico-Philosophicus* (1922: 4.002), "is a part of the

human organism and is not less complicated than it." Even for philosophical logicians who do not seek to replace ordinary language with an ideal language like Frege's *Begriffsschrift* or Whitehead and Russell's *Principia Mathematica*, but, like Wittgenstein, hope to understand how language generally is capable of expressing meaning, the use of symbolic logic has remained an indispensable instrument in philosophy of language. The fact that logic lends itself to more sharply and univocally defined distinctions makes it convenient for the analysis of concepts in philosophy, including the semantic principles by which logical formulas are themselves precisely interpreted. The usefulness of logic in philosophical applications has played a major role in the development of symbolic logic, which in turn has opened up new possibilities for logic's use in refinements of philosophical techniques.

How, then, has the partnership between philosophical logic and philosophy of language taken shape? In too many ways for the story to be told in a summary that does not distort the true riches of ingenuity, invention, and discovery on the part of philosophers and logicians in the annals of recent and contemporary analytic philosophy. Nevertheless, something of the flavor of work in this exciting field can be conveyed from a brief discussion of a few well-chosen examples. We turn next to consider some instructive concrete possibilities.

3 Modes and Methods of Philosophical Logic

Logic is formal, and by itself has no content. It applies at most only indirectly to the world, as the formal theory of thoughts about and descriptions of the world. Logic can be used in many ways to state, clarify, and express ideas, and to authorize the derivation of consequences, when its formulas are assigned substantive content in application. Although logic in its pure form is unfreighted with philosophical truths, it can contribute in definite ways to the clarification and solution of philosophical problems.

Philosophical logic often combines an application of logical symbolisms with a commitment to specific philosophical ideas. Symbolic logic, even in its purest form, is also not entirely free of philosophical ideology, although some logicians have made it their mission to try to make logic as neutral a vehicle as possible for the unbiased expression of the logical form of philosophical disagreements on every conceivable topic, including those most closely related to the conceptual presuppositions of classical logic. To the extent that substantive philosophical positions are built into the interpretation of symbolic logic, the use of logic in addressing philosophical problems may seem highly effective and convincing. In that case, of course, it is not logic alone that is doing the work, but whatever philosophical theses have been packed into its symbolism.

There is often a temptation to use philosophical logic in this way. A logical notation is loaded with philosophical cargo to enable it to appear at least to make progress against outstanding philosophical problems. Logic as a branch of mathematics deservedly carries a certain authority in intellectual disputes. We should recognize, however, that when a logical formalism appears to solve a philosophical problem, it seldom does so by itself, but only by virtue of the philosophical ideas it is used to express. That being the case, we need to question whether the philosophy shouldered by philo-

sophical logic is sound or faulty, just as we would need to do if we had set about considering the philosophical issues directly without the intervention of a symbolic logical notation. If logic helps the cause of clarifying and solving or avoiding philosophical problems, it does so thanks largely to the ability of its formal structures to sort out and more clearly represent a choice of philosophical ideas, and not by means of substantive philosophical assumptions hidden in the background of a particular logical system.

In his "Introduction" to Wittgenstein's *Tractatus*, Russell recognizes the potential of a logical symbolism to clarify philosophical concepts. He states: "a good notation has a subtlety and suggestiveness which at times make it seem almost like a live teacher. Notational irregularities are often the first sign of philosophical errors" (1922: 17–18). The value of an adequate logical notation is that it provides information about the logical form of the ideas it expresses. It can call attention to logical structures that might otherwise be overlooked in informal expression, including tipoffs about conceptual inconsistencies. This, after all, is a primary pragmatic justification for the use of symbolic logic. It teaches us things that we could not (or not as easily) learn without its formalisms. Such discoveries are often made as logicians explore the scope and expressive flexibility of a formal system. They emerge in the study of a formalism's mathematical multiplicity, in Wittgenstein's terminology, its shared isomorphism or lack thereof with the features of thought or discourse it is supposed to formalize, together with its internal logical interrelations and deductive consequences.

Russell, in his own celebrated application of philosophical logic in the analysis of definite descriptions, in his essay "On Denoting" (*Mind* 1905), seems nevertheless to have decanted a significant amount of philosophy into a logical vessel in order to gain philosophical mileage from what appears to be purely logical distinctions. Russell's theory of descriptions has been enormously influential in the rise of analytic philosophy, to such a degree that F. P. Ramsey in his essay "Philosophy" was moved to eulogize it as "that paradigm of philosophy." The theory has indeed been a model for some of the best work in philosophical logic for over a century. It is worthwhile, therefore, to consider the theory in detail, to understand how it combines philosophy with logic, and the amount of labor borne by logic as opposed to the prior philosophical commitments deeply integrated into Russell's logic.

4 Logic as Philosophy in Philosophical Logic

We can identify at least three characteristics of Russell's theory that provide enduring guidelines for philosophical logic. Russell's breakdown of definite descriptions into an existence clause, uniqueness clause, and predication of a property to a uniquely denoted entity, using the devices of symbolic logic to conjoin these three formalized conditions, demonstrate the power of symbolic logic to present the analysis of a complex concept into more basic components for philosophical purposes. Russell's method has very properly been compared to that of an optical prism that takes a single beam of white light and breaks it up into its constituent spectrum of colors. The colors are not added or produced by the prism, but are there all along, inherent in the white light, although it takes a special instrument to reveal their presence. The same is true of definite descriptions, to which Russell applies symbolic logic in order to break apart

and discover by reflection the three conditions concealed within the apparently simple word 'the.'

This observation leads to the second noteworthy feature of Russell's analysis. Russell makes an inestimable contribution to the flowering of analytic philosophy by suggesting that the logical form of a proposition, as judged in terms of its superficial grammatical structure, is not necessarily its real, underlying form, appreciated by means of logical analysis. I cannot put the point better than Wittgenstein in *Tractatus* (1922: 4.0031), when he declares: "Russell's merit is to have shown that the apparent logical form of the proposition need not be its real form." Wittgenstein no doubt puts his finger on a major ingredient in the appeal of Russell's theory of descriptions. By suggesting that philosophical logic has as part of its project to uncover the real underlying or ulterior logical form of sentences in ordinary thought and language, Russell inspired generations of philosophers with a vision of logical analysis excavating the subterranean logical structures beneath the surface of colloquial discourse.

Third, Russell's theory is rightly dignified as a wellspring of contemporary analytic philosophy because of its dramatic use of logical methods in disambiguating philosophically equivocal linguistic expressions. Russell considers among others the problem of interpreting the sentence, 'The present king of France is not bald.' The dilemma he intuits is that if the sentence is taken to mean that there is a present king of France who is not bald, then the sentence should be false. To declare the sentence false, at least when we are operating within the parameters of ordinary language, wrongly seems to entail that there is a present hirsute king of France. Russell's genius in the theory of definite descriptions is partly seen in his recognition that symbolic logic permits the exact disambiguation of the scope of the negation operator that is blurred in everyday speech. He accordingly distinguishes between saying 'There exists one and only one present king of France and it is not the case that he is bald,' versus 'It is not the case that there exists one and only one present king of France and he is bald (or, it is not the case that he is bald).' The first sentence is false, but its proper negation is the second sentence, which does not commit the speaker to the existence of a hirsute present king of France.

Although the distinction can also be indicated as here in a modified form of ordinary English, Russell finds that it is only in symbolic logic that the full force of placing the negation sign externally, with the entire proposition in its scope, as opposed to internally, governing only the predication of the property of being bald in the third clause of the formal analysis of the definite description, can be fully and unequivocally appreciated. In standard logical notation, the difference is formalized as that between $\sim(\exists x)(Kxf \,\&\, (\forall y)((Kyf \equiv x = y) \,\&\, Bx))$ as opposed to $(\exists x)(Kxf \,\&\, (\forall y)((Kyf \equiv x = y) \,\&\, \sim Bx))$. The difference in the scope of the negation, and the difference it makes in the truth values of the two propositions, is so immediately apparent as to powerfully iconically recommend the use of symbolic logic as a general method of clarifying logical obscurities and circumventing conceptual confusions.

Having acknowledged the strength of Russell's analytic paradigm, it may also be worthwhile to consider its underlying philosophical assumptions. Russell is interested not only in the truth value of sentences ostensibly designating nonexistent objects like the present king of France, but also in understanding predications of properties to fictional creatures, like Pegasus, the flying horse of ancient Greek mythology. Russell

regards proper names like 'Pegasus' as disguised definite descriptions, which he interprets according to his three-part analysis as consisting of an existence claim, a uniqueness claim, and the predication of a property to the uniquely designated entity. If I say, then, that 'Pegasus is winged,' Russell interprets this sentence as falsely asserting that there exists a flying horse, there is only one flying horse, and it is winged. From this it appears to follow that something of metaphysical significance has been derived from Russell's skillful use of philosophical logic; namely, that it is false to say of any nonexistent object like Pegasus that the object has any of the properties attributed to it in myths, legends, or storytelling contexts.

If we look at the logical symbolism Russell employs, we see that in this case it reads: $(\exists x)(Fx \,\&\, (\forall y)(Fy \equiv x = y) \,\&\, Wx)$. The formula, it must be said, is supposed to be judged false only because the quantifier in $(\exists x)(Fx \ldots)$ is interpreted as meaning that there actually exists such an object in the logic's semantic domain that truly possesses the property F, of being a flying horse. Russell as a matter of fact has no way to construe an object like Pegasus in his logic other than as the value of an existentially loaded quantifier-bound variable. This is probably not the place to dispute with Russell about whether such a logical treatment of names like 'Pegasus' is philosophically justified or not. It is nevertheless important to recognize that Russell's evaluation of such sentences as false is predetermined by his existence presuppositional semantics for the 'existential' quantifier, and by the fact that his logic permits no alternative means of considering the semantic status of sentences ostensibly containing proper names for nonexistent objects. This makes it an altogether philosophically foregone conclusion that sentences like 'Pegasus is winged,' which many logicians would otherwise consider to be true propositions of mythology, are false. The point is that Russell is able to produce this philosophical result from his logical analysis of the meaning of the sentence only because the position is already loaded into the presuppositions of the syntax and semantics of his interpretation of formal symbolic logic. The interesting philosophical question that Russell would be hard-pressed to answer satisfactorily is whether his logic is philosophically adequate to the proper analysis of problematic sentences in this category. It is not a conclusion of logic alone that Russell advocates, whether correct or incorrect, but of an applied philosophical logic that is heavily but not inevitably imbued with a prior metaphysical commitment to an existence-presuppositional extensional syntax and semantics.

A good logical notation, as Russell says, can function philosophically much like a living teacher. As a pure formalism, however, logic is not an autonomous authority on any matter of philosophical truth. It has, in itself, no philosophical implications, and in its applications in philosophical logic, as Russell's example illustrates, it is capable of supporting only those philosophical conclusions with which it is deliberately or inadvertently invested by logicians. This, then, is another sense in which Russell in his most important contributions to philosophical logic identifies logic with philosophy.

5 On Philosophical Presuppositions and Copia of Logical Systems

The perspective we have arrived at in understanding the relation between logic and philosophy can help to answer a difficult question about the nature of logic and the status

of multiple logical systems. Why are there so many different systems of logic? Is there just one underlying logic, of which all the various systems are alternative partial expressions? Or are there many different logics that are related to one another by a network of partially overlapping family resemblances?

If we consider work in contemporary theoretical logic at face value, there seem to be indefinitely many logics. Alethic modal logics are concerned with matters of necessity and possibility; doxastic logics are designed to explain the logical structures of belief states; epistemic logics are offered to formalize valid inferences about knowledge. There are specialized logics of quantum physical phenomena, deontic logics of obligation and permission, and many others. An important source of the proliferation of logical systems in contemporary logic and philosophy is in philosophical issues arising from dissatisfaction with classical logics in dealing with specific aspects of scientific and everyday reasoning. This is the basis for work in many-valued logics, free logic, relevance, and paraconsistent logics, and logics of beingless intended objects, that do not limit logical inference to existent entities in referring to and truly predicating properties of objects, and for the paraconsistent stance that logical inconsistencies need not explosively entail any and every proposition, but that contradictions can be tolerated without trivializing all inferences.

Applications of logic to philosophical problems of these kinds are a continuing basis for innovations in formal symbolic logic and the development of new nonstandard systems of logic. Logic is also concerned with abstract theoretical matters concerning its own formal symbolisms and the properties, such as the scope and limits of logical and mathematical systems considered as a whole, in the study of logical metatheory. The advance of logic has been nourished by its theoretical and practical applications in set theory, computer engineering, artificial intelligence modeling, formal semantics and linguistic analysis of scientific theory, philosophical argument, and colloquial language. There is valuable feedback between logical theory and practice, much as there is in pure and applied mathematics. The need for new formalisms is sometimes made urgent by the limitations of received systems that are only discovered when we try to apply them to real problems. At the same time, developments in symbolic logic that are undertaken purely for the sake of their theoretical interest frequently suggest new applications of logical analysis for which no need had previously been perceived.

The number of distinct logical systems inevitably raises the philosophical question of how the multiplicity of logics should be understood. Some logicians are partisan defenders of particular logical formalisms as the ideal single correct logic. Others are tolerant of many logics, adopting an attitude according to which particular formal systems may be appropriate for particular analytic tasks, but that no single logic or cluster or family of logics deserves to be called the one and only correct system of logic. Those who favor a single correct system of logic must either regard alternative logics as incorrect, however formally interesting, or else interpret them as representing conflicting incompatible opinions about the best and uniquely correct logical system. Such a contrast of philosophical positions about the nature of logic and the uniquely correct logic or plurality of alternative logics has positive analogies in the opposition of moral absolutism and moral relativism, and in questions of privileged objective truth versus subjectivism, perspectivalism, and syncretism in the theory of knowledge. It would not be surprising to find philosophers who incline toward relativism in ethics or epistemol-

ogy also to prefer a tolerant attitude about the peaceful coexistence of many different logical systems, and for their adversaries who think in terms of moral and epistemic absolutes to embrace a single correct logic that either defeats the ostensible alternatives, or resolves apparent conflicts between many if not all of them in a greater overarching synthesis.

Philosophy thrives on just such tensions and ambiguities, and philosophical logic is no exception. All of the diverse formal syntactical distinctions available in contemporary symbolic logic can be put to good use in clarifying philosophical ideas and drawing more precisely interpreted distinctions than are otherwise possible in ordinary language, or even in specialized but nonsymbolic philosophical terminologies. The methods of set theory, model set theoretical semantics, and axiomatizations of many types of philosophical concepts are among the widely used formalisms in present-day philosophical logic. The future will likely see more sophisticated logical machinery, and with it an even greater upsurge in the number and variety of logical systems and distinctive categories of logic and philosophical logics. If there is a logic of knowledge and a logic of moral obligation, then there can surely be multiple logics of deductively valid inference, each tailored to a particular philosophical conception of how even the most basic logical operations may be thought to function. We can nonetheless continue to expect that partisan champions in philosophical logic will want to refer to a preferred formalism as logic full stop, or as *the* one and only correct or underlying primary or essential logic. The awareness of philosophical commitment and presupposition even in the most rigorous abstract logical symbolisms, and of philosophical logic as an application of logic in which philosophical ideas are already deeply infused, can help to make logic a more powerful ally of philosophical analysis.

References

Frege, G. (1879) *Begriffsschrift*. Halle: Louis Nebert.

Ramsey, F. P. (1978) Philosophy. In D. H. Mellor (ed.), *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*. Atlantic Highlands, NJ: Humanities Press.

Russell, B. (1905) On denoting. *Mind*, 104, 827–62.

Russell, B. (1914) *Our Knowledge of the External World*. London: Allen & Unwin.

Whitehead, A. N. and Russell, B. (1927) *Principia Mathematica*. Cambridge: Cambridge University Press.

Wittgenstein, L. (1922) *Tractatus Logica-Philosophicus*, ed. C. K. Ogden. London: Routledge & Kegan Paul.

Part I

HISTORICAL DEVELOPMENT OF LOGIC

This page intentionally left blank

1

Ancient Greek Philosophical Logic

ROBIN SMITH

Ancient Greek logic was inseparable from ancient Greek philosophy. The formal theories developed by major logicians such as Aristotle, Diodorus Cronus, and Chrysippus were in large part influenced by metaphysical and epistemological concerns. In this brief essay, I will try to give some picture of this interrelationship. For reasons of space, I make no attempt to cover, or even to mention, every aspect of ancient Greek logic. I have preferred instead to concentrate on illustrating its philosophical aspects.

1 The Origins: Parmenides and Zeno

Greek philosophical logic originates with Parmenides (c. 510–c. 440 BCE). Though Parmenides cannot be said to have had a logic, or even an interest in studying the validity of arguments, his views did much to set the agenda out of which many things in Greek philosophy, including logic, later arose. His philosophical position is both simple and mystifying: being is, whereas not being is not and cannot either be thought or said. Consequently, any type of expression that implies that being is not or that not being is must be dismissed as nonsense. For Parmenides, this includes any reference to change (since it must involve the coming to be of what is not and the not being of what is) or multiplicity (since to say that there are two things is to say that something is not something else). The conclusion is that what is is one, unchanging, and uniform, without distinctions. Much of subsequent Greek philosophy is an effort to avoid these consequences and defend the coherence of talk of motion and multiplicity.

A second, and more explicitly logical, impact of Parmenides' thought on Greek philosophy is through its defense by Parmenides' follower Zeno of Elea (c. 490–c. 430 BCE). According to Plato's *Parmenides*, Zeno's goal was to defend Parmenides' views from the objection that they were absurd or in contradiction to our ordinary beliefs. In response, Zeno argued that the beliefs that there is motion and that there is a multiplicity of entities have consequences that are even more absurd because self-contradictory. This was the point of his celebrated arguments against motion and multiplicity.

To consider one example, Zeno gives the following argument (paraphrased) that motion is impossible:

> In order to move from point A to point B, you must first reach the point halfway between them. But before you can reach that point, you must reach the point halfway to it. Continuing in this way, we see that before you can reach any point, you must already have reached an infinity of points, which is impossible. Therefore, motion is impossible.

This argument rests only on the assumptions that motion is possible, that in order to move from one point to another one must first pass through the point halfway between, and that there is a point halfway between any two points.

Zeno's arguments take a particular form: beginning with premises accepted by his opponent, they derive conclusions that the opponent must recognize as impossible. Aristotle says that in introducing this form of argument, Zeno was the originator of 'dialectic'. The meaning of this word is contested by scholars, but we may note three features of Zeno's argument: (1) it is directed at someone else; (2) it takes its start from premises accepted by that other party; (3) its goal is the refutation of a view of that other party. These three characteristics can serve as a rough definition of a dialectical argument.

2 Dialectic and the Beginnings of Logical Theory

In the later fifth century BCE, professional teachers of oratory appeared in Athens. These were most often the same people called (by us, by their contemporaries, and often by themselves) 'Sophists'. We know that a number of the Sophists had interesting (and quite divergent) views on philosophical matters. Teaching oratory was a profitable occupation, and several Sophists seem to have amassed fortunes from it. The content of their instruction, to judge by later treatises on rhetoric, would have included such things as style and diction, but it would also have included some training in argumentation. That could have ranged from teaching set pieces of argument useful for specific situations, all the way to teaching some kind of method for devising arguments according to principles. One theme that emerges in several sophistic thinkers is a kind of relativism about truth. This is most forcefully put by Protagoras (c. 485–415 BCE), who began his treatise entitled *Truth* with the line, "Man is the measure of all things; of things that are, that they are, and of things that are not, that they are not." Plato tells us in his *Theaetetus* that this meant "whatever seems to be true to anyone is true to that person": he denied that there is any truth apart from the opinions of individuals. For Protagoras, this appears to have been connected with a thesis about the functioning of argument in a political situation. Whoever has the most skill at argument can make it seem (and thus be) to others however he wishes: in Protagoras' world, persuasive speech creates not merely belief but also truth.

Even apart from this perhaps extreme view, we find the themes of the variability of human opinion and the power of argument widespread in fifth-century Athens. Herodotus' history of the Persian Wars present a picture of opinions about right and wrong as merely matters of custom by displaying the variability in customs from one people to another. The treatise known as the *Twofold Arguments* (*Dissoi Logoi*) gives a series of arguments for and against each of a group of propositions; the implication is that argument can equally well support any view and its contradictory.

Contemporary with the Sophists was Socrates (469–399 BCE), whose fellow Athenians probably regarded him as another Sophist. Socrates did not teach oratory (nor indeed does he appear to have taught anything for a fee). Instead, he engaged people he encountered in a distinctive type of argument: beginning by asking them questions about matters they claimed to have knowledge of, he would lead them, on the basis of their own answers to further questions, to conclusions they found absurd or to contradictions of their earlier admissions. This process, which Plato and Aristotle both saw as a form of dialectical argument, usually goes by the name of 'Socratic refutation.' In overall form, it exactly resembles Zeno's arguments in support of Parmenides. Socrates insisted that he knew nothing himself and that his refutations were merely a tool for detecting ignorance in others.

Plato (428/7–348/7 BCE) did not develop a logical theory in any significant sense. However, he did try to respond to some of the issues raised by Parmenides, Protagoras, and others. In his *Theaetetus*, he argues that Protagoras' relativistic conception of truth is self-refuting in the sense that if Protagoras intends it to apply universally, then it must apply to opinions about Protagoras' theory of truth itself; moreover, it implies that the same opinions are both true and false simultaneously. He also partially rejects Parmenides' thesis that only what is can be thought or said by distinguishing a realm of 'becoming' that is not simply non-being but also cannot be said simply to be without qualification.

Plato's most celebrated philosophical doctrine, his theory of Forms or Ideas, can be seen as a theory of predication, that is, a theory of what it is for a thing to have a property or attribute. In very crude outline, Plato's response is that what it is for x (e.g. Socrates) to be F (e.g. tall) is for x to stand in a certain relation (usually called 'participation') to an entity, 'the tall itself,' which just is tall. In his *Sophist*, Plato begins to develop a semantic theory for predications. He observes that truth and falsehood are not properties of names standing alone but only of sentences produced by combining words. 'Theaetetus' and 'is sitting' are, in isolation, meaningful in some way but neither true nor false. We find truth or falsehood only in their combination: 'Theaetetus is sitting.' For Plato, a major achievement of this analysis is that it allows him to understand falsehoods as meaningful. In the sentence 'Theaetetus is flying,' both 'Theaetetus' and 'is flying' are meaningful; their combination is false, but it is still meaningful.

Aristotle (384–322 BCE), Plato's student, developed the first logical theory of which we know. He follows Plato in analyzing simple sentences into noun and verb, or subject and predicate, but he develops it in far greater detail and extends it to sentences which have general or universal (*katholou*, 'of a whole': the term seems to originate with Aristotle) subjects and predicates.

Aristotle also gives an answer to Protagoras and to related positions. Specifically, in Book IV of his *Metaphysics*, he argues that there is a proposition which is in a way prior to every other truth: it is prior because it is a proposition which anyone who knows anything must accept and because it is impossible actually to disbelieve it. The proposition in question is what we usually call the principle of non-contradiction: "it is impossible for the same thing to be both affirmed and denied of the same thing at the same time and in the same way" (*Met.* IV.3, 1005b19–20). He argues that it follows from this principle itself that no one can disbelieve it. At the same time, since it is prior to every other truth, it cannot itself be proved. However, Aristotle holds that anyone who claims

to deny it (or indeed claims anything at all) already presupposes it, and he undertakes to show this through what he calls a "refutative demonstration" (*Met.* IV.4).

3 Aristotle and the Theory of Demonstration

When Aristotle says that the principle of non-contradiction cannot be proved because there is nothing prior from which it could be proved, he appeals to a more general thesis concerning demonstration or proof: no system of demonstrations can prove its own first principles. His argument for this appears in his *Posterior Analytics*, a work best regarded as the oldest extant treatise on the nature of mathematical proof. The subject of the *Posterior Analytics* is *demonstrative sciences*: a demonstrative science is a body of knowledge organized into demonstrations (proofs), which in turn are deductive arguments from premises already established. If a truth is demonstrable, then for Aristotle to know it just is to possess its demonstration: proofs are neither a means of finding out new truths nor an expository or pedagogical device for presenting results, but rather are constitutive of knowledge. Though he does not limit demonstrative sciences to mathematics, it is clear that he regards arithmetic and geometry as the clearest examples of them. Both historical and terminological affinities with Greek mathematics confirm this close association.

A demonstration, for Aristotle, is a deduction that shows why something is necessarily so. This at once imposes two critical limits on demonstrations: nothing can be demonstrated except what is necessarily so, and nothing can be demonstrated except that which has a cause or explanation (the force of the latter restriction will be evident shortly).

Since demonstrations are valid arguments, whatever holds of valid arguments in general will hold of them. Therefore, a natural place to begin the discussion of demonstrations would be with a general account of validity. Aristotle announces exactly that intention at the beginning of his *Prior Analytics*, the principal subject of which is the 'syllogism', a term defined by Aristotle as "an argument in which, some things being supposed, something else follows of necessity because of the things supposed." This is obviously a general definition of 'valid argument.' However, Aristotle thought that all valid arguments could be 'reduced' to a relatively limited set of valid forms which he usually refers to as 'arguments in the figures' (modern terminology refers to these forms as 'syllogisms'; this can lead to confusion in discussing Aristotle's theory).

Aristotle maintained that a single proposition was always either the *affirmation* or the *denial* of a single predicate of a single subject: 'Socrates is sitting' affirms 'sitting' of Socrates, 'Plato is not flying' denies 'flying' of Plato. In addition to simple predications such as those illustrated here, with individuals as subjects, he also regarded sentences with general subjects as predications: 'All Greeks are humans,' 'Dogs are mammals,' 'Cats are not bipeds.' (Here he parts company from modern logic, which since Frege has seen such sentences as having a radically different structure from predications.) Aristotle's logical theory is in effect the theory of general predications. In addition to the distinction between affirmation and denial, general predications can also be divided according as the predicate is affirmed or denied of all (universal) or only part (particular) of its subject. There are then four types of general predications:

	Affirmed (affirmative)	*Denied (negative)*
Universal	'Every human is mortal'	'No human is mortal'
Particular	'Some human is mortal'	'Not every human is mortal'

Aristotle then explores which combinations of two premises that share a term will imply a third sentence having the two non-shared terms as its subject and predicate. He distinguishes three possibilities based on the role of the shared term (the 'middle,' in his terminology) in the premises: it can be predicate of one and subject of the other (he calls this the 'first figure'), predicate of both ('second figure'), or subject of both ('third figure'). He carries out his investigation by first taking four combinations in the first figure as basic. He then systematically examines all other combinations in all the figures, doing one of two things for each of them: (1) in some cases, he shows that a conclusion follows by deducing that conclusion from the premises, using as resources one of the four basic forms and a limited stock of rules of inference; (2) in other cases, he shows that no conclusion follows by giving a set of counterexamples to any possible form of conclusion. As a result, he not only has an enumeration of all the valid forms of 'argument in the figures,' he also has shown that all of them can be 'reduced' to the basic four forms. He even shows that two of the basic forms can be derived from the other two using somewhat longer deductions. Following this treatment, he argues that every valid argument whatsoever can be 'reduced' to the valid forms of argument 'in the figures.' His defense of this is necessarily more complex, since it includes analysis of a variety of forms of arguments, for each of which he proposes ways to extract a figured argument.

I will not pursue here the details of his theory (see Corcoran 1973; Łukasiewicz 1957; Smiley 1974; Smith 1989). My concern instead is with the character of the whole enterprise. Aristotle's overriding concern is with demonstrating that every valid argument whatsoever can be reduced to a very small number of valid forms. This is not the sort of result that an author of a handbook for testing arguments for validity would want. It is, however, precisely the kind of result that someone interested in studying the structures of proofs would find valuable. And that is precisely the use we find Aristotle making of it. The only work of his that makes substantive use of the results proved in the *Prior Analytics* is the *Posterior Analytics*. Aristotle uses those results as the basis for a crucial argument to establish his position on the structures of demonstrative sciences. On this basis, I am persuaded that the theory contained in the *Prior Analytics* was developed largely to serve the needs of Aristotle's theory of demonstration, especially this argument: here, as in much of the early history of modern symbolic logic, logical theory arose to meet the needs of the philosophy of mathematics.

4 The Regress Argument of *Posterior Analytics* I.3

The argument to which I am referring is Aristotle's response to a problem about the possibility of demonstration that he presents in *Posterior Analytics* I.3: if demonstrations must rest on premises already demonstrated, then how is demonstration possible at all? Here is Aristotle's presentation of the positions in the debate:

> Some think that, because of the need to know the first things scientifically, there is no scientific knowledge. Others think that there is and that there is a demonstration of them all. Neither of these views is either true or necessary. Now, as for those who suppose that there is no scientific knowledge at all, they claim that it can be led into infinity, so that we do not know the posterior things from prior things of which none are first (and they are right, for it is impossible to go through infinite things). And if they do come to a stop and there are starting points, these will not be known just because there is no demonstration of them (which alone they say is scientific knowledge). And if it is not possible to know the first things, then neither is it possible to know those which follow from them scientifically, in the absolute or correct sense, but only from the assumption 'if these are so.' The other group agrees about scientific knowledge (that is, that it comes only through demonstration) but think that nothing prevents there being demonstration of everything because demonstration can be in a circle, that is, reciprocal. (*Posterior Analytics* I.3, 72b5–18)

Though this regress argument is frequently used as an early example of the kind of skeptical problem central to modern epistemology, a careful study of Aristotle's response to it shows that he has rather different concerns. He is really setting the stage for a complex and sophisticated argument about the structures of systems of mathematical proofs.

Even before Aristotle arrived in Athens, Plato's Academy was becoming a focal point for new developments in mathematics. In addition to proving new results and searching for the solutions to outstanding puzzles, Greek mathematicians had begun to arrange their accumulated knowledge systematically as a single structure of proofs. The ultimate outcome of this process, a century after Aristotle, was Euclid's *Elements*. However, though we do not know its contents, Hippocrates of Chios (fl. 440 BCE) composed an *Elements* in the late fifth or early fourth century, and Theudius of Magnesia (fl. c. 350? BCE) put together a treatise during Aristotle's lifetime that incorporated work by a number of other prominent mathematicians, including Archytas (428–347 BCE), Eudoxus (400–347 BCE), Leodamas (fl. c. 380 BCE), Theaetetus (c. 415–c. 369 BCE), and Menaechmus (c. 350? BCE). Euclid's *Elements* (c. 295 BCE) presupposes a certain overall structure for a mathematical system. At its basis are propositions which are not proved in the system; some of these are definitions, some are 'common conceptions' (*koinai ennoiai*), and some are 'things asked for' (*aitemata*: the customary translation is 'postulates'). Further propositions are added to the system by logical deduction from these first propositions and any others already proved; these are called theorems. Now, it is precisely this picture of a demonstrative system that is at issue in the passage quoted above from *Posterior Analytics* I.3, and one of the main goals of the treatise is to argue for it. Specifically, Aristotle argues that any demonstrative system must contain first propositions which are not demonstrated, or even demonstrable, in that system.

Aristotle's response to the regress argument appears at first to be a mere assertion: there are first principles that can be known without being demonstrated. We should then expect him to tell us straightaway what this other means of knowledge of these first principles is. Instead, he expends a great deal of argument trying to prove that the regress of premises always 'comes to a stop,' and it is in this argument that he needs the results established in the *Prior Analytics*. In order to appreciate the significance of this, we need to take note of an important difference between Aristotle's logical system

and modern predicate (and propositional) logic. In Aristotle's logic, it is possible for there to be true propositions which cannot be deduced from any other set of true propositions whatsoever that does not already contain them. Aristotle's logic contains only predications, and the only rules of inference it knows about are those of the arguments in the figures. Now, a true sentence 'A belongs to every B' can only be deduced from premises of exactly one type: two premises of the forms 'A belongs to every C' and 'C belongs to every B.' If there are no such true premises, then 'A belongs to every B,' though true, is absolutely undeducible, and thus indemonstrable in a purely logical or semantic sense. Similar results hold for the other forms of sentence, though they are more complicated because there are multiple ways of deducing each of them.

Aristotle calls such true but undeducible sentences 'unmiddled' (*amesos*: the standard translation 'immediate,' though etymologically correct, is highly misleading). Since an unmiddled proposition cannot be deduced from anything, it obviously cannot be the object of a demonstration. Moreover, any premise regress that encounters such a proposition will come to a stop at that point. If every premise regress comes to a stop in unmiddled premises, then it might seem that we have a serious problem for the notion of demonstration, just as the anti-demonstrators of Aristotle's regress argument claimed. However, notice that it is a matter of objective fact which propositions are unmiddled in this way: given the sum total of all the true propositions, we can apply a set of mechanical procedures to find out which ones are unmiddled (Aristotle in effect gives us such a set of procedures in *Prior Analytics* I.27). Moreover, if we did have knowledge of just exactly the unmiddled propositions, then since they are the propositions in which every regress comes to a stop, and since a regress can be reversed to become a deduction, we would have knowledge of premises from which every other proposition could be deduced. Since unmiddled propositions cannot be known except by non-demonstrative means, it follows that the possibility of non-demonstrative knowledge of the unmiddled propositions is both a necessary and a sufficient condition for the possibility of demonstrations. Since there is no middle term explaining why an unmiddled proposition is true, there is no explanation of its truth: it is, in effect, uncaused and unexplained. Aristotle's view is precisely this: demonstrations, which give the causes why their conclusions must be true, ultimately rest on first premises for the truth of which there is no further explanation or cause.

This brief account of Aristotle's theory raises a host of important questions, most critically the question of how it is possible to have knowledge of these first indemonstrable premises. I will not try to pursue that issue further here (see Smith 1986 for a little more detail). The point I wish to emphasize is that Aristotle's logical theory arose in response to a philosophical question about the possibility of proof. Aristotle's logic is, at its core, a philosophical logic.

5 Time and Modality: The Sea-Battle and the Master Argument

Necessity and possibility were subjects of major importance for ancient logicians. This might be seen as part of the Parmenidean legacy, since Parmenides asserted that what is must be and what is not cannot be: from there it is not a long distance to the view that what is the case is necessary and what is not the case is impossible. On such a view,

possibility and necessity collapse into one another. Only that which is, is possible; thus, what is possible is simply what is necessary, and there are no possibilities that are not actual. In other words, Parmenides' position appears to lead to a universal determinism or fatalism. Since such a view seems to rule out such things as free choice and deliberation, it runs into conflict both with common sense and with many philosophical views. Not surprisingly, we find considerable discussion of necessity and possibility in Greek philosophy. A good deal of that discussion involves the attempt to deal with these concepts in a logical system. Once again, we find that Greek logical theory developed in response to philosophical questions.

In *Metaphysics* IX.3, Aristotle ascribes the view that the modalities all collapse into one another to "the Megarians" and is at some pains to argue against it. Though he does not tell us who these Megarians were, we can supply a little history from other sources. Euclid of Megara (c. 430–c. 360 BCE), an approximate contemporary of Plato, was a member of Socrates' circle. He is said to have been influenced by Parmenides' views and to have maintained that "the good is one." We are told that he attacked arguments "not from their premises but from their conclusions;" what this means is not clear, but one possible interpretation is that Euclid followed Zeno in attacking rival positions by showing that they led to unacceptable consequences. A small circle of followers assembled around him, and from the beginning they appear to have had a strong interest in argumentation, especially in its dialectical form, in refutations, and in logical puzzles and paradoxes. Kleinomachus of Thurii, perhaps one of the first generation of Megarians, is said to have been the first to write on 'predications and propositions.' Eubulides, coming a generation or two later, is credited with the discovery of a number of paradoxes, including two of the most durable and difficult: the Liar and the Sorites. Eubulides engaged in a somewhat vitriolic controversy with Aristotle.

Now, Aristotle thought that the solution to Eleatic and Megarian arguments against motion and change could be found in a robust notion of potentiality. Aristotelian potentialities might be described as properties that point outside the present time. A lump of bronze, for instance, has the potentiality of being a statue, even though it is not one now, because it could, while remaining the same bronze, acquire the appropriate shape. Socrates, who is now seated, has the potentiality of standing up because he could, at some other time, acquire the property of standing up without ceasing to be Socrates. An intact garment has the potentiality of being cut up; a stone at the top of a hill has the potentiality of being at the bottom of the hill; a log has the potentiality of burning; an illiterate person has the potentiality of learning to read.

Potentialities make change possible, for Aristotle, since they allow him to describe change not at the coming to be of what was not but merely as the actualization of what was already in potentiality. For the bronze to become a statue, it is not necessary (as the Megarians might have it) that the lump of bronze cease to be and a new bronze statue emerges *ex nihilo*; instead, the same bronze persists, but a shape already possessed by it in potentiality becomes its actual shape. Aristotle extends this to a general definition of motion as "the actuality of what is in potentiality insofar as it is in potentiality." On this basis, he thinks that he can respond to Zeno's paradoxes of motion by claiming that a body in motion, while it is in motion, is never actually at any location: it is actually only in motion, only potentially at any of the points along its path. Were

it to stop, of course, it would actually be located at some point; but then, it would no longer be in motion.

I will not discuss here whether this is an effective response to Zeno: what is important is that it depends on a notion of potentialities as properties which things can have at a given time without exhibiting them at that time. The potentiality (capacity, ability) which Socrates has of standing up does not manifest itself while he is seated, but it is there nonetheless: when he stands, of course, it is no longer a potentiality but an actuality. Precisely this point is what the Megarians denied. They held that the only possible evidence for the claim that Socrates can stand up is for him actually to do so: however, his standing will provide no evidence that he could have stood up a moment ago while he was sitting, but only evidence that he can stand now while he is standing.

So far, this may seem to be primarily a matter of metaphysics. In *On Interpretation* 9, however, Aristotle presents us with an argument resting on logical principles. The background of the argument is the notion of a 'contradiction' or 'contradictory pair' (*antiphasis*): two propositions with the same subject, one of which denies of that subject exactly what the other affirms of it (for example, 'Socrates is seated,' 'Socrates is not seated'). In general, Aristotle says that for any contradictory pair at any time, one of the pair is true and the other false. He finds a problem, however, if we allow this to extend to propositions about the future. All we need is the additional thesis that whatever is true about the past is now necessarily true and the general semantical principle that if a proposition is true, then whatever it says is the case is indeed the case. Imagine now that yesterday, I said, 'There will be a sea-battle tomorrow.' By the general principle governing contradictory pairs, either this sentence or its contradictory 'There will not be a sea-battle tomorrow' must have been true when I made my statement. If the sentence was true, then it is now a truth about the past that it was true, and therefore it is now necessary that it was true; therefore, it is now necessarily true that there is a sea-battle today. If, on the other hand, my statement was false, then by similar reasoning it is now necessarily false that there is a sea-battle today. Since my statement was either true or false, then it is now either necessary or impossible that there is a sea-battle today. But this can be generalized to any event at any time, since (as Aristotle says) surely it does not matter whether anyone actually uttered the sentence: thus, everything which happens happens of necessity, and there are no possibilities which do not become actual. It is far from clear just how Aristotle responds to this puzzle, except that he is certain that its conclusion must be rejected. One interpretation is that in order to avoid the repugnant conclusion, he restricts the application of the law of excluded middle to future propositions (the literature on this argument is enormous: see the Suggested Further Reading below for a few places to start).

Aristotle does not tell us the source of the argument to which he is responding in *On Interpretation* 9, though it is a reasonable guess that its author was Megarian. One piece of evidence in favor of that is the 'Master' argument developed by Diodorus Cronus (c. 360–c. 290 BCE). Our sources identify Diodorus as a Megarian (though some scholars have disagreed); his dates are unclear, and it is just possible that Aristotle is actually responding to Diodorus, though I think it more likely that he is replying to an ancestor of the Master developed by other Megarians. In any event, the Master began with a proof that the following three propositions form an inconsistent triad, so that the affirmation of any two entails the denial of the third:

1. What is past is necessary.
2. The impossible does not follow from the possible.
3. There is something possible which neither is nor will be true.

The first of these recalls the argument of *On Interpretation* 9. What the second means is not totally clear, but one reading is 'a possible proposition cannot entail an impossible one.' We do not know how Diodorus argued for the incompatibility of this triad, but we do know the conclusion he drew from it: he affirmed the first two propositions and deduced the denial of the third, so that for him 'possible' was equivalent to 'either now true or true in the future.' His view here conflicts directly with Aristotle, who asserts that there are possibilities that never become actual. What Diodorus may have been doing, in addition to defending a Megarian view of universal necessitation, was finding a way to talk about possibilities in a Megarian view of the world. That is, his position would allow him to assert that there is indeed a meaning for the word 'possible,' even though nothing can happen except what does happen.

The later history of the Master is closely associated with the Stoic school, which began with Zeno of Citium (335–263 BCE). Zeno learned logic from Megarian teachers, and Zeno and his follower Cleanthes (331–232 BCE) responded to the Master. Subsequently, Chrysippus (c. 280–207 BCE), the most distinguished logician among the Stoics and probably the most gifted and prolific logician of the Hellenistic period, affirmed the first and third propositions of the Master and denied the second: he argued that 'an impossible can follow from a possible.' To understand his response, we need first a brief sketch of his theory of propositions. For Chrysippus, a proposition – that is, what is true or false – is really an incorporeal entity, roughly the meaning of a sentence that expresses it (the Stoics called this a *lekton*, 'sayable,' which might plausibly be translated 'meaning' or 'sense'). There are similarities between this notion and, say, a Fregean notion of the sense of a proposition, though there are important differences. One important difference is that the Stoics thought of at least some propositions as changing their truth values over time, for example the proposition expressed by 'It is day' is at one time true and at another false while remaining the same proposition. Another Stoic thesis, and one that is crucial to Chrysippus' solution, is that propositions about individuals specified by demonstratives 'perished' when the individuals ceased to exist. If I point to Dion and say 'He is alive,' then I utter a proposition the subject of which is fixed by a demonstrative (in modern terms, an indexical). However, if Dion dies, then I can no longer point to Dion at all, since he does not exist; therefore, the proposition that was formerly expressed by 'He is alive' also ceases to exist rather than becoming false. Now, Chrysippus offers for consideration the proposition 'If Dion has died, then this one has died' (pointing to a living Dion, obviously). Since 'this one' refers to Dion, this conditional sentence is obviously true: its consequent follows from its antecedent. However, when Dion has died, the antecedent of the conditional becomes true while its consequent perishes: in fact, it is in a sense impossible for 'This one has died' ever to be true, since the condition for its truth is also the condition for its perishing. Therefore, we have an example of something impossible following from something possible.

Both the Stoics and Aristotle, then, investigated logical modalities in order to reconcile logical theory with their views about determinism. Chrysippus, who was a

determinist, could nevertheless argue that his views did not entail that only what is necessary is possible, since he can produce an example of a proposition that is possible but that neither is nor will be true: 'This one has died.' Aristotle, who rejects universal necessitarianism and develops a complex theory of potentialities to accommodate his views on motion and deliberation, at least recognizes that his position will require some radical modification of his logical theory. (For more on the Master argument, see the readings cited below, especially Fine 1984; Gaskin 1995; Prior 1967.)

6 Sentential Logic in Aristotle and Afterwards

Aristotle never developed an account of sentential logic (the inferences that rest on sentential operators such as 'and,' 'or,' 'if,' 'not'). In my opinion, this is closely connected with his use of his logical theory in the *Posterior Analytics*. His argument that 'every regress terminates' can only work if the logic of arguments 'in the figures' is the only logic there is; and for that to be so, every proposition must either affirm or deny a predicate of a subject. In fact, Aristotle thinks that this is so, and he undertakes to show it in the *Prior Analytics*. This requires him to reject sentential composition: he does not recognize conjunctions, disjunctions, or conditionals as individual propositions. Precisely how this is to work is not clear, though we can discern a few details. For instance, because he treats affirmations and denials as two basic types of sentence, he does not think of negations as compound sentences; he appears to regard conjunctions not as single compound sentences but only as, in effect, collections of sentences (i.e. their conjuncts); and he treats conditionals not as assertions but as agreements to the effect that one sentence (the antecedent of the conditional) entails another (the consequent). Subsequent logicians, including Aristotle's own close associate Theophrastus, did not follow him in this and instead offered analyses of the role of sentential composition in arguments. With Chrysippus, this develops into a full-fledged sentential logic, resting on five 'indemonstrable' forms of inference. The Stoics stated these using ordinal numbers as place-holders for propositions:

1. If the first, then the second; the first; therefore the second.
2. If the first then the second; not the first; therefore not the second.
3. Not both the first and the second; the first; therefore not the second.
4. Either the first or the second; the first; therefore not the second.
5. Either the first or the second; not the first; therefore the second.

The Stoics then demonstrated the validity of other valid arguments by means of these indemonstrables (unfortunately, our knowledge of their views is very fragmentary: see Kneale and Kneale 1978; Mates 1953; Mueller 1978 for reconstructions). There may be some connection between the Stoic acceptance of sententially compound propositions and their views on the nature of propositions.

Aristotle may have another reason for being concerned about sentential logic. He wanted to allow for possibilities that never become actual, and to do that he analyzed possibility in terms of a notion of potentiality. This works best with subject–predicate sentences, where possibility can be seen as a matter of the subject possessing a poten-

tiality; it is very difficult to extend it to compound propositions. In fact, Aristotle appears to have had some reservations about treating propositions as entities at all, perhaps because this appeared to give support to the argument from the necessity of past truth in *On Interpretation* 9. The Stoics, with their theory of 'sayables' as the bearers of truth and falsehood and their acceptance of a kind of determinism, had a much easier time developing a logic of sentential composition. Here again, a difference in logical theory may have been closely entwined with a difference in philosophical standpoint.

References

Primary texts

Aristotle, Prior Analytics, Posterior Analytics, On Interpretation, Categories, Topics, On Sophistical Refutations.

Plato, Theaetetus, Sophist.

Translations with commentary

Ackrill, J. L. (1963) *Aristotle, Categories and De Interpretatione*. Oxford: Clarendon Press (Clarendon Aristotle Series).

Barnes, Jonathan (1993) *Aristotle, Posterior Analytics*. 2nd edn. Oxford: Clarendon Press (Clarendon Aristotle Series).

Smith, Robin (1989) *Aristotle, Prior Analytics*. Indianapolis: Hackett.

General history of Greek logic

Kneale, William, and Kneale, Martha (1978) *The Development of Logic*. 2nd edn. Oxford: Clarendon Press.

On Aristotle's logic

Corcoran, John (1972) Completeness of an ancient logic. *Journal of Symbolic Logic*, 37, 696–705.

Corcoran, John (1973) A mathematical model of Aristotle's syllogistic. *Archiv für Geschichte der Philosophie*, 55, 191–219.

Lear, Jonathan (1980) *Aristotle and Logical Theory*. Cambridge: Cambridge University Press.

Łukasiewicz, Jan (1957) *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*. 2nd edn. Oxford: Clarendon Press. (Classic study that initiated modern interpretation of Aristotle's logic.)

Smiley, Timothy (1974) What is a syllogism? *Journal of Philosophical Logic*, 1, 136–54.

Smith, Robin (1986) Immediate propositions and Aristotle's proof theory. *Ancient Philosophy*, 6, 47–86.

Whitaker, C. W. A. (1996) *Aristotle's De Interpretatione: Contradiction and Dialectic*. Oxford: Clarendon Press. (Includes a discussion of *De Interpretatione* 9.)

On Megarian and Stoic logic

Baltzly, Dirk (2000) Stoicism. *Stanford Encyclopedia of Philosophy*. (Includes a very readable summary of Stoic logic.)

Döring, Klaus (1974) *Die Megariker: kommentierte Sammlung der Testimonien*. (Definitive study of the Megarians; not available in English.)

Frede, Michael (1974) *Die stoische Logik*. Vandenheck and Ruprecht. (The most comprehensive account available of Stoic logic.)

Long, A. A. and Sedley, D. N. (1987) *The Hellenistic Philosophers*. 2 vols. Cambridge: Cambridge University Press. (Comprehensive discussion of all the Hellenistic schools, including the Stoics, as well as Diodorus Cronus. Volume 2 contains an extensive bibliography of scholarly books and articles.)

Mates, Benson (1953) *Stoic Logic*. Berkeley: University of California Press.

Mueller, Ian (1978) An introduction to Stoic logic. Chapter 1 in J. M. Rist (ed.), *The Stoics*. Berkeley, Los Angeles, London: University of California Press.

On the Master argument and the argument of On Interpretation *9*

Anscombe, G. E. M. (1956) Aristotle and the sea battle. *Mind*, 65, 1–15.

Fine, Gail (1984) Truth and necessity in *De Interpretatione* 9. *History of Philosophy Quarterly*, 1, 23–47.

Gaskin, Richard (1995) *The Sea Battle and the Master Argument: Aristotle and Diodorus Cronus on the Metaphysics of the Future*. Berlin and New York: Walter de Gruyter.

Hintikka, Jaakko (1973) *Time and Necessity: Studies in Aristotle's Logic of Modality*. Oxford: Clarendon Press.

Prior, Arthur (1967) *Past, Present and Future*. Oxford: Clarendon Press.

Rescher, Nicholas (1966) A version of the "Master Argument" of Diodorus. *Journal of Philosophy*, 63, 438–45.

Sorabji, Richard (1978) *Necessity, Cause, and Blame*. London.

2

History of Logic: Medieval

E. P. BOS AND B. G. SUNDHOLM

Seven 'liberal arts' constituted the curriculum at a medieval arts faculty. The three 'trivial' arts Grammar, Logic (*Dialectica*), and Rhetoric deal with the use of words rather than with (real) things. These are dealt with in the four mathematical arts – Geometry, Arithmetic, Astronomy, and Harmony (Music) – that comprise the *quadrivium*. The specific logical art is concerned with *reasoning*. The logical tradition is as old as Aristotle and history knows periods of intense logical activity. Thus the subject is known under many names and, at different times, knows varying boundaries. Aristotle did not use the Greek *logikê* for the logical art, but preferred *ta analytika* (from the verb *analuo*: to resolve (into premises or principles), from which the names of his 'sweet Analytics,' that is *Analytica priora and posteriora* derive. The Greek *logos* can be found in the writings of both Plato and Aristotle, where it stands for (the smallest meaningful parts of) 'speech' whereby something can be said. The Greek logical terminology was latinized by Cicero and Boethius, and the honour of having named the subject belongs to the former who coined *Logica*. 'Dialectica', the alternative Platonic and Stoic name for logic as part of the *trivium*, derives from the Greek for conversation, since, in this tradition, thinking is seen as the soul's conversation with itself. The dialectician investigates relations between (eternal) ideas which have to be respected if the thinking were to be proper. In the sixth century the logical works of Aristotle – *Categories*, *On Interpretation*, the two *Analytics*, the *Topics*, and *On Fallacies* – came to be seen as an *Organon* (instrument, tool), and the term has stuck, for example in *Novum Organon* (1620), that is, Francis Bacon's attempt to emend Aristotle's instruments for reasoning.

These names, under which the discipline has been known, relate to different aspects of logic, or of how the subject should be seen. 'Logic,' thus, would be the study of (the use of words for making) reasoned claims, and 'Analytics' resolves reasoning into simpler parts in order to provide grounds. 'Dialectics' grounds reasoning in (eternal) relations between logical entities, whereas when logic is thought of as an organon, it serves as the tool for multiplying knowledge through the use of reasoning.

The purely *formal* logic of today is regularly confined to theory of (logical) consequence between well-formed formulas (WFFs). An analogous position within medieval logic would cover only the topics dealt with in the Prior Analytics. Medieval logic, however, covers a much wider range: it comprises also topics from philosophy of

language, for example the theories of signification and supposition (reference), epistemology, for example the theory of demonstration, and philosophy of science (methodology), for example the method of analysis and synthesis. Indeed, logic is sometimes divided into Formal logic versus Material logic, which correspond to Aristotle's two *Analytics*, and cover, respectively, the theory of consequence and the theory of demonstrations (or proofs). Today's logician is primarily a 'dialectician' who studies relations among logical entities, be they meaningful sentences, (abstract) propositions, or the well-formed formulae of a formal language. The medieval logician, on the other hand, was primarily concerned with the exercise of the faculties of the intellect. The use of reasoning as part of the (human) act of demonstration was his main concern. Today the theory of consequence holds pride of place in logic over and above the theory of demonstration (which is commonly not even seen as a part of logic), but in medieval logic their order of priority was the opposite. The Posterior Analytics was in no way inferior to the Prior Analytics. The medieval logician does not primarily study consequence-relations between logical entities; his concern is the act of knowledge that is directed toward real things.

However, prior to studying proper acts of reason, one has to take into account also two other kinds of acts, since reasoning proceeds from judgments that are built from terms. In the first instance, the latter two notions are also the products of mental acts according to certain operations of the intellect, namely apprehension and judgment.

The medieval teaching on the act of reason can be summarized in tabular form:

	Operation of the intellect	*Inner product of the act*	*Outward sign*
I	*(Simple) Apprehending, Grasping*	Concept, Idea, Notion, (Mental) Term	(Written/spoken) Term
II	*Judging,* Composition/Division of two (mental)terms	Judgment (made), (Mental) Proposition: *S* is *P*	(Written/spoken) Assertion, Proposition
III	*Reasoning*, Inferring	(Mental) Inference	(Written/spoken) Inference, Reasoning

Its influence is still visible in the nineteenth century, after half a millennium, when traditional textbooks still show the time-honored structure, comprising the three parts: Of Terms, Of Judgement and Of Inference (sometimes adding a fourth, post-*Port Royal Logic* (1662), part: Of Method). It must be stressed that the medieval notion of 'proposition' that occurs twice in the second row, either as the traditional subject/copula/predicate judgment made, that is, the mental proposition, or as its outward linguistic guise, is *not* the modern one. The term *proposition* enters contemporary logic as Bertrand Russell's unfortunate (mis-)translation of Frege's *Gedanke* ('Thought'). Thus, modern propositions are not judgments, but *contents* of judgments. As such they may be given by nominalized that-clauses, for instance

that snow is white,

which emphasizes their being abstract contents. This, though, is not the way to think of medieval propositions, which are not contents, but combinations of terms *S* and *P*, for instance,

[snow is white], and [Sortes is a man].

(The fourteenth-century *complexe significabile*, though, plays a role that is somewhat analogous to that of the modern notions of proposition (content).)

In medieval logic there is a complete parallelism between thought and reality, between mind and world. The important idea of carrying out purely mechanical, 'formal,' proofs, irrespective of content, emerges only with Leibniz, and does not yet form part of the medieval tradition in logic. Owing to this logical 'picture theory' *avant la lettre* for the relation between mind and world, the theory of categories, especially in the form of simple predications, or *categorizations*, [a is an α], is sometimes seen as part of logic (as well as of metaphysics).

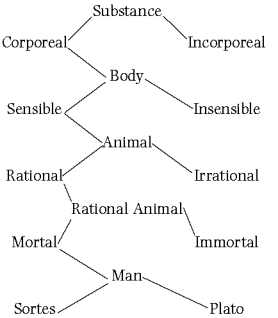
The medieval theories as to the truth of propositional combinations of terms – categorical predications – vary. According to one theory, the (extensional) *identity* theory, the proposition [S is P] is true when the supposition of both terms is the same, that is, when both terms stand for the same entity. Thus, for instance, the predication [Sortes is a man] is true when [Sortes] and [man] both supposit for the same entity, namely Socrates. The main rival of the identity theory of truth is the (intensional) *inherence* theory. According to it, the proposition [Sortes is a man] is true when *humanity*, the property of being a man 'inheres' in (is contained in) the nature of what *Sortes* stands for, namely, Socrates. In modern historical studies the rivalry between these medieval theories is sometimes seen as absolute. However, sometimes a philosopher is committed to (uses of) both conceptions. It seems more likely, though, that the alternative conceptions of truth-conditions pertain to different kinds of predication, than that the philosopher in question wavers between two absolute, all-encompassing theories. For instance, the substantival predication [Man is an animal] is held to be true because the terms man and animal stand for the same entity, whereas the denominative predication [A man is white] is deemed true because whiteness inheres in what man stands for.

A propositional combination of terms can be just apprehended, that is, grasped or understood; it need not be judged, or, when considered in the exterior mode, asserted. Of course, the medieval logicians also realized that not all traditional judgments have categorical [S is P] form. There are also hypothetical and disjunctive judgments, which take, respectively, the forms

[if J_1, then J_2] and [J_1 or J_2],

where J_1 and J_2 are judgments.

Terms can be divided into *general*, for instance, *man*, and *singular*, for instance, *Sortes*. Accordingly, by the correlation between world and mind/language, so can their significations, that is, there is a matching division of singular and general natures. We then get hierarchies of terms that can be ordered in a so-called *Porphyrian tree:*



With respect to such trees, we encounter reasonings based on predications:

Sortes is a man, and man is a rational animal. *Therefore*: Sortes is an animal.

We can, however, ascend in the Porphyrian tree:

An animal is a animate living body. *Therefore*: Sortes is a living body.

Apparently, predication is transitive when climbing in a Porphyrian tree: what is predicated of a predicate of a subject, can be predicated also of the original subject.

However, not all categorical predication is transitive: the two premises

Sortes is a man and Man is a sort,

obviously, do not allow for the nonsensical conclusion

Sortes is a sort.

In order to account for the failure of transitivity in the case of iterated predication, contemporary logical semantics relies only on a (meager) *reference* relation, both relata of which, namely, the expression and its reference, are construed as *things*. Medieval logic, to its credit and great advantage, draws upon a richer spectrum of semantic notions. In effect, the medievals split our modern notion of reference into two notions, namely signification and supposition. The language studied by medieval logicians is a highly stylized, technical Latin, with rigid syntactic rules and clear meaning and in this it resembles, not our current metalinguistic predicate-calculus, but rather those interpreted formal languages that were used by Frege and others to inaugurate modern logic. The carefully crafted systems of the Polish logician Stanislaw Lesniewski are particularly close to the medieval perspective, since they were cast in the mold of traditional logic, using the [S is P] propositional form, rather than the modern, Fregean function/argument form [P(a)], as their point of departure. The expressions of these

formal languages were not seen just as things, but as *signs*, where a sign signifies by making manifest its signification to mind. The notion of *signification* is the closest medieval counterpart to our modern notion of reference. Thus, for instance, the signification of the name *Sortes* is the man Socrates and the signification of the general name *man* is such that the name can be rightly predicated of men. Signification is context-independent, but medieval logic also knows a context-sensitive notion, namely that of *supposition*. Supposition primarily applies to terms that occupy the subject position in [S is P] propositions. The supposition of a term, in a certain propositional context, is what the term stands for in the context in question. What supposition the subject term S takes depends on the signification of the predicate P. In the proposition

[Sortes is a man]

the term *Sortes* has personal supposition, because it stands for the individual Socrates. If we consider the true propositions

[Man is a sort] and [Man is a word]

the term *man* has moved from predicate to subject position. In the proposition

[Man is a word]

it has *material* supposition, because it stands for the word and not the person whence the modern use of quotation-marks is superfluous. It is the term man that has material supposition and not the term 'man.' This reverses current (Carnapian) terminology, where, when speaking about the word, one uses the 'formal,' rather than 'the material mode of speech.' The medieval terminology *material* and *formal* supposition probably derives from the fact that, under the influence of Aristotle's theory of hylemorphism, the subject S is seen as the *matter* of the categorical [S is P]-proposition, and the predicate is its *form*. Similarly, in the proposition

Man is a sort

the term man has *simple* supposition; here it stands for the *species* of men rather than for individual men. The failure of transitivity in the above inferences can then be accounted for by observing that a shift in supposition occurs in the premises: in one the supposition of *man* is formal whereas in the other it is simple, and so the inference is barred.

The theory of consequence in medieval logic, of course, treats of the Aristotelian theory of the syllogism, that is the theory of inference among categorical judgments. Such judgments have the S is P form, but they are not just simple predications such as [Sortes is (a) man]. The copula can vary both in quality and quantity. An affirmative judgment has the form [S is P] and a negative one has the form [S is not P], whereas a universal judgment has the form [all S are P] and a particular one has the form [some S are P]. Thus, for instance, a particular negative judgment takes the form [some S are not P]. Medieval logic summarized the basic inferential properties between such cate-

gorical judgments in the Aristotelian *square of opposition*. In *An. Pr.* Aristotle had organized the syllogism according to three 'figures' (subsequently also a fourth figure was considered by Galen) and determined the 'valid syllogistic modes' by means of reducing the valid modes in later figures to the 'perfect' syllogisms in the first mode. The well-known mnemonic descriptions '*Barbara, Darii, Celarent*, etc.' of the valid modes of inference were given in the Middle Ages; these descriptions provide codes for the reduction of the validity of modes in the later figures to the primitive validity of the perfect modes in the first figure. Decent expositions can be found in any number of texts on traditional logic.

As is well-known, the Aristotelian theory validates inferences that are not held to be valid in current logic. First among these is the instantiation of universal judgments:

All swans are white. *Therefore*: there is a white swan.

Aristotelian terms were reached by *epagogé* (Aristotelian induction). You grasp the concept *swan* by seeing an instance thereof, which particular exemplar serves as an *exempla gratia* for the sort in question. Thus the inference is valid and the universal categorical judgments carry 'existential import.' Today, within current predicate logic the example would be regimented as

$\forall x(\text{Swan}(x) \supset \text{White}(x))$. *Therefore*: $\exists x(\text{Swan}(x) \,\&\, \text{White}(x))$

which inference is not valid. Only the step to the conclusion

$\exists x(\text{Swan}(x) \supset \text{White}(x))$

is valid. This, however, is not a regimentation of 'there is a white swan,' but only of 'there is something which is such that if it is a swan then it is white,' and this claim, given the premise that everything is such that if it is swan then it is white, is completely trivial as long as the universe of discourse is not empty: *any* object is such an object. The inference from an affirmative universal proposition to an affirmative particular one is an example of 'alternation.' Other similar kinds of inference concern 'descent' from the universal judgments to a conjunctive one:

All men are mortal. *Therefore*: Peter is mortal and John is mortal.

(Of course, there is no need to limit ourselves to just two conjuncts here. *Mutatis mutandis* this remark applies also to the examples given in the sequel.) Similarly,

Some men are mortal. *Therefore*: Peter is mortal or John is mortal.

is a descent to a disjunctive proposition. One can also descend with respect to *terms*:

All men are mortal. *Therefore*: John and Simon are mortal.

Aristotelian logic, when cast in the mold of traditional syllogistic theory, is a *term-logic*, rather than a logic of propositions. The medievals liberated themselves from the term-logical straitjacket of the Aristotelian syllogistics, first by considering also

syllogisms with *singular* judgments, that is, categorical [S is P] propositions of the form [s is P], where s is a *singular term*. Here the so-called *expository syllogism* played an important role:

This thing (*hoc*) is a man, but this thing runs. *Therefore*: A man runs.

However, gradually also other forms of inference than term-logical syllogisms were studied by medieval logicians, including the pure and mixed *hypothetical* syllogisms. A pure hypothetical syllogism takes the form

If P then Q and if Q, then R. *Therefore*: If P, then R.

The mixed forms of the hypothetical syllogism include the well-known *modus (ponendo) ponens* inference:

If P, then Q, but P. *Therefore* Q.

Here we have left the term-logic of syllogistic theory; the connections are here not between terms, but between propositions. This shift in perspective led, (± 1300) to the appearance of a new logical genre. Then tracts bearing the title *On Consequence* begin to appear, and consequence becomes the main topic of study in medieval logic.

In such tracts rules for the holding of consequences were set out. Today, in elementary logic classes, when the analysis of natural language arguments is treated, students are taught to search for argument indicator words, such as 'thus,' 'therefore,' 'hence,' 'whence,' 'because,' etc. However, today we also make a clear distinction between implication, consequence, inference and causal grounding:

- 'implies' is an indicator-word for *implication*, which is a propositional connection between proposition(al content)s.
- 'follows from,' 'is a consequence of' and 'if . . . is true, then – is true' are indicator-phrases for *consequence*, which is a relation between proposition(al content)s.
- 'thus,' 'therefore' are indicator words for *inference*, which is a passage from premise judgment[s] (assertion[s]) to a conclusion judgment (assertion).
- 'because,' 'is a cause (ground, reason) for' are indicator words for *causal grounding*, which is a relation between events, or states of affairs.

However, in medieval logic, *si* (if), *igitur* (therefore), *sequitur* (follows) and *quia* (because) are all indicator-words for one and the same notion of a *consequentia*. This notion survives terminologically in modern logic under two different guises, namely, on the one hand, as the notion of *(logical) consequence* between WFFs that derive from Bolzano's *Ableitbarkeit* and that was made famous by Tarski, and, on the other hand, as the *sequents* (German *Sequenzen*) that were used by Gentzen. The medieval theory of consequences, accordingly, can rightly be seen as a partial anticipation of contemporary sequent-calculus renderings of logical systems. The modern notion of logical consequence has its medieval counterpart in the notion of a *formal* consequence, that is, one that holds 'in all terms,' for instance:

All men are mortal. Sortes is a man. *Therefore*: Sortes is mortal.

This consequence remains valid under all (uniform) substitutions (*salva congruitate*) of other terms put in place of *Sortes*, *mortal*, and *man*. Formal consequence is opposed to *material* consequence, for instance the consequence

Sortes is a man. *Therefore*: Sortes is mortal.

holds only materially, since it does not hold 'in all terms.' Material consequence can be compared to (Carnap's contemporary notion of) 'meaning postulates.'

Another very interesting, late addition to medieval logic is the theory of *obligations*, which is concerned with the proper rules for disputation and questioning. Thus, for instance, if I have asserted a conjunctive proposition, I have incurred an obligation and might be held to be asserting each conjunct separately. This theory lies on the borderline between logic, semantics, and pragmatics, incorporating also elements of the theory of speech acts. To an amazing extent, it constitutes an anticipation of the current *dialogicial* approach to logic and semantics that was designed by Lorenzen and Lorenz, or the *game-theoretical semantics* that we owe to Hintikka.

In contemporary philosophical logic, logical paradoxes and their resolution – their diagnosis and prevention – are treated if and when they arise. Their treatment does not constitute a separate branch of logic. In (late) medieval logic, however, a novel genre was added to the standard logical repertoire and tracts devoted solely to the treatment of *Insolubilia* begin to appear.

Not all of medieval logic is confined to logic texts, though. The role that philosophy served in medieval academic life was primarily that of an *ancilla theologicae* ('a servant of theology'). Therefore, one can often find passages that are highly relevant from a logico-semantical point of view also outside tracts that are devoted specifically to matters logical. In particular, treatments of delicate theological questions, for instance, in the *Commentaries* on Peter Lombard's *Sentences* (that is, the obligatory introductory compendium to the study of theology), often contain material that is highly illuminating from a logical point of view. The vexing questions concerning the nature of the Trinity and the interrelations of Its Persons illustrate this sufficiently. Two other topics that stand out in this respect are the question whether God's existence can be demonstrated and the treatments of the various Names of God. Thomas Aquinas does not enjoy a high reputation as a logician; his fame rests on his contribution to metaphysics and the philosophy of mind. Nevertheless, his *Summa Theologica* contains much that is of great relevance for contemporary philosophy of logic and language. Thus, for instance, in his discussion of the Names of God in Question 13 Aquinas anticipates Frege's ideas concerning names with different modes of presentation of the same object.

Furthermore, concerning the demonstrability of God's existence we read:

> A proposition is *per se nota* because the predicate is included in the nature of the subject: for instance, *Man is (an) animal*, for *animal* is contained in the nature of *man*. (*Summa Theologica*, I.ii.)

This passage ought to yield a *déjà lu* experience. Most of us, certainly, will have read this explanation of a proposition *per se nota*. The German text from which we know it is not

medieval, but was published 500 years later, in 1781, by a professor of philosophy at Königsberg in Eastern Prussia. There, though, the same formulation is used to explain the notion of an *analytic judgment*.

A Timeline of Medieval Logicians

Before XI
Porphyry (232–305)
Augustinus (354–430)
Boethius (480–524)

XI
Abbo of Fleury
Garlandus Compotista
Anselm of Canterbury (d.1109)

XII
Peter Abailard, 1079–1142
Adam Parvipontanus
Gilbert of Poitiers, 1080–1154
Alberic van Reims
John of Salisbury, c. 1120–1180

XIII
Peter of Spain (d.1277)
William of Sherwood (1210?–66/70)
Robert Kilwardby (d. 1279)
Albert the Great (1200–80)
Roger Bacon (1215–94)

XIII (*cont.*)
Boethius of Dacia (c. 1270)
Henry of Ghent (c. 1217–93)
Ralph Brito (c. 1290–1330)
Siger of Kortrijk (d. 1341)
Simon of Faversham (c. 1300)
John Duns Scotus (1265–1308/9)

XIV
Walter Burleigh (c.1275–1344/5)
William of Ockham (1285–1347)
Robert Holkot (c.1290–1349)
William of Heytesbury (d.1272/3)
Gregory of Rimini (c.1300–1358)
John Buridan (c.1300–after 1358)
Nicholas of Autrecourt (c.1300–after 1358)
Richard Billingham, (c.1350–60)
Albert of Saxony (1316–1390)
Marsilius of Inghen (c.1340–1396)
Vincent Ferrer (c.1350–1420)
Peter of Ailly (1350–1420/1)
Paul of Venice (1369–1429)
Paul of Pergola (1380–1455)
Peter of Mantua (d. 1400)

A Guide to the Literature

The Aristotelian *Organon* is, of course, a prerequisite for medieval logic. G. Patzig, *Aristotle's Theory of the Syllogism* (First German edn 1959) English translation by J. Barnes (Reidel: Dordrecht, 1969) is still the classical treatment of Aristotle's theory, and Paul Thom, *The Syllogism* (Munich: Philosophia Verlag, 1981) offers a most thorough modern presentation. A. N. Prior's lemma "Logic, Traditional" in: Paul Edwards (ed.), *Encyclopaedia of Philosophy* (New York: Macmillan, 1967) gives a compact, yet lucid overview. H. W. Joseph and R. D. McKirahan, *Principles and Proofs* (Princeton University Press, 1992) treats of Aristotelian demonstrative science, a topic of paramount importance for medieval logic. Valuable surveys of medieval logic can be found in the general histories by W. Kneale and M. Kneale, *The Development of Logic* (Oxford: Clarendon, 1962) and I. M. Bochenski, *Formale Logik*, English tr. by Ivo Thomas: *A*

History of Formal Logic (Notre Dame University Press, 1963). Surveys of medieval logic have been offered by E. A. Moody, *Truth and Consequence in Medieval Logic* (Amsterdam: North-Holland, 1953), Norman Kretzmann, "Semantics, History of" in: Paul Edwards (ed.), *Encyclopaedia of Philosophy* (New York: Macmillan, 1967), Jan Pinborg, *Logik and Semantik im Mittelalter* (Stuttgart-Bad Cannstatt: Fromann-Holzboog, 1972). Of these we have found the trenchant studies of Pinborg and Kretzmann especially useful. Moody draws liberally upon the notations and conceptual resources of modern (Frege–Russellian) predicate logic for his exposition of medieval notions, but the extent of his success in doing so is doubtful, owing to the differences in the forms of judgments used: medieval logic used the form of judgment (S is P) whereas (post-)Fregean logic uses the form of judgment (the judgable content A is true). It is still very much an open question how best to utilize the insights and achievements of modern metamathematical logic (which builds on Fregean logic) for the study of medieval logic in a non-anachronistic way. The systems of Lesniewski are based on traditional rather than Fregean logic, and might work much better here. A standard reference is D. P. Henry's lucid *Medieval Logic and Metaphysics* (London: Hutchinson, 1972) that also serves as an admirable introduction to Lesniewski.

The German *Historisches Wörterbuch der Philosophie* gives an incomparable survey of medieval logic. Individual, detailed lemmas, for instance, those on "Prädikation" and "Logik" have been of great help to us. This dictionary is also an invaluable guide, not just to medieval logic, but to the entire conceptual development of logic.

The Cambridge History of Later Medieval Philosophy, eds. N. Kretzmann, J. Pinborg, and A. Kenny (Cambridge University Press, 1982) is a universal compendium of medieval logic, with a companion volume of original texts *The Cambridge Translations of Medieval Philosophical Texts*: vol. I, *Logic and the Philosophy of Language*, eds. N. Kretzmann and E. Stumpf (Cambridge University Press, 1988). The equally monumental *Logica Modernorum*, vol. II (two parts), (Assen: Van Gorcum, 1967) by L. M. de Rijk, contains the original sources for the theory of supposition and other basic properties of terms.

Among original works we have found the William of Sherwood's thirteenth-century textbook *Introduction to Logic* (English translation by Norman Kretzmann), (Minneapolis: University of Minnesota Press, 1966) a useful general introduction to most issues covered in the present chapter. A later treatment, by almost a century and a half (±1400), of roughly the same material is offered by Paul of Venice in the *Logica Parva* (ed. and tr. by A. Perreiah), Philosophia Verlag (Washington: Catholic University of America Press, 1984). The British Academy supports a multi-volume edition/translation of the magisterial *Logica Magna* by the same Paul of Venice. William of Ockham's *Summa Logicae* has been partly rendered into English: part I (tr. M. Loux) and part II (tr. A. Freddoso and H. Schurmann) (Notre Dame University Press, 1974, 1980). Furthermore, the series *Philosophisches Bibliothek*, published by Felix Meiner Verlag, (Hamburg, contains many bilingual (Latin/German) editions, with introductions and careful annotations, of important works in medieval logic.

The Routledge series *Topics in Medieval Philosophy* contains volumes of interest for the general philosopher: Ivan Boh, *Epistemic Logic in the Later Middle Ages* (London, 1993) is particularly interesting on the epistemological aspects of the theory of consequences, while A. Kenny, *Aquinas on Mind* (London, 1993) spells out interesting par-

allels between medieval conceptions and those of Wittgenstein. Simo Knuuttila, *Modalities in Medieval Philosophy* (London, 1993) contains much that is of interest for the modern theory of modality, as does John Duns Scotus, *Contingency and Freedom: Lectura I 39* (ed. and tr. by A. Vos Jaczn. et al.), *New Synthese Historical Library*, vol. 42 (Dordrecht: Kluwer, 1994). Mikko Yrjönsauuri's Helsinki dissertation *Obligationes – 14th Century Logic of Disputational Duties*, in: *Acta Philosopphica Fennica*, 55 (1994), summarizes much of what is known about the theory of obligations. G. E. Hughes, *John Buridan on Self-Reference* (Cambridge University Press, 1982) is a perfect example of a medieval treatment of logical paradoxes.

There are two (English language) journals devoted to medieval philosophy, namely *Vivarium* and *Medieval Philosophy and Theology*. Of these, the first has a long tradition of articles within medieval logic and semantics. The *History and Philosophy of Logic*, *The Journal of Philosophical Logic*, and *The Notre Dame Journal of Formal Logic* also publish articles on medieval logic.

3

The Rise of Modern Logic

ROLF GEORGE AND JAMES VAN EVRA

The history of some sciences can be represented as a single progression, with each dominant theory coming to the fore, then eventually falling, replaced by another in succession through the centuries. The development of physics, for instance, can be understood as such a chain, connecting Newton in the seventeenth century with Einstein in the twentieth. Logic did not progress in this way; no dominant theory commanded it (a tapestry more than a chain) until the first decades of the twentieth century. No self-sustaining internal theory held sway before then, nor was there much rigor externally imposed. Even Aristotle, as one commentator put it, was more venerated than read, and most versions of syllogistic logic proposed after the Middle Ages did not measure up to the sophistication of his own system.

1 The Dark Ages of Logic

In 1543 the French humanist and logician Peter Ramus (1515–72), who had made a name for himself with his dissertation *Whatever Aristotle Has Said is False*, published his *Dialectic*, a slim book that went through 262 editions in several countries and became a model for many other textbooks. Ramus gratified the taste of the times by writing an elegant Latin, drawing his examples from Cicero and other classical authors, and by neglecting most of the finer points of medieval logic and the associated 'barbarous' technical vocabulary. The book was committed not to logic as we now know it, but to the art of exposition and disputation. Its first sentence, in an early English translation, reads "Dialecticke otherwise called Logicke, is an arte which teachethe to dispute well." In the next centuries, logic as the art of rhetoric and disputation, became the domain of textbook writers and schoolteachers, a prerequisite for careers in law or the church. The major authors of modern philosophy and literature did not advance or even concern themselves with logic so conceived, and generally treated it with derision. John Milton thought it a subject in which "young Novices . . . [are] mockt and deluded . . . with ragged Notions and Babblements, while they expected worthy and delightful knowledge" (*On Education*).

This was an age also of discovery in the sciences and mathematics. The textbook logic 'of the schools' played no role in this. Francis Bacon claimed in the *Novum*

Organum that the "logic we now have" does not help us to *discover* new things, but "has done more to . . . fasten errors upon us, than to open the way to truth" (Book 1, Aphorism xii). He advocated instead rules of induction, a methodology of scientific investigation. In the *Discourse on Method* Descartes made similar remarks and John Locke, more radically, thought unaided natural reason to be more powerful than any logical methodology:

> Native rustic reason . . . is likelier to open a way to, and add to the common stock of mankind, rather than any scholastic proceeding. . . . For beaten tracks lead this sort of cattle . . . not where we ought to go, but where we have been. (*Essay Concerning Human Understanding*, 4.17.7)

The "cattle," poor drudges who taught logic to undergraduates, struck back by proposing to ban Locke's *Essay* from Oxford, since "there was a great decay of logical exercises . . . which could not be attributed to anything so much as the new philosophy, which was too much read" (Cranston 1957: 465ff).

Hume continued Locke's attack: "Our scholastic headpieces shew no . . . superiority above the mere vulgar in their reason and ability" (*Treatise on Human Nature*, 1.3.15). Denis Diderot's article on logic in the *Encyclopédie*, the most widely consulted reference work of the century, claimed that reasoning is a *natural* ability; to conduct logical inquiries is like "setting oneself the task of dissecting the human leg in order to learn how to walk" (*Encyclopédie*, Logique).

Gottfried Wilhelm Leibniz was the great exception to the logic bashing of the seventeenth and eighteenth centuries. He saw the general outline of what logic would much later become, but left only fragments of a 'universal characteristic' through which it would be become possible, he thought, to settle philosophical disputes through calculation. In the *New Essays Concerning Human Understanding*, a dialogue in which he responded to Locke, the latter's representative Philateles eventually admits "I regarded [logic] as a scholar's diversion, but I now see that, in the way you understand it, it is like a universal mathematics" (*New Essays* 4.17.9).

Traditionally, an exposition of logic followed the sequence: theory of terms or concepts, their combination into judgments, and the composition of syllogisms from judgments. This was now commonly prefaced by a discussion of the origin of concepts, as inherent in the mind or deriving from sensation and perception. In the end, many logic books contained more of these epistemological preliminaries than logic. There was, further, especially in England, an ongoing emphasis on logic as the art of disputation.

2 Kant and Whately

For the disordered progress of logic to even get on a path that would lead to modern logic, a reorientation and elimination of materials had first to occur. Neither Kant nor Whately contributed substantially to the formal development of logic, but they played a major role in this eliminative exercise.

Kant, unaware of earlier and since forgotten progress in logic, held that logic did not have to set aside any part of Aristotle's theory, but also had not taken a single step

forward, and "is to all appearances finished and complete" (*Critique of Pure Reason*, B viii). But in early lectures, he had shared the general disdain for the subject: "It took great effort to forget [Aristotle's] false propositions. . . . Locke's book *de intellectu* is the ground of all true *logica*" (Kant 1992: 16, 24).

By 1781, the time of the *Critique of Pure Reason*, he had changed his mind; Locke "speaks of the origin of concepts, but this really does not belong to logic" (Kant 1992: 439). While claiming earlier that the logician must know the human soul and cannot proceed without psychology, he now held that "pure logic derives nothing from psychology" (*Critique of Pure Reason* A54/B78).

Kant made two widely accepted distinctions: (1) he contrasted 'organon' and 'canon.' An *organon* (Kant uses the word in the sense Bacon gave it in the *Novum Organum*) attempts to codify methods of discovery. But "logic serves as a critique of the understanding, . . . not for creation." He sensibly held that there is no universal method of discovery, which rather requires a grasp of the special science that is to be advanced. But since logic must be general, attending only to form and not to content, it can only be a *canon*, a method of evaluation (*diiudicatio*). Methodological rules and theories of the origin and association of ideas, though intended as improvements of logic, are not even part of it. (2) Kant further divided logic into theoretical and practical. The latter, important but derivative, dealt with honing the skill of reasoning and disputation, while logic proper is a theoretical inquiry.

In the following decades nearly every German logic text was written by a student or follower of Kant. A contemporary could rightly observe that Kant gained a pervasive influence upon the history of logic. Regrettably, the overburden of psychology and epistemology in German logic treatises increased again in the course of the century, while its formal development stagnated, in part because of Kant's claim that it was a finished science.

Richard Whately (1787–1863) contributed to logic at the level of theory rather than formal detail. *Elements of Logic* (1827), an enormously popular response to the unrelenting criticism of the subject, was widely credited with reviving logic in England. Rather than fault logic for not doing what it cannot do (be an engine for discovery, or an "art of rightly employing the rational faculties"), it is better to focus on formal structures. In Whately's view, logic is an objective science like chemistry or mathematics, and its point (like that of the others) is the enunciation of principle apart from application. Faulting logic for not making people think better, "is as if one should object to the science of optics for not giving sight to the blind" (Whately 1827: 12).

Whately considered logic to be immediately about language, rather than vaguely conceived 'thought.' Unlike many of its loosely written predecessors, his book contains a formally adequate presentation of the categorical syllogism. A syllogism is a 'peculiar form of expression' into which any specific argument can be translated for testing validity. Properly understood, it is to an articulated argument as grammar is to language. The 'grammatical' analysis of any argument will lead to syllogistic form, just as the analytic devices of chemistry can be used on any compound and lead to basic elements. He also pushed an analogy with mathematics: just as the variables in mathematics stand for any number, so the letter variables used in stating syllogistic form stand for any term.

While Whately's theory is nearer to our present conception of logic, his critics faulted him for confining it within too narrow a scope. No longer would logic be the great sprawling subject that could be redefined almost at will, and many longed for that latitude. He prepared logic for innovation at the formal level.

3 Bernard Bolzano

At about the same time, Bernard Bolzano (1781–1848), "one of the greatest Logicians of all time" (Edmund Husserl), published his four-volume *Theory of Science* (*Wissenschaftslehre* (WL) 1837). It is the finest original contribution to logic since Aristotle, and a rich source for the history of the subject. In WL no formal calculus or system is developed; it is, rather, a treatise on the semantic concepts of logic. It was celebrated for its resolute avoidance of psychology in the development of these concepts.

Bolzano defines a spoken or written sentence as a speech act that is either true or false. Its *content*, that which is asserted or denied, is a proposition 'in itself,' explained as "any claim [*Aussage*] that something is or is not the case, regardless whether someone has put it into words, . . . or even has formulated it in thought" (WL § 19). He had little interest in the ontological status of these abstract propositions and meant to assert nothing deeper than we all do when we say that there *are* truths that are not yet known, or mathematical theorems not yet proved.

Any component of such a proposition not itself a proposition is a *Vorstellung* (idea or representation) in itself. The common sequence of first introducing terms or ideas and then propositions as compounds of them is here reversed. Bolzano noted that no one had successfully defined the type of combination of terms that generates a proposition. Several of the attempts he examined did not distinguish propositions from complex terms, 'the man is tall' from 'the tall man,' and others defined it in terms of 'acts of the mind,' contaminating logic with psychology (WL §§ 21–3).

Others (Hobbes, Condillac) identified propositions with equations, sometimes writing 'Caius is a man' as 'Caius = man.' Condillac and others maintained further that the principle on which all syllogisms rest is that two things equal to a third are equal to each other. But, Bolzano notes, while all equations are propositions, not all propositions are equations (WL §§ 23.20) and paid no further attention to this doctrine.

Identifying propositions with equations demanded further adjustments, the 'quantification of the predicate.' The German logician Ploucquet (1716–90) thought that in an affirmative proposition the predicate cannot be different from the subject. Hence he understood the proposition 'All lions are animals' as 'All lions are *some* animals.' In the same vein George Bentham (1800–84), in a commentary on Whately's book, symbolized 'All X are Y' as 'X *in toto* = Y *ex parte*' or 'All of X = Part of Y' (Bentham 1827: 133). The doctrine is now usually associated with the name of William Hamilton (1788–1856) who disingenuously claimed to have discovered it and gave it wide currency.

Back to Bolzano. He held that many propositions are not adequately expressed in common language. For instance, the proposition corresponding to the utterance

'I have a toothache' identifies speaker and time and is more adequately phrased as 'Neurath has a toothache at t.' Also, 'There is an A' is not, as it seems, about A's, but about the *idea* A; it means that this idea refers to an object (cf. Frege on quantifiers, below).

Bolzano's most important contribution was his definition of logical consequence using the mathematical technique of substitution on variables:

> Propositions M, N, O, . . . *follow from* propositions A, B, C, D, . . . with respect to the variable elements i, j, . . . if every set of ideas [*Vorstellungen*] whose substitution for i, j, . . . makes all of A, B, C, D, . . . true also makes M, N, O, . . . true. (WL § 155)

For example, 'a is larger than b, b is larger than c, therefore a is larger than c' is valid 'with respect to' the set of ideas 'a,' 'b,' 'c.'

It was generally understood, and often stated, that in a valid deductive argument, the conclusion follows *of necessity* from the premises (cf. Aristotle, *Prior Analytics* 24^b18). Bolzano's definition, closely akin to that given a century later by Alfred Tarski, was meant to explain the nature of this necessity.

If the variable elements i, j, . . . include *all* extralogical terms, then the consequence is said to be *logical*, as in a valid categorical syllogism. The unusual *triadic* construction of consequence also allows for enthymemes, or partly 'material' consequences, where only a *subset* of extralogical terms is varied. For example, in the argument 'All men are mortal, therefore Socrates is mortal,' any substitution on 'mortal' that makes the premise true makes the conclusion true: though not a logical consequence, it is valid with respect to 'mortal' (cf. George 1983).

Most logic texts of the period claimed, without supporting argument, that the so-called 'laws of thought' (identity, contradiction, and excluded middle) are the basic principles, the foundation on which all logic rests. While Bolzano agreed that these principles are true – his own logic was bivalent – his understanding of logical consequence showed him that nothing of interest followed from them. Logic, he maintained, *obeys* these laws, but they are not its *first principles* or, as we would now say, axioms (WL § 45).

He objected further to common attempts of grounding these laws in psychological necessities. Typically, the law of contradiction was supported by claims that a whole that is inconsistent cannot be united in a unity of thought, for example that round and quadrangular cannot be thought together because "one representation destroys the other." Against this Bolzano noted that we can, and often do, entertain inconsistent concepts. We can ask, for example, if there are regular dodecahedrons with hexagonal sides. But such a figure is just as impossible as a round square, only not obviously so. There are, in other words inconsistent ideas in themselves in Bolzano's abstract realm, and if entertained in a mind, they do not self-destruct.

Bolzano took mathematics to be a purely conceptual science, and disagreed with Kant's view that it was founded on intuition. Even in a diagram, what matters is what is general in it: the concept and not the intuition. His pioneering contributions to functional analysis entered the mainstream of mathematics in the nineteenth century, while his logical writings were appreciated only in the next.

4 John Stuart Mill

In his *System of Logic* (1843) Mill did not contribute to the development of logic as formal science, but like Bacon, attacked it. He claimed that formal principles, especially the syllogism, are a *petitio principii* since they can generate no new knowledge. One can know that the major premise 'All men are mortal' is true only if one knows the truth of the conclusion 'Socrates is mortal.' If that is still doubtful, the "same degree of uncertainty must hang over the premiss" (*System of Logic*, 2.3.2). When Archbishop Whately said that the object of reasoning is to "unfold the assertions wrapt up . . . in those with which we set out," Mill complained that he did not explain how a science like geometry can all be "wrapt up in a few definitions and axioms" (*System of Logic* 2.2.2). To explain that this is indeed the case had been a main objective of logic and mathematics before and especially after Mill. He thought it a project doomed to fail and claimed that the truths of geometry and arithmetic are empirically discovered by the simplest inductive method, that is *enumeration*. If a large number of instances of, and no exceptions to, A's being B is observed, it is concluded that *all* A's are B. Now if we have two pebbles and add another, then without exception we get three; neither do we ever observe two straight lines enclosing a space, forcing our minds to accept the truth of these and other mathematical propositions. Mill concluded that the "principles of number and geometry are duly and satisfactorily proved" by the inductive method of simple enumeration (*System of Logic* 3.21.2). Gottlob Frege later observed sarcastically that Mill never defined any number other than 3, nor did he illustrate the physical facts underlying 1 or 0, nor what "observed fact is asserted in the definition of the number 777846" (Frege 1884, § 7: 9).

Mill took the same empiricist and psychological approach to logic, whose "theoretic grounds are wholly borrowed from Psychology, and include as much of that science as is required to justify the rules of the [logical] art" (Mill 1865: 359). This holds in particular for the 'laws of thought,' which are grounded either in our psychological constitution, or in universal experience (1865: 381). Echoing earlier claims, he thought it impossible to entertain inconsistent concepts.

The *System of Logic* is best known for formulating rules for the discovery of causes, his famous 'canons': the methods of agreement, difference, residues, and concomitant variation. To illustrate the last: we take the moon to be the cause of tides, because the tides vary in phase with the position of the moon.

For a while, Mill's logic was the dominant text in logic and the philosophy of science in Britain, his eloquence creating much support to the view that logic is methodology and the art of discovery.

5 Boole, De Morgan, and Peirce

George Boole (1815–64) formulated his algebraic logic in conscious opposition to Mill's approach. Taking the mathematical analogy further than the loose suggestion of Whately, he sought to use algebra as a formal structure within which inferences could be perspicuously formulated. Logic should be a branch of mathematics, not of philoso-

phy; this would excise methodology, rhetoric, and epistemology. But logic can be a branch of mathematics only if the latter is not construed, as was common, as the science of quantity, but as the science of symbolic operations in general.

In his *Mathematical Analysis of Logic* of 1847 Boole introduced the notion of an 'elective symbol,' for example 'x', which represents the result of 'electing' the x's from the universe; it is the symbol for the resulting class. xy is the result of electing y's from the class x, hence the intersection of the two classes. It holds that $xy = yx$ and also that $xx = x$. $x + y$ is the union of the two classes, $x - y$ elects the x's that are not y. 0 is the empty class and 1 'the universe,' hence $1 - x$ is the class of non-x's. It follows that $1x = x$, $0x = 0$ and $x(y \pm z) = xy \pm xz$. A universal affirmative, 'All x are y' becomes '$x(1 - y) = 0$,' which says that the class of things that are x and not-y is empty. While this is an equation, it should be noted that it does not identify the subject with the predicate, as we find in earlier attempts of introducing algebraic notation into logic. A proof of the syllogism *Barbara* illustrates the algebraic method:

The syllogism	*Boolean computation*	*Comment*
All M are P	1. $m(1 - p) = 0$	the intersection of m and non-p = 0
All S are M	2. $s(1 - m) = 0$	the intersection of s and non-m = 0
	3. $m = mp$	algebraically from 1.
	4. $s = sm$	algebraically from 2.
	5. $s = smp$	mp for m in 4, licensed by 3.
	6. $s = sp$	s for sm in 5, licensed by 4.
	7. $s - sp = 0$	algebraically from 6.
All S are P	8. $s(1 - p) = 0$	algebraically from 7. QED.

The conclusion follows by 'multiplying' and 'adding,' specifically by maneuvering the middle term into a position where it can be eliminated. Syllogistics becomes part of the algebra of classes and thus an area of mathematics. If every argument can be formulated as a syllogism, then all of logic is a part of algebra.

For every analogy there is some disanalogy, and Boole's link between logic and algebra (as he was fully aware) was no exception. Some arithmetic functions (such as division, and even some cases of addition and subtraction) did not easily admit of logical interpretation. There are also difficulties in Boole's rendition of existential propositions: he wrote 'Some X are Y' as $v = xy$ where v stands for a class whose only defining condition is that it not be empty. But how can one define such a class? Also, his logic was still a logic of terms. The recognition of even so elementary a sentential function as negation came only later in the century.

Augustus De Morgan (1806–71) took a different path, retaining a closer connection with traditional syllogistic logic but moving the subject far beyond its traditional limits. When stripped of unnecessary restrictions, the syllogism would constitute an adequate basis for the representation of all modes of deductive reasoning. In his *Formal Logic* (1847), and in a later series of articles, he pushed the syllogistic structure so far that he called the status of the standard copula – 'is' – into question. If that term could be replaced by any term relating the other components in the statement, the reach of the

syllogism would be broadened: categorical statements would become relational statements.

De Morgan's more general interest in the logic of relations led him to examine inherently relational arguments, such as 'Every man is an animal. Therefore the head of a man is the head of an animal', which traditional syllogistic logic could not accommodate. He also introduced the concept of the 'universe of discourse,' still generally used, as a way of targeting statements to a class of objects under discussion, rather than the entire universe.

Charles Sanders Peirce's (1839–1914) theory of logic was once characterized as wider than anyone's. He was the first to consider himself not primarily a mathematician or philosopher, but a logician, filtering through the sieve of logic every topic he dealt with. On the formal level, he developed the logical lineage of Boole and De Morgan by refining the logic of relations, and devising more abstract systems of algebraic logic. He viewed it as a new and independent stage in the development of logic. The algebra of logic should be self-developed, and "arithmetic should spring out of logic instead of reverting to it." He developed a version of the modern quantifier, and of sentential functions. In both cases, it has been argued that, although Frege is often credited with introducing both notions into logic, it was Peirce and his students who were there first. Earlier he thought that logic is part of 'semiotics,' the theory of signs, their meaning and representation. Later he took it to *be* that theory, and while first taking logic to be descriptive, he later thought it to address cognitive norms.

Peirce introduced the memorable division of arguments into deduction, induction, and hypothesis, the last also called abduction and, more recently, 'inference to the best explanation.' He illustrated them as follows, using the then common terms 'Rule' for the major premise, 'Case' for the minor, and 'Result' for the conclusion of a categorical syllogism (Peirce 1931: 2.623):

Deduction:		*Rule*:	All the beans in this bag are white.
		Case:	These beans are from this bag.
	∴	*Result*:	These beans are white.
Induction:		*Case*:	These beans are from this bag.
		Result:	These beans are white.
	∴	*Rule*:	All the beans in this bag are white.
Hypothesis:		*Rule*:	All the beans in this bag are white.
	...	*Result*:	These beans are white.
	∴	*Case*:	These beans are from this bag.

In the last example the conclusion (the 'case') is accepted because on the available evidence it is the best explanation of why the beans are white.

6 Gottlob Frege

Frege (1848–1925) was a German mathematician and philosopher who set logic on a new path. He sought to connect logic and mathematics not by reducing logic to a form of algebra, but by deriving mathematics, specifically arithmetic, from the laws of logic.

He saw that a philosophy of language was a prerequisite for this and developed much of it in his *Conceptual Notation* (*Begriffsschrift*) of 1879. Like Bolzano, but more polemically, Frege opposed any attempt to import psychology into logic, repeatedly attacking Mill for this confusion. The meaning of sentences, for instance, is not explained by the mental states of speakers, but by investigating the language itself.

From the premise 'Castor is a sibling of Pollux,' two conclusions can be drawn by the *very same principle of inference*: 'Someone is a sibling of Castor' and 'Someone is a sibling of Pollux.' Traditionally, 'Castor' was construed as a different kind of sentential component than 'Pollux,' the first being the subject, the second lodged inside the predicate, so that the two conclusions followed by *different* principles. To correct this and other shortcomings of the traditional analysis of sentences, Frege replaced it with one built on *functions*.

In the equation $\sqrt{4} = |2|$ we distinguish *function* ('$\sqrt{}$'), *argument* ('4'), and *value* ('$|2|$'). The function is said to 'map' the argument to the value. '$\sqrt{(\ )}$' by itself is an 'unsaturated' expression that has a gap (shown as '( )') to be filled by an argument.

Frege construed sentences in the same way, '( ) is a planet' as a *sentential* function. If an argument, here called a *name* (an expression like 'Mercury,' 'Sirius' or 'the planet nearest the Sun') is inserted, a sentence results: 'Mercury is a planet' for example, or 'Sirius is a planet.' Sentential functions, like mathematical functions, can take more than one argument, as in '( ) is a sibling of { }', etc. In the Castor–Pollux example, the two arguments have the same status, and thus the single rule now called $\exists$-introduction, or existential generalization, legitimates both conclusions.

A function symbol refers to, or denotes, a *concept*, the name an *object*. Concepts and objects belong to distinct ontological categories. When a concept-term is an argument in a sentence, as in 'Red is a color,' the sentence is said to be on a 'higher level' than those whose arguments refer to objects.

As in the mathematical case, a sentential function maps its argument(s) to a value, but there are only two of these, the True and the False, the *truth values* of sentences. Thus the concept '( ) is a planet' maps 'Mercury' to Truth, 'Sirius' to Falsehood. In Frege's terms, Mercury 'falls under' the concept, Sirius does not. This is not just a more complicated way of saying that the one sentence is true, the other false. It is, rather, an analysis of what that *means*.

A further profound innovation was the quantifier. In mathematical texts quantification is usually tacit. For instance, '$x + 0 = x$' is true if it holds for every integer. If sentential connectives are brought into play, this no longer works: 'Fx,' if taken in the sense of a mathematical formula, will mean that everything is F, and its denial '$\neg$F(x)' that nothing is F, since it is true if $\neg$F(a) $\neg$F(b) etc. But 'Not everything is F' cannot be expressed in this way. For this, a special sign, a *quantifier* with a *scope* is needed. In current notation we can then distinguish between $\neg\forall$F(x) and $\forall x\neg$F(x). Frege took quantifiers to be higher level functions. The sentence 'There is a planet' is to be rendered as 'There is at least one thing such that [( ) is a planet].' The quantifier is here construed as a function that has another function as its argument.

Frege emphasized the importance of the 'deductive method.' Claims in a deductive science must be justified by a *proof*, which in his and all later logicians' view, is a sequence of propositions, each of which is either an assumption, or follows from previous members of the sequence by clearly articulated steps of deduction.

With this understanding of the structure of propositions, of quantification, and of the nature of a proof, *Begriffsschrift* develops an axiomatic system of sentential logic, based on two principles (actually two sets of axioms), one dealing with conditionals, the second with negation. The rule of *modus ponens* is employed to generate the first consistent and complete (as was shown much later) system of sentential logic.

A third principle, substitutivity, is introduced: if a = b, then F(a) is equivalent (as we now say) to F(b). With the introduction of a fourth principle, now 'universal instantiation' or ∀-elimination, a system of second order predicate logic is developed.

It seems that substitutivity fails in so-called *oblique* (or as we now say *opaque*) contexts. According to Frege, they are dependent clauses introduced by such words as 'to say,' 'to hear,' 'to believe,' 'to be convinced,' 'to conclude,' and the like. Now 'N believes that the morning star is a planet' may be true, while 'N believes that the evening star is a planet' false, even thought the two heavenly bodies are identical, apparently violating substitutivity. To save this principle, Frege introduced the important distinction between *sense* (*Sinn*) and *reference* (*Bedeutung*) (1892). 'The morning star' refers to the same object as 'The evening star' but they have a different sense. This is not the mental content associated with the signs, but their 'common meaning,' an objective entity determining the reference. Frege made the attractive assumption that in opaque contexts such expressions do not name an object, but their own sense, allowing substitution with any name of identical sense. Consider the sentence 'K believed that the evening star is a planet illuminated by the sun.' Here 'the evening star' may be replaced, *salva veritate* by 'the brightest star-like heavenly body in the evening sky,' provided the two expressions have the same sense for K. Similarly, sentences in oblique contexts have as their reference not their truth value, but the *thought* or sense they express. In this way, substitutivity, for Frege an incontrovertible principle of logic, can be made to work in opaque contexts.

Frege's main object was to show that arithmetic can be derived from logic alone, a project now called 'logicism.' For this he needed a definition of 'number' (in the sense of 'positive integer'), which he tried to provide in his famous monograph *The Foundations of Arithmetic* (1884).

> How, then, are numbers to be given to us, if we cannot have any ideas or intuitions of them? Since it is only in the context of a sentence that words have any meaning, our problem becomes this: To define the sense of a sentence in which a number word occurs. (Frege 1884: § 62)

This illustrates Frege's 'linguistic turn,' foreshadowing and inspiring twentieth century analytic philosophy: the question how we come to know numbers is transformed into one about the meaning of sentences in which number words occur. No further intuition or idea is needed or even possible. The quotation also states Frege's 'context principle': that only in the context of a sentence does a word have meaning. We have already seen that it makes no sense to ask for the meaning of 'red' if we do not know whether it occurs as function or as argument. Only in a sentence can we discern the grammatical role of its elements, and thus their meaning. As well, to determine the meaning of a word, one must know whether or not it occurs in an opaque context.

To give a definition of number, Frege used 'Hume's Principle': "When two numbers are so combined as that the one has always a unit answering to every unit of the other,

we pronounce them equal" (*Foundations* § 63, Hume, *Treatise of Human Nature* 1.3.1). Plainly, though true and obvious, this is not a principle of *logic*. He therefore tried to deduce it from what he took to be such a principle, the notorious Fifth Principle (in addition to the four of *Begriffsschrift*) which he introduced in his later work, *The Basic Laws of Arithmetic* of 1894. This is the so-called unrestricted comprehension (or abstraction) axiom, to the effect that any concept determines a set that has as its elements the objects that fall under the concept. While he expressed some uneasiness about the principle, he thought it a law of logic that one always has in mind when speaking about the extensions of concepts. Bertrand Russell discovered that a paradox (which bears his name) results from this. The concept '( ) is not a horse' determines the set of all objects not a horse, which includes that set itself. It is thus a set that has itself as an element. Consider now the set S determined by the predicate '( ) is not an element of itself'. If S is an element of itself, then it is not. But if S is *not* an element of itself, then it is, a contradiction from which in Frege's and all 'classical' systems of logic any conclusion whatever follows, rendering the system worthless. A postscript to the second volume of his *Basic Laws* (1903) states:

> Nothing can be more unwelcome to a scientific author than that, after the conclusion of his work, one of the foundations of his building is made to crumble. A letter from Mr. Bertrand Russell placed me in this situation just as the printing of this volume was almost finished. (Frege 1903)

Russell's discovery showed that the axioms of arithmetic (now commonly stated in the form Guiseppe Peano gave them) cannot be formally and consistently derived from Frege's principles (to say nothing of *all* of arithmetic, which cannot be so derived even *given* the axioms (Gödel 1931). But only in recent years has it been shown that these axioms follow from the principles of logic (minus the ill-fated Fifth) together with Hume's Principle. This is now called 'Frege's Theorem.'

7 The Austrian School

Franz Brentano (1838–1917), observed that all 'psychological phenomena' are targeted on some object: when we think, we think of *something*, when we value, we value *something*. These are *intentional objects* whose existence or nonexistence need not be an issue. Brentano shied away from allowing the contents of mental acts to have a form of being, taking this to be an unseemly Platonism. But his students Kasimir Twardowski (1866–1938) and Edmund Husserl (1859–1938) did just that, following Bolzano. Both distinguished *content* from *object*, with the object determined by the content. This is a distinction analogous to Frege's between sense and reference. Although they used figures of speech like the mind *grasping* its objects, they did not draw on psychological theories, and must be absolved of psychologism. Students of Twardowski formed the distinguished school of Polish logicians of the first part of the twentieth century. Of their many achievements we mention only Lesniewsky's (1886–1939) exploration of *mereology* of 1916, a subject that has only recently come to greater prominence. He distinguished the part–whole relation from that of class membership: an element of a

class is not a 'part' of it, though a subset is. Importantly, membership is not transitive: if s is an element of t, and t of u, then s is not an element of u, whereas a part of a part is a part of the whole.

Alexius Meinong (1853–1920), another of Brentano's students, inquired into the nature of intentional acts that lack existing objects and are 'beyond being and non-being.' When we think or speak of Hamlet, the content does not refer to a mental image, but to a 'subsisting' object that has lots of properties and satisfies certain identity conditions: the same person killed Polonius and loved Ophelia. Such talk does not lack logical structure. Meinong has more recently been credited with inspiring *free logic*: a logic without existence assumptions, and work in the logic of fiction. For a long time, however, he was known only in caricature through Bertrand Russell's famous article "On Denoting" (1905).

8 Bertrand Russell

In 1905 Russell published "On Denoting," his finest philosophical essay, as he thought. It became a milestone in the development of analytic philosophy. A distinction is here made between proper names and expressions like 'the so and so,' which he titled *definite descriptions*. In English grammar, 'The present king of France is bald' has the subject 'the present King of France' and the predicate 'bald.' But this is misleading. According to Russell, a proper understanding should distinguish three components of its meaning: (1) there is now at least one King in France (2) there is now at most one king in France and (3) every object satisfying (1) and (2) is bald. The sentence is true if all three conditions are satisfied, false if there is no king, if there is more than one king, or if there is a single non-bald king. But if this is what the sentence says, then 'the present king of France' is not part of its proper logical phrasing; a language constructed to strict logical standards will not contain a symbol for it. The misleading 'surface structure' of the sentence disguises its underlying logical structure.

Russell's conclusions are these: (1) Definite descriptions are not names, as Frege had thought; if they were, there would have to be objects to which they refer, leading to Meinong's ontological excesses. (2) Natural language structure and grammar are misleading and must be distinguished from the deeper logical structure. This was a landmark discovery, leading many philosophers to argue that metaphysical and even political convictions often gain their plausibility from deceptive natural language expressions. (3) Expressions like definite descriptions, but not only they, can be defined only in their contexts, by *definitions in use*. 'The present king of France' is not treated as a stand-alone expression and given an 'explicit' definition. Rather, the meaning and function of such expressions is conveyed through the analysis of the sentences in which they occur. (4) It is not necessary, as Meinong had thought, to populate the world with nonexisting, merely *subsisting* objects as the referents of definite descriptions. But there are problems. Some apparent names are disguised descriptions: 'Hamlet' is short for 'the Prince of Denmark'. Unfortunately, then, 'Hamlet loves Ophelia' is just as false as 'Hamlet loves Desdemona', since the prince is fictional. Rather than accept this one might wish to introduce a fictional, subsisting object to answer to the 'Hamlet'.

Despite his discovery of the paradox, Russell held that logicism could be made to work, if the comprehension axiom were restricted. He proposed several solutions, eventually the *theory of types*, fully articulated in the monumental *Principia Mathematica* authored by Russell and A. N. Whitehead (1910–13, three volumes, 1,000 pages), through which Frege's contributions entered the mainstream of logic. The preface states that "in all questions of logical analysis our chief debt is to Frege."

The theory of types stratifies expressions in a hierarchical order so that elements of a set are on a lower level than the set, making it impossible for a set to be a member of itself. A 'ramified' theory of types is introduced to solve as well the so-called semantic paradoxes, notably the liar paradox 'what I now say is false'. Russell and Whitehead were more successful in this than Philetas of Cos (third centuy BC) whose gravestone reads "I am Philetas; the lying argument has killed me and the night – long pondering," and more succinct than Chrysippus, who wrote 28 volumes on it (now lost: Bochenski 1961: 131). But their theory was burdened by the need to recognize a separate definition for truth at each type level and the inability to define a number as the set of all similar (two membered, three membered, etc.) sets. Strictly speaking, every level has different 2s, 3s, 4s, etc., and strictly speaking also different logical principles. They resolve this by using symbols that are 'systematically ambiguous' between types. Further complex adjustments were needed, the axioms of reducibility and choice, which are less than intuitively obvious as they should be for logicism really to succeed. It was also supposed that the vast remainder of mathematics could somehow be reduced to arithmetic, which seems ever more unlikely.

Russell and Whitehead did succeed, however, in deriving a significant portion of mathematics from their principles: a comprehensive theory of relations and order, Cantor's set theory, and a large portion of (finite and transfinite) arithmetic. *Principia* was also meant to be a kind of *Lingua Universalis*, a canonical language pure enough to permit construction of disciplined discourse on the skeleton it provided. Its symbolism was universally accepted, revisions to it addressing problems of readability rather than substance. Some philosophers went farther and proclaimed it the 'ideal language': either translate your claims into *Principia* notation or admit that they are meaningless.

We saw that several distinct areas of study were advanced under the name of logic. There was the view that logic investigates cognitive performance, or else scientific methodology and strategy of discovery, or that it is a branch of rhetoric. Setting aside all these as having contributed little to formal logic as now understood, there were still two distinct types of theory. Until *Principia*, and culminating in that work, the most prominent of them was proof theory, the development of mathematically rigorous *syntactical* procedures for deriving theorems from assumptions. Bolzano, representing the other type of theory, gave a *semantic* definition of logical consequence, which does not dwell on the process of derivation.

The most important development of logic after *Principia* was to bring these two strands together. In propositional logic, for instance, truth tables (introduced by Wittgenstein in 1922) allow a *semantic* test for the validity of formulas and proofs, a continuation of Bolzano's project. It was then proved that the *Principia* version of propositional logic is complete, that is to say that every semantically valid formula can be derived in it and that it is consistent, that is, that only such formulas (and hence no contradiction) can be derived. Later Kurt Gödel proved that first order predicate logic is

complete as well, but that higher order logic is not. Since the latter is needed to define arithmetic concepts, this spelled the end of the logicist project.

References

Ashworth, E. J. (1974) *Language and Logic in the Post-medieval Period*. Dordrecht and Boston, MA: Reidel. (Synthese Historical Library, vol. 12.)

Bentham, George (1827) *Outline of a New System of Logic*. London.

Bochenski, I. M. (1961) *A History of Formal Logic*. Notre Dame, IN: University of Notre Dame Press.

Bolzano, Bernard (1837) *Wissenschaftslehre*. Sulzbach: Seidel. (Cited from the English translation by Rolf George (1972) *Theory of Science*. Berkeley and Los Angeles: University of California Press; Oxford: Blackwell.)

Boole, George (1952) *Collected Logical Works*. La Salle, IL: Open Court.

Cranston, Maurice William (1957) *Locke*. New York: Macmillan.

Frege, Gottlob (1879) *Begriffsschrift*. Halle: Nebert. Translated as *Conceptual Notation*, with biography and introduction by Terrell Ward Bynum. Oxford: Clarendon Press, 1972.

Frege, Gottlob (1884) *Grundlagen der Arithmetik*. Translated as *The Foundations of Arithmetic: A Logico-mathematical Enquiry into the Concept of Number* by J. L. Austin. Oxford: Blackwell, 1959.

Frege, Gottlob (1892) Ueber Sinn und Bedeutung. *Zeitschrift fuer Philosophie und Philosophische Kritik*, 100, 25–50. Translated as *On Sense and Reference* in Michael Beaney (ed.), *The Frege Reader*. Oxford: Blackwell, 1997.

Frege, Gottlob (1894, 1903) *Grundgesetze der Arithmetik*. Jena: Pohle. Trans. as *The Basic Laws of Arithmetik* by Montgomery Furth. Berkeley and Los Angeles: University of California Press, 1967.

George, Rolf (1983) Bolzano's consequence, relevance and enthymemes. *Journal of Philosophical Logic*, 12, 299–318.

Gödel, Kurt (1931) On formally undecidable propositions in *Principia Mathematica*. In *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*, ed. Jean Van Heijenoort. Cambridge, MA: Harvard University Press, 1967.

Kant, Immanuel (1992) *Lectures on Logic*. Translated and edited by J. Michael Young. Cambridge: Cambridge University Press.

Kneale, William and Martha (1962) *The Development of Logic*. Oxford: Clarendon.

Meinong, Alexius (1902) *Ueber Annahmen*. Reprinted in *Gesamtausgabe*, ed. Rudolf Haller and Rudolf Kindlinger. Graz: Akademische Druck und Verlagsanstalt, 1968–78, vol. IV. Trans. as *On Assumptions*, by James Heanue. Berkeley and Los Angeles: University of California Press, 1983.

Mill, John Stuart (1843) *A System of Logic*. New edition by J. M. Robson in *Collected Works*. Toronto: University of Toronto Press; London: Routledge, 1963–, vols. VII and VIII.

Mill, John Stuart (1865) *An Examination of Sir William Hamilton's Philosophy*. New edition by J. M. Robson in *Collected Works*. Toronto: University of Toronto Press; London: Routledge, 1963–, vol. IX.

Peirce, Charles Sanders (1931) *Collected Papers*, ed. Charles Hartshorne and Paul Weiss. 3rd edn. Cambridge, MA: Harvard University Press, 1974. Vol. II: *Elements of Logic*.

Russell, Bertrand (1905) On denoting. *Mind*. Reprinted in *Logic and Knowledge*, ed. Robert Charles Marsh. London: George Allen & Unwin, 1956.

Whately, R. (1827) *Elements of Logic*. New edition ed. Paola Dessì. Bologna: Editrice CLUEB, 1988.

Zalta, Edward N. (1999) Frege. *The Stanford Encyclopaedia of Philosophy*, http//plato.stanford.edu.

Part II

SYMBOLIC LOGIC AND ORDINARY LANGUAGE

This page intentionally left blank

4

Language, Logic, and Form

KENT BACH

Despite their diversity, natural languages have many fundamental features in common. From the perspective of Universal Grammar (see, e.g., Chomsky 1986), such languages as English, Navajo, Japanese, Swahili, and Turkish are far more similar to one another than they are to the formal languages of logic. Most obviously, natural language expressions fall into lexical categories (parts of speech) that do not correspond to the categories of logical notation, and some of them have affixes, including prefixes, suffixes, and markings for tense, aspect, number, gender, and case. Moreover, logical formalisms have features that languages lack, such as the overt presence of variables and the use of parentheses to set off constituents. The conditions on well-formed formulas in logic (WFFs) are far simpler than those on well-formed (grammatical) sentences of natural languages, and the rules for interpreting WFFs are far simpler than those for interpreting grammatical sentences. Compare any book on syntax and any book on formal logic and you will find many further differences between natural languages and formal languages. There are too many approaches to the syntax of natural languages to document these differences in detail. Fortunately, we will be able to discuss particular examples and some general issues without assuming any particular syntactic framework.

We will focus mainly on logically significant expressions (in English), such as 'and,' 'or,' 'if,' 'some,' and 'all' and consider to what extent their semantics is captured by the logical behavior of their formal counterparts, '&' (or '$\wedge$'), '$\vee$,' '$\supset$' (or '$\rightarrow$'), '$\exists$,' and '$\forall$.' Rendering 'if' as the material conditional '$\supset$' is notoriously problematic, but, as we shall see, there are problems with the others as well. In many cases, however, the problems are more apparent than real. To see this, we will need to take into account the fact that there is a pragmatic dimension to natural language.

Sentences of English, as opposed to (interpreted) formulas of logic, not only have semantic contents but also are produced and perceived by speakers (or writers) and listeners (or readers) in concrete communicative contexts. To be sure, logical formulas are also produced and perceived by particular people, but nothing hangs on the fact that they are so produced and perceived. In ordinary speech (or writing), it is not just what a sentence means but the fact that someone utters (or writes) it plays a role in determining what its utterance conveys (Bach 1999a). So, for example, there is a difference between what is likely to be conveyed by utterances of (1) and (2),

(1) Abe felt lousy and ate some chicken soup.
(2) Abe ate some chicken soup and felt lousy.

and the difference is due to the order of the conjuncts. Yet 'and' is standardly symbolized by the conjunction '&,' and in logic the order of conjuncts doesn't matter. However, it is arguable that (1) and (2) have the same semantic content and that it is the fact that the conjuncts are *uttered* in a certain order, not the meaning of 'and,' that explains the difference in how the utterances are likely to be taken.

One recurrent question in our discussion is to what extent rendering natural language sentences into logical notation exhibits the logical forms of those sentences. In addressing this question, we will need to observe a distinction that is often overlooked. It is one thing for a sentence to be rendered into a logical formula and quite another for the sentence itself to have a certain logical form. When philosophers refer to the logical form of a sentence, often all they mean is the form of the (interpreted) logical or semi-logical formula used to paraphrase it, often for some ulterior philosophical purpose, for example to avoid any undesirable ontological commitments (see Quine 1960) or to reveal the supposedly true structure of the proposition it expresses. A logical paraphrase of a natural language sentence does not necessarily reveal inherent properties of the sentence itself. However, as linguists construe logical form, it is a level of syntactic structure, the level that provides the input to semantic interpretation. The logical form of a sentence is a property of the sentence itself, not just of the proposition it expresses or of the formula used to symbolize it.

The difference is evident if we consider a couple of simple sentences and how they are standardly symbolized:

(3) There are quarks.
(4) Some quarks are strange.

In first-order predicate logic (3) and (4) would be symbolized as (3_{PL}) and (4_{PL}):

(3_{PL}) $(\exists x)\, Qx$
(4_{PL}) $(\exists x)\, (Qx \,\&\, Sx)$

Whereas (3) expresses an existential proposition and (4) apparently does not, both sentences are symbolized by means of formulas containing an existential quantifier. Not only that, there appears to be nothing in (4) corresponding to the conjunction ('&') in (4_{PL}). These discrepancies do not, however, deter many philosophers and logic texts from proclaiming that a formula like (4_{PL}) captures the logical form of a sentence like (3). Obviously they are not referring to logical form as a level of syntactic structure.

1 Sentential Connectives

In the propositional calculus, the words 'and' and 'or' are commonly rendered as truth-functional, binary sentential connectives. 'S_1 and S_2' is symbolized as 'p & q,' true iff 'p'

is true and 'q' is true, and 'S_1 or S_2' as 'p $\vee$ q,' true iff 'p' is true or 'q' is true. There are two obvious difficulties with these renderings. For one thing, there is no limit to the number of clauses that 'and' and 'or' can connect (not that their usual truth-functional analysis cannot be extended accordingly). Moreover, 'and' and 'or' do not function exclusively as sentential connectives, for example as in (5) and (6):

(5) Laurel and Hardy lifted a piano.
(6) Abe wants lamb or halibut.

Clearly those sentences are not elliptical versions of these:

(5+) Laurel lifted a piano and Hardy lifted a piano.
(6+) Abe wants lamb or Abe wants halibut.

So the use of 'and' and 'or' as subsentential connectives cannot be reduced to their use as sentential connectives. It could be replied that this difficulty poses no problem for the standard truth-functional analysis of 'and' and 'or' when used as sentential connectives. However, such a reply implausibly suggests that these terms are ambiguous, with one meaning when functioning as sentential connectives and another meaning when connecting words or phrases. These connectives seem to have unitary meanings regardless of what they connect.

A further difficulty, perhaps of marginal significance, is that the truth-functional analysis of 'and' and 'or' does not seem to handle sentences like 'Give me your money and I won't hurt you' and 'Your money or your life,' or, more domestically:

(7) Mow the lawn and I'll double your allowance.
(8) Mow the lawn or you won't get your allowance.

It might seem that these sentences involve a promissory use of 'and' and a threatening use of 'or.' But that's not accurate, because there are similar cases that do not involve promises or threats:

(9) George Jr. mows the lawn and George Sr. will double his allowance.
(10) George Jr. mows the lawn or he won't get his allowance.

Here the speaker is just a bystander. The 'and' in (9) seems to have the force of a conditional, that is 'If George Jr. mows the lawn, George Sr. will double his allowance.' This makes the 'and' in (9) weaker than the ordinary 'and.' And the 'or' in (10) has the force of a conditional with the antecedent negated, that is 'if George Jr. does not mow the lawn, he won't get his allowance.'

If we can put these difficulties aside, although they may not be as superficial as they seem, the standard truth-functional analysis of 'and' and 'or' does seem plausible. Grice's (1989: ch. 2) theory of conversational implicature inspires the hypothesis that any counterintuitive features of this analysis can be explained away pragmatically.

'And'

As observed by Strawson (1952: 81) and many others since, the order of conjuncts seems to matter, even though the logical '&' is commutative: (p&q) ≡ (q&p). Although there is no significant difference between (11a) and (11b),

(11) a. Uzbekistan is in Asia and Uruguay is in South America.
b. Uruguay is in South America and Uzbekistan is in Asia.

there does seem to be a difference between (12a) and (12b):

(12) a. Carly got married and got pregnant.
b. Carly got pregnant and got married.

and between (13a) and (13b):

(13) a. Henry had sex and got infected.
b. Henry got infected and had sex.

However, it is arguable that any suggestion of temporal order or even causal connection, as in (13a), is not a part of the literal content of the sentence but is merely implicit in its utterance (Levinson 2000: 122–7). One strong indication of this is that such a suggestion may be explicitly canceled (Grice 1989: 39). One could utter any of the sentences in (12) or (13) and continue, 'but not in that order' without contradicting or taking back what one has just said. One would be merely canceling any suggestion, due to the order of presentation, that the two events occurred in that order.

However, it has been argued that passing Grice's cancelability test does not suffice to show the differences between the (a) and (b) sentences above is a not a matter of linguistic meaning. Cohen (1971) appealed to the fact that the difference is preserved when the conjunctions are embedded in the antecedent of a conditional:

(14) a. If Carly got married and got pregnant, her mother was thrilled.
b. If Carly got pregnant and got married, her mother was relieved.
(15) a. If Henry had sex and got infected, he needs a doctor.
b. If Henry got infected and had sex, he needs a lawyer.

Also, the difference is apparent when the two conjunctions are combined, as here:

(16) I'd rather get married and get pregnant than get pregnant and get married.
(17) It's better to have sex and get infected than to get infected and have sex.

However, these examples do not show that the relevant differences are a matter of linguistic meaning. A simpler hypothesis, one that does not ascribe multiple meanings to 'and,' is that these examples, like the simpler ones in (12) and (13), are instances of the widespread phenomenon of conversational impliciture (Bach 1994), as opposed to Grice's implic-a-ture, in which what the speaker means is an implicitly qualified version

of what he says. Here are versions of (14a) and (16) with the implicit 'then' made explicit:

(14a+) If Carly got married and *then* got pregnant, her mother was thrilled.
(16+) I'd rather get married and *then* get pregnant than get pregnant and *then* get married.

(14a) and (16) are likely to be uttered as if they included an implicit 'then,' and are likely to be taken as such. The speaker is exploiting Grice's (1989: 28) maxim of manner. Notice that if the contrasts in the pairs of conjunctions were a matter of linguistic meaning, then 'and' (and sentences containing it) would be semantically ambiguous. There would be a sequential 'and,' a causal 'and,' and a merely truth-functional 'and,' as in (11). Each of our examples would be multiply ambiguous and would require disambiguation. (13b), for example, would have a causal reading, even if that is not the one likely to be intended. An additional meaning of 'and' would have to be posited to account for cases like (18):

(18) He was five minutes late and he got fired?

where what is questioned is only the second conjunct. The pragmatic approach, which assimilates these cases to the general phenomenon of meaning something more specific than what one's words mean, treats 'and' as unambiguously truth-functional and supposes that speakers intend, and hearers take them to intend, an implicit 'then' or 'as a result' or something else, as the case may be, to be understood along with what is said explicitly.

'Or'

Even though it is often supposed that there is both an inclusive 'or' and an exclusive 'or' in English, in the propositional calculus 'or' is symbolized as the inclusive '$\vee$.' A disjunction is true just in case at least one of its disjuncts is true. Of course, if there were an exclusive 'or' in English, it would also be truth-functional – an exclusive disjunction is true just in case exactly one of its disjuncts is true – but the simpler hypothesis is that the English 'or' is unambiguously inclusive, like '$\vee$.' But does this comport with the following examples?

(19) Sam is in Cincinnati or he's is Toledo.
(20) Sam is in Cincinnati or Sally (his wife) will hire a lawyer.

An utterance of (19) is likely to be taken as exclusive. However, this is not a consequence of the presence of an exclusive 'or' but of the fact that one can't be in two places at once. Also, it might seem that there is an epistemic aspect to 'or,' for in uttering (19), the speaker is implying that she doesn't know whether Sam is in Cincinnati or Toledo. Surely, though, this implication is not due to the meaning of the word 'or' but rather to the presumption that the speaker is supplying as much relevant and reliable information as she has (see Grice 1989: ch. 2). The speaker wouldn't be contradicting herself

if, preferring not to reveal Sam's exact whereabouts, she added, "I know where he is, but I can't tell you."

The case of (20) requires a different story. Here the order of the disjuncts matters, since an utterance of "Sally will hire a lawyer or Sam is in Cincinnati" would not be taken in the way that (20) is likely to be. Because the disjuncts in (20) are ostensibly unrelated, its utterance would be hard to explain unless they are actually connected somehow. In a suitable context, an utterance of (20) would likely be taken as if it contained 'else' after 'or,' that is as a conditional of sorts. That is, the speaker means that if Sam is *not* in Cincinnati, Sally will hire a lawyer, and might be implicating further that the reason Sally will hire a lawyer is that she suspects Sam is really seeing his girlfriend in Toledo. The reason that order matters in this case is not that 'or' does not mean inclusive disjunction but that in (20) it is intended as elliptical for 'or else,' which is not symmetrical.

One indication that 'or' is univocally inclusive is that it is never contradictory to add 'but not both' to the utterance of a disjunction, as in (21),

(21) You can have cake or cookies but not both.

However, it might be argued that 'or' cannot be inclusive, or at least not exclusively so, since there seems to be nothing redundant in saying,

(22) Max went to the store or the library, or perhaps both.

The obvious reply is that adding 'or perhaps both' serves to cancel any implication on the part of the speaker that only one of the disjuncts holds and to raise to salience the possibility that both hold.

'If'

Since the literature on conditionals is huge, they cannot be discussed in detail here. But we must reckon with the fact – nothing is more puzzling to beginning logic students than this – that on the rendering of 'if S_1, then S_2' as '$p \supset q$,' a conditional is true just in case its antecedent is false or its consequent is true. This means that if the antecedent is false, it doesn't matter whether the consequent is true or false, and if the consequent is true, it doesn't matter whether the antecedent is true or false. Thus, both (23) and (24) count as true,

(23) If Madonna is a virgin, she has no children.
(24) If Madonna is a virgin, she has children.

and so do both (25) and (26),

(25) If Madonna is married, she has children.
(26) If Madonna is not married, she has children.

Apparently the basic problem with the material conditional analysis of 'if' sentences is that it imposes no constraint on the relationship between the proposition expressed by

the antecedent and the one expressed by the consequent. On this analysis (27)–(30) are as true as (23)–(26),

(27) If Madonna is a virgin, she is a multi-millionaire.
(28) If Madonna is a virgin, she is not a multi-millionaire.
(29) If Madonna is married, she is a pop singer.
(30) If Madonna is not married, she is a pop singer.

This might suggest that 'if' sentences are not truth-functional (indeed, Edgington (1991) has argued that they are not even truth-valued).

However, it is arguable that the connection (what Strawson (1986) calls a "ground-consequent relation") between antecedent and consequent is not part of the conventional meaning of an 'if' sentence. Perhaps the implication of such a connection can be explained pragmatically. So suppose that an 'if' sentence is equivalent to a material conditional, '$p \supset q$,' true just in case either its antecedent is false or its consequent is true. It is thus equivalent to '$\neg p \vee q$.' Now as Strawson sketches the story, one would not utter a conditional if one could categorically assert the consequent or the negation of the antecedent. That would violate the presumption, to put it roughly, that a speaker makes as strong a relevantly informative statement as he has a basis for making. As we saw above, it would be misleading to assert a disjunction if you are in a position to assert a disjunct, unless you have independent reason for withholding it. In the present case, you wouldn't assert the equivalent of '$\neg p \vee q$' if you could either assert '$\neg p$' or assert 'q.' But then why assert the equivalent of '$\neg p \vee q$'? The only evident reason for this is that you're in a position to deny '$(p \,\&\, \neg q)$' – '$\neg(p \,\&\, \neg q)$' is equivalent to '$\neg p \vee q$' – on grounds that are independent of reasons for either asserting '$\neg p$' or asserting 'q.' And such grounds would involve a ground-consequent relation. So, for example, you wouldn't utter (23) if you could assert that Madonna is not a virgin or that she has no children. However, in the case of (31),

(31) If Madonna has many more children, she will retire by 2005.

where you're not in a position to deny the antecedent or categorically assert the consequent, you would assert it to indicate a ground-consequent relation between them.

Although Strawson's account is plausible so far as it goes, sometimes we have occasion for asserting a conditional without implicating any ground-consequent relation between its antecedent and consequent. Indeed, we may implicate the absence of such a relation. This happens, for example, when one conditional is asserted and then another is asserted with a contrary antecedent and the same consequent, as in the following dialogue:

Guest: The TV isn't working.
Host: If the TV isn't plugged in, it doesn't work.
Guest: The TV *is* plugged in.
Host: If the TV is plugged in, it doesn't work.

Clearly the host's second utterance does not implicate any ground-consequent relation. As the propositional calculus predicts, the host's two statements together entail that the TV doesn't work, period.

One last bit of support for the truth-functional account of conditionals comes from cases like "If you can lift that, I'm a monkey's uncle" or (32),

(32) If Saddam Hussein wins the Albert Schweitzer Humanitarian Award, Dr. Dre will win the Nobel Prize for medicine.

In such cases, the antecedent is obviously false, and the speaker is exploiting this fact. There is no entailment of a ground-consequent connection between the antecedent and consequent, and the speaker is not implicating any. Rather, he is implicating that the consequent is false, indeed preposterous.

One last point about conditionals is that sometimes they are used as if they were biconditionals (symbolized by '≡' rather than '⊃'). For example, it might be argued that 'if' can sometimes mean 'if and only if,' as in (33),

(33) If Harry works hard, he'll get promoted.

where there seems to be an implication that if Harry doesn't work hard, he won't get promoted, that is that he'll get promoted only if he works hard.

We have not addressed the case of so-called subjunctive or counterfactual conditionals (I say 'so-called' because, as Dudman (e.g. 1991) has repeatedly pointed out, they need not be either subjunctive or counterfactual). The conditions on their truth is a complex and controversial question (see the relevant essays in Jackson 1991), but clearly the following conditionals differ in content:

(34) a. If Oswald didn't shoot Kennedy, someone else did.
b. If Oswald hadn't shot Kennedy, someone else would have.

Whatever the explanation of the difference, presumably it is not due to any ambiguity in 'if' but to something else.

There are a great many sentential connectives that we will not consider, such as 'after,' 'although,' 'because,' 'before,' 'but,' 'consequently,' 'despite the fact that,' 'even though,' 'however,' 'inasmuch as,' 'nevertheless,' 'provided that,' 'since,' 'so,' 'therefore,' 'unless,' and 'until.' We cannot take them up here, but it is interesting to consider which ones are truth-functional and which are not.

2 Quantifiers and Quantified Noun Phrases

Only the existential and universal quantifiers are included in standard first-order predicate logic. The existential quantifier is commonly used to capture the logical properties of 'some' and 'a' and the universal quantifier those of 'every,' 'each,' and 'all' ('any' is a tricky case because it seems to function sometimes as a universal and sometimes as an existential quantifier). But there are differences between 'some' and 'a' and between

'every,' 'each,' and 'all' that are not captured by their formal symbolizations. For example, only 'some' and 'all' can combine with plural nouns. Also, 'some' but not 'a' can be used with mass terms, as in 'Max drank some milk' as opposed to 'Max drank a milk' ('Max drank a beer' is all right, but only because the mass term 'beer' is used here as a count noun, as in 'Max drank three beers'). But these differences are superficial as compared with two deeper difficulties with the symbolization of quantifiers in first-order predicate logic.

One difficulty was mentioned at the outset. A simple sentence like (4) is standardly symbolized with existential quantification, as in (4_{PL}):

(4) Some quarks are strange.
(4_{PL}) $(\exists x)(Qx \,\&\, Sx)$

The difficulty is that there is nothing in (4) corresponding to the connnective '&' in (4_{PL}) or to the two open sentences it conjoins. There is no constituent of (4_{PL}) that corresponds to the quantified noun phrase 'some quarks' in (4). The situation with universal quantification is similar, illustrated by the symbolization of a sentence like (35) as (35_{PL}):

(35) All fish are garish.
(35_{PL}) $(\forall x)(Fx \supset Gx)$

In fact, not only is there is nothing in (35) that corresponds to the connective '$\supset$' in (35_{PL}), but (35_{PL}) is true if there are no Fs, as with (36),

(36) All four-legged fish are gymnasts.

This is not a difficulty only if (36) is equivalent to (37),

(37) Anything that is a four-legged fish is a gymnast.

and intuitions differ on that. In standard predicate logic, universal sentences of the form 'All Fs are G' are true if there are no Fs, and, according to Russell's theory of descriptions, sentences of the form 'The F is G' are true if there is no unique F. Of course, one would not assert such a sentence if one believed there to be no F or no unique F, but logic need not concern itself with that. In any case, clearly the forms of (4_{PL}) and (35_{PL}) do not correspond to the grammatical forms of the sentences they symbolize.

These discrepancies might be thought to reveal a problem with English rather than with predicate logic. Indeed, Russell regarded it as a virtue of his theory of descriptions that the structure of the formal rendering of a description sentence does not mirror that of the sentence it symbolizes. A sentence like (38),

(38) The director of *Star Wars* is rich.

should not be symbolized with 'Rd,' where 'R' stands for 'is rich' and 'd' stands for 'the director of *Star Wars*,' but with the more complex but logically revealing (38_{PL}):

(38_{PL}) $(\exists x)(Dx \,\&\, (y)(Dy \supset (y = x) \,\&\, Rx)$

(This is not Russell's notation but one of several ways in modern predicate logic to render his analysis.) Whereas (38) has 'the director of *Star Wars*' as its grammatical subject and 'is rich' as its grammatical predicate, it is revealed by logical analysis not to be of subject-predicate logical form. Hence the grammatical form of a sentence like (38) is "misleading as to logical form," as Russell was paraphrased by Strawson (1952: 51). The definite description 'the director of *Star Wars*' does not correspond to any constituent of the proposition expressed by (38). Definite descriptions "disappear on analysis." The contribution they make to the propositions in which they occur is a complex quantificational structure of the sort contained in (38_{PL}).

Although Russell's theory of descriptions is often taken as the paradigm of how grammatical form can be misleading as to logical form, as we have seen, sentences like (4) and (35), when symbolized in the standard ways, seem to be examples of the same thing. However, it is arguable that this alleged misleadingness is entirely an artifact of the notation being used. Indeed, as Barwise and Cooper (1981) have shown, the notation of first-order logic is not adequate for symbolizing such quantificational expressions as 'most,' 'many,' 'several,' 'few.' And there are numerical quantifiers to contend with, like 'eleven' and 'a dozen,' and more complex quantificational expressions, such as 'all but one,' 'three or four,' 'fewer than ten,' 'between ten and twenty,' 'at most ninety-nine,' and 'infinitely many.' The notation of restricted quantification can uniformly handle this rich diversity of locutions (see Neale (1990: 41ff.) for a clear explanation of how restricted quantification works). Not only that, it does so in a way that respects the structural integrity of the quantified noun phrases that it symbolizes. So, for example, the sentences in (39) may be symbolized by the corresponding formulas in (39_{RQ}), where for simplicity the predicates are symbolized with predicate letters:

(39) a. Most baseball players like golf.
b. Many philosophers like wine.
c. Few pro-lifers support gun control.
d. Eleven jurors voted guilty.

(39_{RQ}) a. [Most x: Bx] Gx
b. [Many x: Px] Wx
c. [Few x: Lx] Cx
d. [Eleven x: Jx] Gx

Restricted quantification notation thus avoids first-order logic's "notorious mismatch between the syntax of noun phrases of natural languages like English and their usual representations in traditional predicate logic" (Barwise and Cooper 1981: 165), and instead symbolizes constituents with constituents, thus facilitating a more straightforward compositional semantics. In particular, it does not separate quantifiers from their nominal complements. As a result, it removes any suggestion that grammatical form is misleading as to logical form. This holds even for definite descriptions, which do not disappear on the restricted quantification analysis.

The terms 'only' and 'even' pose some special problems. What propositions are expressed by (40) and (41)?

(40) Only Ernie eats turnips.
(41) Even Ernie eats turnips.

(41) seems to entail that Ernie is not the sole individual who eats turnips, even though there is no explicit indication who the other people are, much less an explicit quantification over the group in question. (41) seems to say, in effect, that Ernie eats turnips and that, of an unspecified group of people who eat turnips, Ernie is the least likely to do so. Exactly what it says is a matter of some debate (see, e.g., Kay 1990, and Francescotti 1995), but even if the paraphrase is correct, it is not obvious how to render that into the notation of first-order logic or even restricted quantification. Even if it could be so rendered, such a symbolization would have to contain structure that is not present, or at least not evident, in (41) itself.

Let us focus on the somewhat simpler case of 'only.' Offhand, (40) seems to express the proposition that Ernie and no one (in the contextually relevant group) other than Ernie eats turnips. In first-order predicate logic, this can be rendered as (40_{PL}):

(40_{PL}) $Te \,\&\, (\forall x)(x \neq e \supset \neg Tx)$

Like (40), (40_{PL}) entails both that Ernie eats turnips and that no one else does. A logically equivalent but distinct rendering of (40) is ($40'_{PL}$),

($40'_{PL}$) $Te \,\&\, (\forall x)(Tx \supset x = e)$

which says that Ernie eats turnips and anyone who does is Ernie. There has been a debate in the literature about whether this is entirely accurate (see Horn (1996) and references there), but the relevant question here concerns the relationship between (40) and the first-order formula used to symbolize it. Both (40_{PL}) and ($40'_{PL}$) contain elements of structure that are not present, at least not obviously so, in (40). This can be avoided somewhat if we render the second conjuncts of (40_{PL}) and ($40'_{PL}$) in restricted quantificational notation. Then (40_{PL}) becomes (40_{RQ}) and ($40'_{PL}$) becomes ($40'_{RQ}$).

(40_{RQ}) $Te \,\&\, [\text{every } x\text{: } x \neq e]\ \neg Tx$
($40'_{RQ}$) $Te \,\&\, [\text{every } x\text{: } Tx]\ x = e$

But still there are elements not ostensibly present in (40): conjunction, a universal quantifier, an identity sign, and, in the case of (40_{RQ}), a negation sign. We can eliminate most of these elements and the structure they require if we treat 'only' as itself a quantifier,

($40''_{RQ}$) $[\text{Only } x\text{: } x = e]\ Tx$

Here the proper name 'Ernie' is treated as a nominal that combines (together with a variable and an identity sign) with a quantifier to yield a quantified noun phrase.

There is a further problem posed by 'only.' Consider (42):

(42) Only Bernie loves his mother.

What is the property that no one else (in the contextually relevant group) possesses? On one reading of (42), it is the property of loving Bernie's mother; on another, it is the property of loving some contextually relevant male's mother). These two readings may be represented with the help of indices.

(42) a. Only Bernie$_1$ loves his$_1$ mother.
b. Only Bernie$_1$ loves his$_2$ mother.

But there is a third, reflexive reading of (42), on which the property in question is that of loving one's own mother (for discussion of different approaches to reflexivity, see Salmon 1992). There is no obvious way to use indices to reflect that (the indices in (42) cover the options), but restricted quantificational notation can do the trick:

(42) c. [Only x: x = b] (x loves x's mother).

Notice that, as in ($40_{RQ'}$) above, 'only' is treated here as a quantifier and the proper name as a nominal that combines with the quantifier to yield a quantified noun phrase.

3 Proper Names and Individual Constants

It is customary in logic to use individual constants to symbolize proper names, and to assign only one such constant to a given individual. Doing so obliterates semantic differences between co-referring proper names. It implicitly treats names as essentially Millian, as contributing only their bearers to the semantic contents of sentences in which they occur. From a logical point of view there is no difference between the propositions expressed by (43) and (44),

(43) Queen Noor skis.
(44) Lisa Halaby skis.

since Queen Noor is Lisa Halaby. They could be symbolized as 'Sn' and 'Sh' respectively, but this would not exhibit any semantic difference, given that n = h. It might seem that there is no such difference, insofar as co-referring names may be substituted for one another without affecting truth value, but such substitution does seem to affect propositional content. As Frege (1892) pointed out, a sentence like (45) seems to be informative in a way that (46) is not:

(45) Queen Noor is Lisa Halaby.
(46) Queen Noor is Queen Noor.

Millianism, which provides the rationale for symbolizing proper names as individual constants, must deny that there is any difference in propositional content between (45) and (46), even if it concedes a cognitive, but non-semantic, difference between

them, or between (43) and (44). However, replacing a name with a co-referring one does seem to affect both truth value and propositional content in the context of attitude ascriptions:

(47) Prince Rainier believes that Queen Noor skis.
(48) Prince Rainier believes that Lisa Halaby skis.

It seems that (47) might be true while (48) is false and that they have different contents, since they ascribe to Prince Rainier belief in two different things. Millians must reject this, and explain away the appearance of substitution failure as based on some sort of pragmatic or psychological confusion (see Salmon 1986; Braun 1998; Soames 2001), but many philosophers find such explanations, however ingenious, to be implausible (see Bach 2000).

A further problem for Millianism is posed by existential sentences containing proper names. If the contribution that a proper name makes to sentences in which it occurs is its referent (if it has one) and nothing else, then how are sentences like the following to be understood or symbolized?

(49) Bigfoot does not exist.
(50) Sting exists.

As first remarked by Kant, existence is not a property and 'exists' is not a predicate. Bigfoot is not a creature which lacks a property, existence, that Sting possesses. That is why sentences like (49) and (50) are ordinarily not symbolized as '¬Eb' and 'Ep.' But what is the alternative? In first-order predicate logic, there is no straightforward way to symbolize such sentences, since 'exists' is symbolized by the existential quantifier, not by a predicate, and combines with open sentences, not individual constants. A common trick for symbolizing sentences like (49) and (50) is with identity, as in (49_{FL}) and (50_{FL}):

(49_{FL}) $\neg(\exists x)\ x = b$
(50_{FL}) $(\exists x)\ x = s$

However, (49) and (50) do not seem to contain anything corresponding to the variable-binding existential quantifier '∃x' or to the identity sign '='. It is not evident from their grammatical form that (49) says that nothing is identical to Bigfoot and that (50) says that something is identical to Sting.

In any case, in claiming that the meaning of a proper name is its referent, Millianism has the unfortunate implication that a sentence like (49), which contains a name that lacks a referent, is not fully meaningful but is nevertheless true. And if the meaning of a proper name is its referent, then (50) presupposes the very proposition it asserts; indeed, its meaningfulness depends on its truth.

The case of non-referring names has an important consequence for logic. In standard first-order logic, individual constants are assumed to refer, so that, by existential generalization, 'Fa' entails '(∃x)Fx.' This assumption conflicts with the fact that some proper names do not refer. So-called free logics, which do not take existential general-

ization as axiomatic, have been devised to accommodate empty names. However, adopting a free logic does not help explain how different empty names, like 'Bigfoot' and 'Pegasus,' can differ semantically. It provides no explanation for the difference in content between (49) and (51),

(51) Pegasus does not exist.

Leaving aside the common controversies about proper names, consider uses of proper names that tend to be overlooked by philosophers and logicians. For example, names can be used as predicates (Lockwood 1975). Also, they can be pluralized and combined with quantifiers as in (52),

(52) Many Kennedys have died tragically.

This conflicts the treatment of proper names as individual constants or logically singular terms, and suggests that proper names are more like other nominals than is commonly supposed. In syntax, it is common to treat nominals as constituents of noun phrases, which included a position for a determiner as well, as in 'a man,' 'few tigers,' 'all reptiles,' and 'some water.' And note that in some languages, such as Italian and German, names are often used with definite articles.

A further complication is that proper names seem to function as variable binders. To see this, notice that in the following two sentences,

(53) Marvin$_1$ hates his$_1$ supervisor.
(54) Every employee$_1$ hates his$_1$ supervisor.

the relation between the pronoun and the noun phrase that syntactically binds it appears to be the same. It is sometimes suggested that the pronoun 'his$_1$' is an anaphor when bound by a singular referring expression, such as a proper name, and is a variable when bound by a quantificational phrase. However, it is difficult to see what the relevant difference here could be. Notice further that there are readings of (55) and (56) in which the pronoun functions as a bound variable:

(55) Marvin and every other employee hates his supervisor
(56) Only Marvin hates his supervisor.

Against the suggestion that a proper name is a variable binder it could be argued, I suppose, that in (55) and (56) it is the phrase in which the proper name occurs that binds the pronoun, but consider the following example, involving ellipsis:

(57) Marvin hates his supervisor, and so does every other employee.

If the pronoun is not a bound variable, then (57) could only mean that every other employee hates Marvin's supervisor. It could not have a reading on which it says that every other employee hates his respective supervisor.

4 Adjectives

When a noun is modified by a adjective, it is customary to symbolize this by means of conjunction. A sentence like (58) is standardly symbolized by (58_{PL}) or in restricted quantifier notation by (58_{RQ}):

(58) Enzo has a red car.
(58_{PL}) $(\exists x)(Hex \,\&\, (Rx \,\&\, Cx)$
(58_{RQ}) [an x: Rx & Cx] Hex

Leaving aside the difference between (58_{PL}) and (58_{RQ}), notice that they both render the modification as predicate conjunction. In effect, something is a red car just in case it is a car and it is red. Intuitively, however, it seems that the modification restricts the sort of thing in question. That is, just as 'car' applies to cars, so 'red car' applies to those cars that are red. (58_{PL}) and (58_{RQ}) do not quite capture this.

Even so, using conjunction to adjectival modification does seem to explain why (59) entails (60),

(59) Garfield is a fat cat.
(60) Garfield is a cat.

where 'Garfield' is the name of a child's pet. As (59_{PL}) and (60_{PL}) represent these sentences,

(59_{PL}) $Fg \,\&\, Cg$
(60_{PL}) Cg

the entailment is from conjunction to conjunct, and that is a formal entailment. However, there is a problem here, as illustrated by (61) and (62),

(61) Springfield is a plastic cat.
(62) Springfield is a cat.

where 'Springfield' is the name of a child's toy. (61) does not entail (62), since plastic cats aren't cats. Whether or not (59_{PL}) is the best way to symbolize (59), surely (61_{PL}),

(61_{PL}) $Ps \,\&\, Cs$

is not even a good way to symbolize (61). Plastic cats are not cats that are plastic (just as counterfeit money is not money). Notice, however, that when 'plastic' modifies, say, 'hat,' the resulting phrase applies to a subcategory of hats. So sometimes the entailment from 'x is a plastic K' to 'x is a K' holds, and sometimes it does not. This shows that when the entailment does hold, it is not a formal entailment, and not explained by logic alone. (For further discussion of these and other issues involving adjectives, see Partee 1995.)

5 Adverbs and Events

Consider the fact that (63) entails (64) and (65):

(63) Jack is touching Jill gently with a feather.
(64) Jack is touching Jill gently.
(65) Jack is touching Jill.

Standard symbolizations of these sentences make these entailments problematic, because 'touch gently' is treated is a distinct predicate from 'touch' and whereas in (63) the predicate is treated as three-place predicate, in (64) and (65) it represented as two-place. Then these sentences come out (semi-formalized) as:

(63′) Touch gently (Jack, Jill, a feather)
(64′) Touch gently (Jack, Jill)
(65′) Touch (Jack, Jill)

Special meaning postulates are needed to account for the entailments. It needs to be assumed that to touch someone with something is to touch someone and that to touch someone gently is to touch someone. Davidson (1967) suggested that such entailments can best be explained on the supposition that sentences containing action verbs (or other verbs implying change) involve implicit quantification to events. Then these sentences can be symbolized as:

(63_e) $\exists e$(Touching(Jack, Jill, e) & Gentle(e) & With(a feather, e).
(64_e) $\exists e$(Touching(Jack, Jill, e) & Gentle(e)).
(65_e) $\exists e$(Touching(Jack, Jill, e).

Given these symbolizations, (64) and (65) are formal entailments of (63).

Implicit event quantification also helps handle what Lewis (1975) calls adverbs of quantification, such as 'always,' 'never,' 'often,' 'rarely,' 'sometimes,' and 'usually.' For example, (66) can be symbolized as (66_e):

(66) Jack always touches Jill gently.
(66_e) $\forall e$(Touching(Jack, Jill, e) & Gentle(e)).

Despite the perspicuousness of this symbolization and the explanatory value of the previous ones, they all seem to suffer from a familiar problem: they introduce structure that does not seem to be present in the sentences they purport to symbolize. However, Parsons (1990) and Higginbotham (2000) have offered various reasons for supposing that this problem is not genuine.

6 Utterance Modifiers

There are certain expressions that do not contribute to the propositional contents of the sentences in which they occur and thus fall outside the scope of logical symboliza-

tion. I don't mean interjections like 'Oh' and 'Ah' but a wide range of expressions that may be called 'utterance modifiers.' These locutions, like 'moreover,' 'in other words,' and 'now that you mention it,' are used to comment on the main part of the utterance in which they occur, as in:

(67) *Moreover*, Bill is honest.
(68) *In other words*, Bill is a liar.
(69) New York is, *now that you mention it*, a great place to visit.

Such locutions are vehicles for the performance of second-order speech acts. Thus, for example, 'moreover' is used to indicate that the rest of the utterance adds to what was previously said, and 'in other words' indicates that the balance of the utterance will reformulate something just said.

Because of the second-order function of an utterance modifier, it is not semantically coordinate, though syntactically coordinate, with the rest of the sentence. If it is a connective, it is a *discourse* as opposed to a *content* connective. To appreciate the difference, compare the uses of 'although' in the following two utterances:

(70) *Although* he didn't do it, my client will plead guilty.
(71) *Although* I shouldn't tell you, my client will plead guilty.

In (70), the content of the main clause contrasts with the content of the subordinate clause. The use of 'although' indicates that there is some sort of clash between the two. In (71), on the other hand, there is no suggestion of any contrast between the client's pleading guilty and his lawyer's divulging it. Here the speaker (the lawyer) is using the 'although' clause to perform the second-order speech act of indicating that he shouldn't be performing the first-order speech act of revealing that his client will plead guilty.

There are a great many utterance modifiers, and I have catalogued and classified them elsewhere (Bach 1999b: sec. 5). They can pertain to the topic of conversation, the point of the utterance or its relation to what preceded, the manner of expression, or various other features of the utterance. To illustrate their diversity, here are a few more examples of them:

by the way, to sum up, in a nutshell, figuratively speaking, in a word, frankly, off the record, to be specific, by the same token, be that as it may

It should be understood that these locutions do not function exclusively as utterance modifiers. They function as such only when they occur at the beginning of a sentence or are otherwise set off. But when they do so function, they do not contribute to the primary propositional content of the sentence that contains them and therefore fall outside the scope of logical symbolization.

7 Logical Form as Grammatical Form

Ever since Frege, Russell, and the early Wittgenstein, many philosophers have thought that the structures of sentences of natural languages do not mirror the structures of

the propositions they express. Whether their goal is to develop a language adequate to science, to avoid unwanted ontological commitments, to provide a framework for the analysis of propositions, or merely to adopt a notation that makes the logical powers (formal entailment relations) of sentences explicit and perspicuous, philosophers have generally not supposed that logical forms are intrinsic to natural language sentences themselves. They have supposed, as Russell did in the case of sentences containing definite descriptions, that grammatical form is often misleading as to logical form. From a linguistic point of view, however, logical form is a level of syntactic structure. The logical form of a sentence is a property of the sentence itself, not just of the proposition it expresses or of the formula used to symbolize it. From this perspective, it makes no sense to say that grammatical form is misleading as to logical form.

If logical form is a property of sentences themselves and not merely of the propositions they express or of the formulas used to symbolize them, it must be a level of grammatical form. It is that level which provides the input to semantic interpretation, the output of which consists of interpreted logical forms. This is on the supposition that natural language semantics is compositional, and that the semantics of a sentence is a projection of its syntax. Anything short of that puts the notion of logical form in a different light. If it is essentially a property of propositions, not sentences, or merely a property of logical formulas, then two structurally different sentences, or a sentence and a formula, can express the same proposition, in which case to say that a sentence has a certain logical form is just to say that it expresses a proposition of that form or can be symbolized by a formula with that form. If logical form is not a property of sentences themselves, any reference to the logical form of a sentence is just an elliptical way of talking about a property of the proposition it expresses or of the logical formula used to symbolize it.

There are various sorts of linguistic evidence for a syntactic level of logical form. Consider first the case of scope ambiguity, as in (72),

(72) Most boys love some girl.

Its two readings are captured in semi-English restricted quantifiers as follows,

(73) a. [most x: boy x] ([some y: girl y] (x loves y))
b. [some y: girl y] ([most x: boy x] (x loves y))

where the order of the quantifiers determines relative scope. Here it might be objected that this notation does not respect syntax because it moves the quantificational phrases to the front, leaving variables in the argument positions of the verb. However, as May (1985) explains, working within the syntactic framework of GB theory, such movement of quantificational phrases parallels the overt movement of wh-phrases in question formation, as in (74),

(74) [which x: girl x] (does Marvin love x)?

The transitive verb 'love' requires an object, and the variable marks the position from which the wh-phrase 'which girl' has moved. There are constraints on wh-movement,

and these, in conjunction with other syntactic constraints, explain why, for example, (75a) and (75b) are grammatical and (75c) is not,

(75) a. Who does Jack believe helped Jill?
b. Who does Jack believe that Jill helped?
c. *Who does Jack believe that helped Jill?

why (76a) is ambiguous and (76b) is not,

(76) a. What did everyone see?
b. Who saw everything?

and why (77a) but not (77b) is possible with a co-referential interpretation,

(77) a. Who_1 saw his_1 dog?
b. *Who_1 did his_1 dog see?

May (1985) presents compelling arguments to show that what he calls "quantifier raising" (QR) can explain not only scope ambiguities but a variety of other phenomena. Of course, QR differs from wh-movement in that it is not overt, occurring only at the level of logical form (LF). Positing QR at LF explains the bound-variable interpretation of VP-ellipsis, as in (78),

(78) Cal loves his mother, and so do Hal and Sal.

on which Hal and Sal are being said to love their own mothers, not Cal's. It also explains the phenomenon of antecedent-contained deletion, illustrated by (79),

(79) Clara visited every town that Carla visited.

(79) would be subject to an interpretive regress unless it has, at the level of LF, something like the following form,

(80) [every x: (town that Carla visited) x] (Clara visited x)

which is clearly interpretable. The linguistic arguments based on data like these cannot be presented here, but suffice it to say that they all appeal to independently motivated principles to explain the phenomena in question. The syntactic level of logical form is supported by the same sorts of empirical and theoretical considerations that support other levels of grammatical representation.

8 Summary

There are many topics we haven't even touched on here (see Further reading), including negation, modalities, mass terms, plural quantifiers, quantificational adverbs,

higher-order quantification, quantifier domain restriction, implicit arguments, pronouns and anaphora, prepositions, tense and aspect, context-dependence, vagueness, and semantic underdetermination (sentences that do not express complete propositions, even with context-sensitive references fixed). We have not examined the linguistic arguments for a syntactic level of logical form. Moreover, there are many different syntactic frameworks and, as later chapters explain, many different types of logic and various approaches to each. In short, there is an open-ended range of linguistic phenomena for a diversity of syntactic frameworks and logical theories to take into account. Even so, as suggested by the limited range of phenomena we have discussed, apparent divergences between the behavior of logically important expressions or constructions in natural languages and their logical counterparts are often much narrower than they seem. And where grammatical form appears misleading as to logical form, this appearance is often the result of limiting consideration to standard logic systems, such as first-order predicate logic, and failing to appreciate that insofar as logical form is a property of natural language sentences and not just a property of artificial forms used to symbolize them, logical form is a level of grammatical form.

References

Bach, K. (1994) Conversational impliciture. *Mind & Language*, 9, 124–62.

Bach, K. (1999a) The semantics-pragmatics distinction: What it is and why it matters. In K. Turner (ed.), *The Semantics-Pragmatics Interface from Different Points of View* (pp. 65–84). Oxford: Elsevier.

Bach, K. (1999b) The myth of conventional implicature. *Linguistics and Philosophy*, 22, 327–66.

Bach, K. (2000) A puzzle about belief reports. In K. M. Jaszczolt (ed.), *The Pragmatics of Belief Reports* (pp. 99–109). Oxford: Elsevier.

Barwise, J. and Cooper R. (1981) Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4, 159–219.

Braun, D. (1998) Understanding belief reports. *Philosophical Review*, 107, 555–95.

Chomsky, N. (1986) *Knowledge of Language*. New York: Praeger.

Cohen, L. J. (1971) The logical particles of natural language. In Y. Bar-Hillel (ed.), *Pragmatics of Natural Language* (pp. 50–68). Dordrecht: Reidel.

Davidson, D. (1967) The logical form of action sentences. In N. Rescher (ed.), *The Logical of Decision and Action* (pp. 81–95). Pittsburgh: University of Pittsburgh Press.

Dudman, V. H. (1991) Interpretations of 'if'-sentences. In Jackson (ed.), *Conditionals* (pp. 202–32). Oxford: Oxford University Press.

Edgington, D. (1991) Do conditionals have truth conditions? In Jackson (ed.), *Conditionals* (pp. 176–201). Oxford: Oxford University Press.

Francescotti, R. M. (1995) EVEN: The conventional implicature approach reconsidered. *Linguistics and Philosophy*, 18, 153–73.

Frege, G. (1892) On sense and reference. In P. Geach and M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*, 3rd edn (pp. 56–78). Oxford: Blackwell, 1980.

Grice, P. (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Higginbotham, J. (2000) On events in linguistic semantics. In J. Higginbotham, F. Planesi, and A. C. Varzi (eds.), *Speaking of Events* (pp. 49–79). Oxford: Oxford University Press.

Horn, L. (1996) Exclusive company: *Only* and the dynamics of vertical inference. *Journal of Semantics*, 13, 1–40.

Jackson F. (ed.) (1991) *Conditionals*. Oxford: Oxford University Press.
Kaplan, D. (1979) On the logic of demonstratives. *Journal of Philosophical Logic*, 8, 81–98.
Kay, P. (1990) Even. *Linguistics and Philosophy*, 13, 59–111.
Levinson, S. (2000) *Default Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
Lewis, D. (1975) Adverbs of quantification. In E. Keenan (ed.), *Formal Semantics of Natural Language* (pp. 3–15). Dordrecht: Reidel.
Lockwood, M. (1975) On predicating proper names. *Philosophical Review*, 84, 471–98.
May, R. (1985) *Logical Form: Its Structure and Derivation*. Cambridge, MA: MIT Press.
Neale, S. (1990) *Descriptions*. Cambridge, MA: MIT Press.
Neale, S. (1994) Logical form and LF. In C. Otero (ed.), *Noam Chomsky: Critical Assessments* (pp. 788–838). London: Routledge.
Parsons, T. (1990) *Events in the Semantics of English*. Cambridge, MA: MIT Press.
Partee, B. (1995) Lexical semantics and compositionality. In L. R. Gleitman and M. Liberman (eds.), *An Invitation to Cognitive Science*, 2nd edn, vol. 1: *Language* (pp. 311–60). Cambridge, MA: MIT Press.
Quine, W. V. (1960) *Word and Object*. Cambridge, MA: MIT Press.
Russell, B. (1905) On denoting. *Mind*, 14, 479–93.
Salmon, N. (1986) *Frege's Puzzle*. Cambridge, MA: MIT Press.
Salmon, N. (1992) Reflections on reflexivity. *Linguistics and Philosophy*, 15, 53–63.
Soames, S. (2001) *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford: Oxford University Press.
Strawson, P. (1952) *Introduction to Logical Theory*. London: Methuen.
Strawson, P. F. (1986) 'If' and '⊃'. In R. Grandy and R. Warner (eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends* (pp. 229–42). Oxford: Oxford University Press.

Further Reading

Bach, K. (2000) Quantification, qualification, and context: A reply to Stanley and Szabó. *Mind & Language*, 15, 263–83.
Edgington, D. (1995) On conditionals. *Mind*, 104, 235–329.
Frege, G. (1972) *Conceptual Notation and Related Articles*, ed. and trans. T. W. Bynum. Oxford: Oxford University Press.
Horn, L. (1989) *A Natural History of Negation*. Chicago: Chicago University Press.
Hornstein, N. (1994) *Logical Form: From GB to Minimalism*. Oxford: Blackwell.
King, J. (1995) Structured propositions and complex predicates. *Noûs*, 29, 516–35.
Kripke, S. (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Lappin, S. (1991) Concepts of logical form in linguistics and philosophy. In A. Kasher (ed.), *The Chomskyan Turn* (pp. 300–33). Oxford: Blackwell.
Lepore, E. (2000) *Meaning and Argument*. Oxford: Blackwell.
Ludlow, P. and S. Neale (1991) Indefinite descriptions: In defense of Russell. *Linguistics and Philosophy*, 14, 171–202.
May, R. (1991) Syntax, semantics, and logical form. In A. Kasher (ed.), *The Chomskyan Turn* (pp. 334–59). Oxford: Blackwell.
McCawley, J. (1993) *Everything Linguists Have Always Wanted to Know about Logic But Were Afraid to Ask*, 2nd edn. Chicago: Chicago University Press.
Partee, B. (1989) Binding implicit variables in quantified contexts. *Proceedings of the Chicago Linguistics Society*, 25, 342–65.

Perry, J. (1997) Indexicals and demonstratives. In B. Hale and C. Wright (eds.), *A Companion to the Philosophy of Language* (pp. 586–612). Oxford: Blackwell.
Radford, A. (1997) *Syntax: A Minimalist Introduction*. Cambridge: Cambridge University Press.
Sainsbury, M. (1991) *Logical Forms: An Introduction to Philosophical Logic*. London: Routledge.
Sanford, D. (1989) *If P, then Q: Conditionals and the foundations of reasoning*. London: Routledge.
Schein, B. (1993) *Plurals and Events*. Cambridge, MA: MIT Press.
Sells, P. (1985) *Lectures on Contemporary Syntactic Theories*. Stanford: CSLI.
Stanley, J. and Z. Szabó (2000) On quantifier domain restriction. *Mind & Language*, 15, 219–61.
Woods, M. (1997) *Conditionals*. Oxford: Oxford University Press.

5

Puzzles about Intensionality

NATHAN SALMON

I

Nonextensional notions – such as *necessity*, *possibility*, and especially notions of propositional attitude like *believing that* – raise a number of perplexing philosophical questions, some very old. One issue concerns the sorts of objects that are necessary or possible or are believed or disbelieved. What exactly are they? The standard answer is *propositions*, understood as units of information semantically expressed by declarative sentences but not belonging to any particular language, like the common content of 'Snow is white' and the French '*La neige est blanche*.' W. V. Quine (1956) has objected to propositions as the contents of sentences and the objects of belief on grounds of an alleged obscurity of the 'conditions' under which a pair of propositions p and q are the same. Quine proposes replacing a sentence like

(1) Chris believes that the Earth is round,

which evidently entails the existence of a proposition (that the Earth is round), with

(2) Chris believes-true$_{\text{ENG}}$ 'The Earth is round,'

which, Quine says, is committed to the existence of an English sentence but not to any proposition thereby expressed. He cautions that *believing-true* a sentence is not to be confused with believing the sentence to be true, since Chris (who may speak no English) can believe that the Earth is round – or as we now put it, Chris can believe-true$_{\text{ENG}}$ 'The Earth is round' – without believing that the English sentence 'The Earth is round' is true (i.e. without believing-true$_{\text{ENG}}$ ''The Earth is round' is true$_{\text{ENG}}$'). On closer inspection this proposal collapses. Quine's cautionary remark raises the question of just what belief-truth of a sentence is. Quine argues that one who accepts propositions cannot legitimately complain that the notion of belief-truth is obscure, since (2) is definable for the propositionalist as

(3) Chris believes the proposition expressed$_{\text{ENG}}$ by 'The Earth is round.'

On this explanation, the word for Quine's surrogate notion might be more perspicuously spelled 'believes-the-content$_{ENG}$-of.' Truth, it turns out, is beside the point. Contra Quine, however, (3) is exactly how the notion *cannot* be defined. If it is, then (2) is as committed to the proposition that the Earth is round as (1) is. If (2) is to fulfill its mission, its content must be explained without any appeal to the proposition that the Earth is round. Furthermore, Alonzo Church (1956) demonstrated that (3) does not mean the same as (1). Both designate the offending proposition, but (3) merely describes it as whatever is expressed by a certain English sentence whereas (1) identifies the actual proposition more directly. This is easily seen by translating both (1) and (3) into another language, say, French, while preserving literal meaning:

(1′) *Chris croit que la terre est ronde.*
(3′) *Chris croit la proposition exprimée$_{ANG}$ par* 'The Earth is round.'

It is apparent that these sentences do not carry the same information for a French speaker who speaks no English. Quine concedes Church's point, protesting that he does not claim that (2) has the same meaning as (1), only the same truth value. But if (1) and (2) are alike in truth value, it follows once again that (2) is true only if there is a proposition that the Earth is round. The case for propositions is strikingly powerful, while no viable alternative has yet been offered.

Acknowledging propositions as the objects of belief and other attitudes provides an answer to one question, only to raise a host of further questions. *Kripke's Puzzle* about belief concerns a normal French speaker, Pierre, who on reflection sincerely assents to the French sentence '*Londres est jolie.*' Later, Pierre learns the English language through immersion. Aware that 'London' names the city where he now resides, but unaware that it names the same city he calls '*Londres*,' Pierre sincerely and reflectively assents to 'London is not pretty' – while still sincerely and reflectively assenting to '*Londres est jolie.*' Does Pierre believe (the proposition) that London is pretty? Assuming an extremely plausible Principle of Disquotation, and assuming standard literal translation of French into English, any normal French speaker who sincerely and reflectively assents to '*Londres est jolie*' and who is not under any relevant linguistic confusion culminating in misunderstanding, believes that London is pretty. Whereas by the English version of Disquotation, Pierre's assent to 'London is not pretty' likewise indicates a belief that London is not pretty. Yet Pierre evidently does not contradict himself. Worse, assuming a Strengthened Principle of Disquotation – that a normal speaker who is not reticent or under a relevant linguistic confusion sincerely and reflectively assents to a declarative sentence iff the speaker believes the proposition thereby expressed – Pierre's failure to assent to 'London is pretty' indicates he does *not* believe that London is pretty.

II

Another cluster of issues concerns the distinction of *de dicto* and *de re*. Quine noted that a sentence like 'The number of planets might have been even' may be understood two ways. On the *de dicto* reading, it expresses that the prospect of an even number of planets is a possibility. This is true in some ordinary sense of 'possible' or 'might,' since

there might have been ten planets instead of nine. On the *de re* reading the sentence instead asserts something of the actual number of planets, that is nine: that *it* might have been even instead of odd. This is false on any natural understanding of 'might.' The distinction arises also for belief. Thus 'Smith believes the number of planets is even' may be understood as expressing that Jones believes there are an even number of planets (*de dicto*), or alternatively, that Smith believes of the number nine that it is even (*de re*). (A common confusion conflates the distinction of *de dicto* and *de re* with Keith Donnellan's (1966) distinction between two types of uses of definite descriptions: the *attributive* use on which 'the such-and-such' is used to mean *whatever is uniquely such-and-such*, and the *referential* use on which the description is used instead to name something in particular to which the speaker is relevantly connected. That the two distinctions are different is proved by the fact that a *de re* reading allows the description to be used referentially or attributively.) Kripke's Puzzle demonstrates that *de dicto* belief alone generates hard riddles. Adding *de re* attitudes into the mix compounds the mystery. Whether or not Pierre believes that London is pretty, it seems beyond reasonable dispute that Pierre believes *of* London that it is pretty. But if propositions are the objects of *de dicto* belief, *de re* beliefs appear to be something else again. Is there something – some object – common to all who believe of Socrates that, say, if he is a man then he is mortal? There is the man, Socrates himself, but is there anything else? If so, what?

Related questions took on a distinctly logical flavor, and new questions in philosophical logic arose, when Russell introduced his Theory of Descriptions, with its concomitant distinction between *primary* and *secondary occurrence* – a distinction that for all intents and purposes duplicates *de re* and *de dicto*, respectively, where definite or indefinite descriptions ('denoting phrases') are involved. *Russell's Puzzle* of how George IV could wish to know whether Scott is the author of *Waverley* without wishing to know whether Scott is Scott was solved, in part, by recognizing two senses of wondering whether Scott is the author of *Waverley*: King George may wonder whether Scott and no one else wrote *Waverley* (secondary occurrence); or instead (or in addition), George may wonder concerning *Waverley*'s author (i.e. Scott), whether Scott is *him* (primary). The *de re* is aptly represented using a pronoun ('him') or the logician's variable:

($\exists x$)[x is sole author of *Waverley* & George IV wondered whether: Scott = x],
($\exists n$)[there are exactly n planets & it is possible that: n is even]
(λx)[Pierre believes that: x is pretty](London), etc.

Assuming (with Russell, for the sake of illustration) that 'Scott' and 'London' are genuine names, the attributed *de re* attitudes are indeed a wonder whether Scott is Scott and a belief that London is pretty. Russell offered an answer to the question of what interrelations of logical dependence exist, given that Scott = the author of *Waverley*, between believing that Scott is the author of *Waverley* and believing that Scott is Scott. His answer is: none. But deep questions concerning their connections remain.

Characteristic of representing the *de re* using the apparatus of first-order logic is the occurrence of a variable within a nonextensional context bound from outside that context. The question of what it is to believe (or wonder, etc.) something *de re* con-

cerning Scott receives a sharpened formulation: what is the proper way to interpret an open sentence of the form

George believes that: . . . x . . .

under the assignment of Scott as value for the free variable or pronoun? *Quine's Puzzle* about Ralph and Ortcutt is best posed using this apparatus. Given that Ralph believes that the man in the brown hat is a spy but not that the man seen at the beach is a spy, even though it is Ortcutt in both cases, what sense can be made of

(4) Ralph believes that: x is a spy

under the assignment of Ortcutt to 'x'? Consider first an easier question: is (4) true or false (in English, plus variables) under this assignment? Or in the terminology of Alfred Tarski, does Ortcutt *satisfy* (4)? The obvious reply, as Quine set out the case, is that he does. Quine misled a generation of readers into thinking his puzzle is to some extent a puzzle of philosophical psychology, and is less tractable than it is, by objecting on the questionable grounds that if Ortcutt satisfies (4), then Ralph believes that Ortcutt is a spy even while sincerely and vehemently affirming 'Ortcutt is no spy.' *Pace* Quine, the problem is not how to make Ralph come out consistent. The problem is one of philosophical logic, and is concerned not so much with Ralph as with Ortcutt: is he believed to be a spy? The answer is that despite Ralph's denials, Ortcutt is indeed so believed. If it follows from this (I agree that it does, though most might disagree, perhaps even Quine) that Ralph also believes, *de dicto*, that Ortcutt is a spy, then so he does. Ralph's believing that Ortcutt is a spy while failing to assent to 'Ortcutt is a spy' violates Kripke's Strengthened Principle of Disquotation. But Kripke's own examples demonstrate how dubious that principle is. The principle should be measured against the examples, not the other way around. Belief need not always culminate in assent – even belief with understanding, on reflection, without reticence, etc. – witness Kripke's Pierre. Pierre's doxastic disposition with regard to the question of London's pulchritude parallels Ralph's with regard to Ortcutt's participation in unlawful espionage.

Recognizing that Ortcutt satisfies (4) places an important restriction on the answer to the question of how to interpret (4), but the question still needs an answer. *Neo-Fregeanism* encompasses attempts to provide an answer faithful to the idea that the objects of belief are propositions of a particular sort: Fregean *thoughts*, which are purely conceptual through and through. Neo-Fregeanism faces a number of serious difficulties. Indeed, Hilary Putnam's imaginative Twin Earth thought-experiment seems to demonstrate that *de re* belief and other *de re* attitudes are not adequately captured by Fregean thoughts, since any pair of individuals who are molecule-for-molecule duplicates will entertain the very same set of Fregean thoughts despite having different *de re* attitudes. *Neo-Russellianism* provides a simple alternative solution: (4) attributes belief of a *singular proposition*, which is about Ortcutt in virtue of including Ortcutt himself among the proposition's constituents. Neo-Russellianism does not merely avoid the problems inherent in neo-Fregeanism. It is strongly supported by considerations from philosophical syntax and logic. An English sentence of the form

α believes that ϕ,

is true if and only if the individual designated by α believes the proposition expressed by ϕ. Thus, for example, (1) is true_{ENG} if and only if Chris believes the proposition expressed_{ENG} by 'The Earth is round,' to wit, that the Earth is round. Likewise, then, (4) is true_{ENG} under the assignment of Ortcutt as value for the variable 'x' if and only if Ralph believes the proposition expressed_{ENG} by 'x is a spy' under the same assignment of Ortcutt to 'x.' What proposition does 'x is a spy' express_{ENG} under this assignment? (Cf. What does 'He is a spy' express_{ENG} under the assignment of Ortcutt to the pronoun 'he'?) The variable 'x' has an assigned value (viz., Ortcutt) but, unlike the description 'the man in the brown hat,' does not have a Fregean *sense* which determines this value. If it did, (4) would be *de dicto* rather than *de re*. The variable's only semantic content is its value. The proposition expressed is thus exactly as neo-Russellianism says it is: the singular proposition about Ortcutt, that he is a spy.

III

The *de dicto*/*de re* distinction may be tested by anaphoric links to a descriptive phrase. Consider:

> Quine wishes he owned a sloop, but it is a lemon.
> Ralph believes a female spy has stolen his documents; she also tampered with the computer.

These sentences strongly favor a *de re* reading. Appropriately understood, each evidently entails the *de re* reading of its first conjunct, even if the first conjunct itself is (somewhat perversely) read *de dicto*. If, as alleged, it is a lemon, then there must be an *it* that is a lemon, and that *it* must be a sloop that Quine wants. Similarly, if she tampered with the computer, then there must be a *she* who is a spy and whom Ralph suspects of the theft. The *de dicto*/*de re* distinction comes under severe strain, however, when confronted with Peter T. Geach's (1967) ingenious Hob/Nob sentence:

(5) Hob thinks a witch has blighted Bob's mare, and Nob wonders whether she (the same witch) killed Cob's sow.

This puzzling sentence seems to resist both a *de re* and a *de dicto* reading. If there is a *she* whom Nob wonders about, then that *she*, it would appear, must be a witch whom Hob suspects of mare blighting. But the sincere utterer of (5) intuitively does not seem committed in this way to the reality of witches. Barring the existence of witches, though (5) may be true, there is no actual witch about whom Hob suspects and Nob wonders. Any account of the *de dicto*/*de re* that depicts (5) as requiring the existence of a witch is *ipso facto* wrong. There is a natural reading of (5) that carries an ontological commitment to witches, viz., the straightforward *de re* reading. The point is that the intended reading does not.

A tempting response to Geach's Puzzle construes (5) along the lines of

(5_{dd}) (*i*) Hob thinks: a witch has blighted Bob's mare; and (*ii*) Nob wonders whether: the witch that (Hob thinks) blighted Bob's mare also killed Cob's sow.

Yet this will not do; (5) may be neutral concerning whether Nob has a true belief about, let alone shares, Hob's suspicion. Nob's wondering need not take the form "Did the same witch that (Hob thinks) blighted Bob's mare also kill Cob's sow?" It may be that Hob's thought takes the form "Maggoty Meg blighted Bob's mare" while Nob's takes the form "Did Maggoty Meg kill Cob's sow?" If so, (5) would be true, but no fully *de dicto* reading forthcoming.

Worse, Hob's and Nob's thoughts need not involve the same manner of specification. It may be that Hob's thought takes the form "Maggoty Meg has blighted Bob's mare" while Nob's wondering takes the form "Did the Wicked Witch of the West kill Cob's sow?" This appears to preclude a neo-Fregean analysis along the lines of the following:

(*F*) $(\exists\alpha)[\alpha$ **co-represents** for both Hob and Nob & Hob thinks $\ulcorner\alpha$ is a witch who has blighted Bob's mare$\urcorner$ & Nob **thinks** $\ulcorner\alpha$ is a witch$\urcorner$ & Nob **wonders** $\ulcorner$Did α kill Cob's sow?$\urcorner]$.

Geach himself argues that since (5) does not commit its author to the existence of witches, it must have some purely *de dicto* reading or other. He suggests an alternative neo-Fregean analysis, evidently along the lines of the following:

(*G*) $(\exists\alpha)(\exists\beta)[\alpha$ is a witch-representation & β is a witch-representation & α and β **co-represent** for both Hob and Nob & Hob **thinks** $\ulcorner\alpha$ has blighted Bob's mare$\urcorner$ & Nob **wonders** $\ulcorner$Did β kill Cob's sow?$\urcorner]$.

This proposal faces certain serious difficulties, some of which are also problems for (*F*): The relevant notion of a *witch-representation* must be explained in such a way as to allow that an individual representation α (e.g. an individual concept) may be a witch-representation without representing anything at all. More important, the relevant notion of *co-representation* needs to be explained so as to allow that a pair of individual representations α and β may co-represent for two thinkers without representing anything at all for either thinker. Geach does not explicitly employ the notion of co-representation. I include it on his behalf because it, or something like it, is crucial to the proposed analysis. Any analysis, if it is correct, must capture the idea that Hob's and Nob's thoughts have a common focus. Though there is no witch, Hob and Nob are, in some sense, thinking about the *same* witch. It is on this point that *de dicto* analyses generally fail. Even something as strong as (5_{dd}) – already too strong – misses this essential feature of (5). On the other hand, however the notion of vacuously co-representing witch-representations is ultimately explained, by contrast with (*G*), (5) evidently commits its author no more to co-representing witch-representations than to witches. More generally, any analysis along the lines of (*F*) or (*G*) cannot forever avoid facing the well-known difficulties with neo-Fregean analyses generally (e.g. the Twin Earth considerations).

An alternative approach accepts the imposingly apparent *de re* character of (5) at face value, and construes it along the lines of the following:

(6) There is someone whom: (*i*) Hob thinks a witch that has blighted Bob's mare; (*ii*) Nob also thinks a witch; and (*iii*) Nob wonders whether she killed Cob's sow.

This happily avoids commitment to witches. But it does not provide a solution. Hob's and Nob's thoughts need not concern any real person. Maggoty Meg is not a real person, and there may be no one whom either Hob or Nob believe to be the wicked strega herself.

Some proposed solutions to Geach's Puzzle make the unpalatable claim that Hob's and Nob's musings concern a Meinongian Object – a particular witch who is both indeterminate and nonexistent. Many proposed solutions instead reinterpret *de re* attributions of attitude so that they do not make genuine reference to the individuals apparently mentioned therein by name or pronoun. These responses inevitably make equally unpalatable claims involving *de re* constructions – for example, that Nob's wondering literally concerns the very same witch/person as Hob's belief yet neither concerns anyone (or anything) whatsoever, or that *de re* constructions mention or generalize over speech-act tokens and/or connections among speech-act tokens. It would be more sensible to deny that (5) can be literally true on the relevant reading, given that there are no actual witches. The problem with this denial is that its proponent is clearly in denial. As intended, (5) can clearly be true (assuming Hob and Nob are real) even in the absence of witches. Numerous postmodern solutions jump through technical hoops to allow a pronoun ('she') to be a variable bound by a quantifier within a belief context ('a witch') despite standing outside the belief context, hence also outside the quantifier's scope, and despite standing within an entirely separate belief context. These 'solutions' do not satisfy the inquiring mind as much as boggle it. It is one thing to construct an elaborate system on which (5) may be deemed true without 'There is a witch.' It is quite another to provide a satisfying explanation of the content of Nob's attitude, one for which the constructed system is appropriate. How can Nob wonder about a witch, and a particular witch at that – the very one Hob suspects – when there is no witch and, therefore, no particular witch about whom he is wondering? This is the puzzle in a nutshell. It combines elements of intensionality puzzles with puzzles concerning nonexistence and puzzles concerning identity, and has been deemed likely intractable.

IV

The solution I urge takes (5) at face value, and takes seriously the idea that false theories that have been mistakenly believed – what I call *myths* – give rise to fabricated but genuine entities. These entities include such oddities as: Vulcan, the hypothetical planet proposed by Babinet and which Le Verrier believed caused perturbations in Mercury's solar orbit; the ether, once thought to be the physical medium through which light waves propagate; phlogiston, once thought to be the element (material substance) that causes combustion; the Loch Ness Monster; Santa Claus; and Meinong's Golden Mountain. Such *mythical objects* are real things, though they are neither material objects nor mental objects ('ideas'). They come into being with the belief in the myth. Indeed, they are created by the mistaken theory's inventor, albeit without the theorist's knowledge. But they do not exist in physical space, and are, in that sense, abstract entities. They are an unavoidable by-product of human fallibility.

Vulcan is a mythical planet. This is not to say, as one might be tempted to take it, that Vulcan is a planet but one of a rather funny sort, for example a Meinongian Object

that exists in myth but not in reality. On the contrary, Vulcan exists in reality, just as robustly as you the reader. But a mythical planet is no more a planet than a toy duck is a duck or a magician is someone who performs feats of magic. A mythical object is an imposter, a pretender, a stage prop. Vulcan is not a real planet, though it is a very real object – not concrete, not in physical space, but real. One might say that the planet Mercury is also a 'mythical object,' in that it too figures in the Vulcan myth, wrongly depicted as being gravitationally influenced by Vulcan. If we choose to speak this way, then it must be said that some 'mythical planets' are real planets, though not really as depicted in the myth. Vulcan, by contrast with the 'mythical' Mercury, is a *wholly mythical* object, not a real planet but an abstract entity inadvertently fabricated by the inventor of the myth. I shall continue to use the simple word 'mythical' as a shorthand for the notion of something wholly mythical.

The existence of fictional objects, in something close to this sense, has been persuasively urged by Peter van Inwagen (1977) and Saul Kripke (1973) as an ontological commitment of our ordinary discourse about fiction. Their account, however, is significantly different from the one I propose. Kripke contends that a mythical-object name like 'Vulcan' is ambiguous between two uses, one of which is parasitic on the other. It would be less deceptive to replace the ambiguous name with two univocal names, 'Vulcan_1' and 'Vulcan_2.' The name on its primary use, 'Vulcan_1,' was introduced into the language, *sans* subscript, by Babinet as a name for an intra-Mercurial planet. Le Verrier used the name in this way in theorizing about Mercury's perihelion. On this use, the name names nothing; 'Vulcan_1' is entirely vacuous. Giving the name this use, we may say such things as that Le Verrier believed that Vulcan_1 affected Mercury's perihelion. Le Verrier's theory is a myth concerning Vulcan_1. The name on its secondary use, 'Vulcan_2,' is introduced into the language (again *sans* subscript) at a later stage, when the myth has finally been exposed, as a name for the mythical planet erroneously postulated, and thereby inadvertently created, by Babinet. Perhaps it would be better to say that a new *use* of the name 'Vulcan' is introduced into the language. 'Vulcan_2' is fully referential. Using the name in this way, we say such things as that Vulcan_2 was a mythical intra-Mercurial planet hypothesized by Babinet. The difference between Vulcan_1 and Vulcan_2 could not be more stark. The mistaken astronomical theory believed by Babinet and Le Verrier concerns Vulcan_1, which does not exist. Vulcan_2, which does exist, arises from the mistaken theory itself. Vulcan_2 is recognized through reflection not on events in the far-off astronomical heavens but on the more local story of man's intellectual triumphs and defeats, particularly on the history of science.

Kripke's account is vulnerable to a familiar family of thorny problems: the classical problem of true negative existentials and the more general problem of the content and truth value of sentences involving vacuous names. Vulcan_1 does not exist. This sentence is true, and seems to say about something (viz., Vulcan_1) that it fails to exist. Yet the sentence entails that there is nothing for it to attribute nonexistence to. Furthermore, on Kripke's account, Le Verrier believed that Vulcan_1 has an impact on Mercury's perihelion. What can the content of Le Verrier's belief be if there is no such thing as Vulcan_1? Furthermore, is the belief content simply false? If so, then it may be said that Vulcan_1 has no impact on Mercury's perihelion. Yet this claim too seems to attribute something to Vulcan_1, and thus seems equally wrong, and for exactly the same

reason, with the claim that Vulcan_1 does have such an impact. Kripke is aware of these problems but offers no viable solution.

I submit that Kripke's alleged primary use of a mythical-object name is itself a myth. To be sure, Babinet believed himself to be naming a real planet in introducing a use of 'Vulcan' into the language. And other users like Le Verrier believed themselves to be referring to a real planet. But this linguistic theory of the name 'Vulcan' is mistaken, and is in this respect exactly like the astronomical theory that Vulcan is a real planet. The two theories complement each other, and fall together hand in hand. The situation should be viewed instead as follows. Babinet invented the theory – erroneous, as it turns out – that there is an intra-Mercurial planet. In doing this, he inadvertently created Vulcan. Indeed, Babinet even introduced a name for this mythical planet. The name was intended for a real planet, and Babinet believed the name thus referred to a real planet (*de dicto*, not *de re*!). But here again, he was simply mistaken. Other astronomers, most notably Le Verrier, became convinced of Babinet's theory, both as it concerns Vulcan (that it is a very real intra-Mercurial planet) and as it concerns 'Vulcan' (that it names the intra-Mercurial planet). Babinet and Le Verrier both believed, correctly, that the name 'Vulcan', on the relevant use, refers to Vulcan. But they also both believed, mistakenly, that Vulcan is a real planet. They might have expressed the latter belief by means of the French version of the English sentence 'Vulcan is a planet,' or other shared beliefs by means of sentences like 'Vulcan's orbit lies closer to the Sun than Mercury's.' These beliefs are mistakes, and the sentences (whether English or French) are false.

Importantly, there is no relevant use of the name 'Vulcan' by Babinet and Le Verrier that is vacuous. So used the name refers to Vulcan, the mythical planet. Le Verrier did *not* believe that Vulcan_1 is an intra-Mercurial planet – or, to put the point less misleadingly, there is no real use marked by the subscript on 'Vulcan' on which the string of words 'Vulcan_1 is an intra-Mercurial planet' expresses anything for Le Verrier to have believed, disbelieved, or suspended judgment about. To put the matter in terms of Kripke's account, what Le Verrier believed was that Vulcan_2 is a real intra-Mercurial planet. Le Verrier's belief concerns the mythical planet, a very real object that had been inadvertently created, then named 'Vulcan,' by Babinet. Their theory about Vulcan was completely wrong. Vulcan is in fact an abstract object, one that is depicted in myth as a massive physical object.

A common reaction is to charge my proposal with miscasting mythical objects as the objects with which myths are concerned. On the contrary, it is objected, if they exist at all, mythical objects enter the intellectual landscape only at a later stage, not in the myth itself but in the subsequent historical account of the myth. A robust sense of reality demands that the myth itself be not about these abstract objects but about *nothing*, or at most about representations of nothing. No one expresses this sentiment more forcefully than Russell:

> [Many] logicians have been driven to the conclusion that there are unreal objects. . . . In such theories, it seems to me, there is a failure of that feeling for reality which ought to be preserved even in the most abstract studies. Logic, I should maintain, must no more admit a unicorn than zoology can; for logic is concerned with the real world just as truly as zoology, though with its more abstract and general features. To say that unicorns have an

> existence in heraldry, or in literature, or in imagination, is a most pitiful and paltry evasion. What exists in heraldry is not an animal, made of flesh and blood, moving and breathing of its own initiative. What exists is a picture, or a description in words. . . . A robust sense of reality is very necessary in framing a correct analysis of propositions about unicorns . . . and other such pseudo-objects. (Russell 1919: 169–70)

I heartily applaud Russell's eloquent plea for philosophical sobriety. But his attitude toward 'unreal' objects is fundamentally confused. To repeat, a mythical planet is not a massive physical object but an abstract entity, the product of creative astronomizing. Likewise, a mythical unicorn or a mythical winged horse is not a living creature but a fabricated entity, the likely product of blurred or fuzzy vision, just as mermaids are the likely product of a deprived and overactive imagination under the influence of liquor – creatures not really made of flesh and blood and fur or scales, not really moving and breathing of their own initiative, but depicted as such in myth, legend, hallucination, or drunken stupor.

It is frequently objected even by those who countenance mythical objects that the Vulcan theory, for example, is merely the theory that there is an intra-Mercurial planet, not the bizarre hypothesis that the relevant abstract entity is that planet. Babinet and Le Verrier, it is observed, did not believe that an abstract entity is a massive heavenly object. Quite right, but only if the sentence is meant *de dicto*. Understood *de re* – as the claim that, even if there is such an abstract entity as the mythical object that is Vulcan, Babinet and Le Verrier did not believe it to be an intra-Mercurial planet – it turns mythical objects into a philosophical black box. What role are these abstract entities supposed to play, and how exactly are their myth-believers supposed to be related to them in virtue of believing the myth? In fact, this issue provides yet another reason to prefer my account over Kripke's. On my account, in sharp contrast, the role of mythical objects is straightforward: they are the things depicted as such-and-such in myth, the fabrications erroneously believed by wayward believers to be planets or the medium of light-wave propagation or ghosts, the objects the mistaken theory is about when the theory is not about any real planet or any real medium or any real ghost. It is not merely that being depicted as such-and-such is an essential property of a mythical object, a feature the object could not exist without. Rather, being so depicted is the metaphysical function of the mythical object; that is *what* it is, its *raison d'être*. To countenance the existence of Vulcan as a mythical planet while at the same time denying that Babinet and Le Verrier had beliefs about this mythical object, is in a very real sense to miss the point of recognizing Vulcan's existence. It is precisely the astronomers' false beliefs about the mythical planet that makes it a mythical planet; if no one had believed it to be a planet, it would not *be* a mythical planet. Come to that, it would not even exist.

Another important point: I am not *postulating* mythical objects. For example, I am not postulating Vulcan. Even if I wanted to, Babinet beat me to it – though he postulated Vulcan as a real planet, not a mythical one. Mythical objects would exist even if I and everyone else had never countenanced or recognized them, or admitted them into our ontology, etc. Rather, I see myself as uncovering some evidence for their independent and continued existence, in something like the manner of the paleontologist who infers dinosaurs from their fossil remains, rather than the theoretical physicist who

postulates a new category of physical entity in order to make better sense of things (even if what I am actually doing is in important respects more like the latter).

Perhaps the most important evidence in favor of this theory of mythical objects is its logical entailment by our thoughts and beliefs concerning myths. We are sometimes led to say and think things like "An intra-Mercurial planet, Vulcan, was hypothesized by Babinet and believed by Le Verrier to affect Mercury's perihelion, but there has never been a hypothetical planet whose orbit was supposed to lie between Mercury and Venus" and "Some hypothetical species have been hypothesized as linking the evolution of birds from dinosaurs, but no hypothetical species have been postulated to link the evolution of mammals from birds." The distinctions drawn cannot be made without a commitment to mythical objects, that is without attributing existence, in some manner, to mythical objects. No less significant, beliefs are imputed about the mentioned mythical objects, to the effect that they are not mythical. Being wrongly believed not to be mythical is just what it is to be mythical. Furthermore, beliefs are imputed to distinct believers concerning the very same mythical object.

Further evidence – in fact, evidence of precisely the same sort – is provided by the Hob/Nob sentence. Geach's Puzzle is solved by construing (5) on its principal reading, or at least on one of its principal readings, as fully *de re*, not in the manner of (6) but along the lines of:

(7) There is a mythical witch such that (*i*) Hob thinks: she has blighted Bob's mare; and (*ii*) Nob wonders whether: she killed Cob's sow.

This has the distinct advantage over (6) that it does not require that both Hob and Nob believe someone to be the witch in question. In fact, it allows that there be no one in particular whom either Hob or Nob believes to be a witch. It does require something not unrelated to this, but no more than is actually required by (5): that there be something that both Hob and Nob believe to be a witch – some*thing*, not some*one*, not a witch or a person, certainly not an indeterminate Meinongian Object, but a very real entity that Nob thinks a real witch who has blighted Bob's mare. Nob also believes this same mythical witch to be a real witch and wonders about 'her' (really: about *it*) whether she killed Cob's sow. In effect, the proposal substitutes ontological commitment to mythical witches for the ontological commitment to real witches intrinsic to the straightforward *de re* reading of (5) (obtained from (7) by deleting the word 'mythical'). There are other witch-free readings for (5), but I submit that any intended reading is a variant of (7) that equally commits the author to the existence of a mythical witch, such as:

(*i*) Hob thinks: some witch or other has blighted Bob's mare; and (*ii*) the (same) mythical witch that Hob thinks has blighted Bob's mare is such that Nob wonders whether: she killed Cob's sow.

Significantly, one who accepts Kripke's account may not avail him/herself of this solution to Geach's Puzzle. On Kripke's account it may be observed that

(*i*) Hob thinks: Meg_1 has blighted Bob's mare; and (*ii*) Nob wonders whether: Meg_1 killed Cob's sow.

The Hob/Nob sentence (5) is not obtainable by existential generalization on 'Meg_1,' since by Kripke's lights, this name is supposed to be vacuous and to occur in non-extensional ('referentially opaque,' *ungerade*) position. Nor on Kripke's (1973) account can 'Meg_2' be correctly substituted for 'Meg_1'; Hob's and Nob's theories are supposed to concern the nonexistent witch Meg_1 and not the mythical witch Meg_2. Kripke might instead accept the following, as a later-stage observation about the Meg_1 theory:

> Meg_2 is the mythical witch corresponding to Meg_1.

Here the relevant notion of *correspondence* places 'Meg_2' in extensional position. While 'Meg_2' is thus open to existential generalization, 'Meg_1' supposedly remains in a non-extensional position where it is not subject to quantification. It is impossible to deduce (5) from any of this. Geach's Puzzle does not support Kripke's account. On the contrary, the puzzle poses a serious threat to that account, with its denial that Hob's and Nob's thoughts are, respectively, a suspicion and a wondering regarding Meg_2.

On my alternative account, we may instead observe that

> Maggoty Meg is a mythical witch. Hob thinks she has blighted Bob's mare. Nob wonders whether she killed Cob's sow.

We may then conjoin and EG (existential generalize) to obtain (7). In the end, what makes (7) a plausible analysis is that it (or some variant) spells out in more precise language what (5) literally says to begin with. Babinet and Le Verrier provide a real-life case in which the thoughts of different thinkers converge on a single mythical object: Babinet thought he had seen an intra-Mercurial planet, and Le Verrier believed that it (the same 'planet') impacted Mercury's perihelion. The primary lesson of Geach's Puzzle is that when theoretical mistakes are made mythical creatures are conceived, and in acknowledging that misbelievers are sometimes related as Nob to Hob, or as Le Verrier to Babinet, we commit ourselves to their illegitimate progeny.

References

Church, Alonzo (1956) *Introduction to Mathematical Logic I*. Princeton University Press.

Donnellan, Keith (1966) Reference and definite descriptions. *The Philosophical Review*, 75, 3, 281–304.

Frege, Gottlob (1892) Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50; translated as On *Sinn* and *Bedeutung*, in Michael Beaney (ed.), *The Frege Reader*. Oxford: Blackwell, 1997, pp. 51–171.

Frege, Gottlob (1918) Der Gedanke. *Beiträge zur Philosophie des deutschen Idealismus*, 1, pp. 58–77; translated as Thoughts, in N. Salmon and S. Soames (eds.) pp. 33–55.

Geach, Peter T. (1967) Intentional identity. *Journal of Philosophy*, 64, 20, 627–32; reprinted in Geach's *Logic Matters*. Oxford: Basil Blackwell, 146–53.

Geach, Peter T. (1976) Two kinds of intentionality. *Monist*, 59, 306–20.

Kaplan, David (1969) Quantifying in. In D. Davidson and J. Hintikka (eds.), *Words and Objections: Essays on the Work of W. V. Quine*. Dordrecht: Reidel, 206–42.

Kaplan, David (1989) Demonstratives. In J. Almog, J. Perry and H. Wettstein (eds.), *Themes from Kaplan*. Oxford: Oxford University Press, 481–563.
Kripke, Saul (1972) *Naming and Necessity*. Cambridge: Harvard University Press.
Kripke, Saul (1973) *Reference and Existence: The John Locke Lectures for 1973*. Oxford University Press, unpublished.
Saul, Kripke (1979) A puzzle about belief. In A. Margalit (ed.), *Meaning and Use*. Dordrecht: Reidel, 239–83; reprinted in N. Salmon and S. Soames (eds.) 102–48.
Quine, W. V. O. (1956) Quantifiers and propositional attitudes. *Journal of Philosophy*, 53; reprinted in Quine's *The Ways of Paradox*. New York: Random House, 183–94.
Russell, Bertrand (1905) On denoting. *Mind*, 14, 479–93; reprinted in R. M. Harnish, *Basic Topics in the Philosophy of Language*. Englewood Cliffs, NJ: Prentice-Hall, 1994, 161–73.
Russell, Bertrand (1919) *Introduction to Mathematical Philosophy*. London: George Allen & Unwin.
Salmon, Nathan (1986) *Frege's Puzzle*. Atascadero, CA: Ridgeview.
Salmon, Nathan (1989) Illogical belief. In J. Tomberlin (ed.), *Philosophical Perspectives, 3: Philosophy of Mind and Action Theory*. Atascadero, CA: Ridgeview, 243–85.
Salmon, Nathan (1995) Relational belief. In P. Leonardi and M. Santambrogio (eds.), *On Quine: New Essays*. Cambridge University Press, 206–28.
Salmon, Nathan (1998) Nonexistence. *Noûs*, 32, 3, 277–319.
Salmon, Nathan (2000) Mythical objects. Forthcoming in the proceedings of the Inland Northwest Philosophy Conference, 2000 (Davidson Press).
Salmon, Nathan and Scott Soames (eds.) (1988) *Propositions and Attitudes*. Oxford University Press.
van Inwagen, Peter (1977) Creatures of fiction. *American Philosophical Quarterly*, 14, 4, 299–308.

Further Reading

Burge, Tyler (1977) Belief *de re*. *Journal of Philosophy*, 69, 338–62.
Dennett, Daniel C. (1968) Geach on intentional identity. *Journal of Philosophy*, 65, 335–41.
Donnellan, Keith (1974) Speaking of nothing. *The Philosophical Review*, 83, 1, 3–31.
Kaplan, David, Bob and Carol and Ted and Alice. In J. Hintikka, J. Moravcsik and P. Suppes (eds.), *Approaches to Natural Language*. Dordrecht: Reidel, 490–518.
Kaplan, David (1986) Opacity. In L. E. Hahn and P. A. Schilpp (eds.), *The Philosophy of W. V. Quine*. La Salle, IL: Open Court, 229–89.
Russell, Bertrand, Knowledge by acquaintance and knowledge by description. In N. Salmon and S. Soames (eds.), 16–32.
Salmon, Nathan (1998) Is *de re* belief reducible to *de dicto*? In Ali A. Kazmi (ed.), *Meaning and Reference: Canadian Journal of Philosophy*, supplementary vol. 23. Calgary, Alberta: University of Calgary Press, 85–110.
Sosa, Ernest (1975) Propositional attitudes *de dicto* and *de re*. *Journal of Philosophy*, 71, 883–96.

6

Symbolic Logic and Natural Language

EMMA BORG AND ERNEST LEPORE

Initially the connection between the formal notation of symbolic logic and ordinary sentences of natural language might seem opaque. Why on earth would anyone want to draw a parallel between the technical and abstract endeavors of formal logicians and what seems more properly an object of study for linguists? However, it has been a common assumption of twentieth-century Anglo-American philosophy that symbolic logic can reveal something important about language. The reasons for this assumption are, in actual fact, not very hard to see.

Arguments (1) and (2) are deductively valid inasmuch as it is impossible for their premises (their first two sentences) to be true and their conclusions (their last sentence) false:

(1) If the Yankees won, then there will be a parade.
The Yankees won.
So, there will be a parade.

(2) If Socrates is a man, then he is mortal.
Socrates is a man.
So, he is mortal.

Moreover, the reason that (1) is valid does not seem to be independent of the reason that (2) is: both seem valid because they share a common form. Each begins with a conditional statement, followed by another premise that asserts the condition part (the *antecedent*) of the first premise, and concludes with its *consequent* part. By virtue of sharing this form, both arguments (and countless others) are not only valid but are valid *in virtue of this shared form.*

Though (1) is about the Yankees and parades, and (2) is about men and their mortality, when our concern is with inference (i.e. issues about which sentences can be validly deduced, or 'follow,' from which others), it seems best to abstract away from any particular *content* and concentrate instead on *structure.* The structure underlying an argument (and hence, the structure underlying the sentences making up that argument) in virtue of which it has its inferential properties is known as its *logical form.* We arrive at statements of logical form by replacing certain expressions (so-called non-

logical ones) with schematic letters and combining these with symbolic representations of the logical components of the argument;[1] for instance, (1) and (2) share the logical form:

$A \supset B$
A
$\therefore B$[2]

(with '$\supset$' representing the logical component 'if . . . then,' 'A' and 'B' standing for propositional claims, and '$\therefore$' indicating the conclusion). The logical representation of a sentence then might be thought of as *a structure that determines from which sentences it can be validly deduced, and which sentences can be validly deduced from it and other premises.*

The notion of logical form has become commonplace in philosophical discussions of language (at least in the analytic tradition), but theorists are not always explicit about the kind of relationship they envisage between natural language sentences and statements in logical form, or about the role they expect such symbolizations to be playing. Our aim in this chapter, then, is to explore these questions; in Section 1 we will concentrate on the question of constraints on logical representations, while in Section 2 we concentrate on the nature of the relationship between natural language and logical form.

1 What are the Constraints on Formal Representations?

Given what we have said so far, the only constraint that must be respected in mapping natural language sentences onto a symbolic notation is that whatever form we assign to a sentence, relative to an argument, must underwrite the logical properties of that argument. However, this condition can lead to some *prima facie* surprising results; to see this, let's look at Frege's system of predicate logic. Frege's logical system was designed to be able to cope with the sort of generality evidenced in sentences like 'All men are mortal' or 'Some girl is happy' (i.e. claims which tell us about the range of objects in the world which possess certain properties, rather than telling us any particulars about specific objects which possess those properties). He attempted to achieve this end with two explicit quantifier symbols, '$\forall$,' '$\exists$' (introduced to stand for the English counterparts 'all' and 'some' respectively), which could combine with predicates (e.g. 'is a man') and variables (given by lower case letters from the end of the alphabet like 'x' or 'y') in order to represent general claims.

A standard practice for representing a universal sentence like 'All men are mortal' in the language of predicate logic is as '$(\forall x)(\text{Man}(x) \supset \text{Mortal}(x))$'; which says in 'logicians' English': *for all things, x, if x is a man then x is mortal.* Although such a claim, if true, entails something about individual males, it does not assert anything about one particular man. If corresponding representations for 'Socrates is a man' and 'Socrates is mortal' render the inference from the first two sentences to the third in argument (2) valid in virtue of form, our techniques have achieved their end.

Suppose, though, that someone complains about the occurrence of the symbol '$\supset$', the notational counterpart, recall, for an English conditional (typically, an 'if . . . then'

statement). Unlike the first premises in (1) and (2), the universal English sentence 'All men are mortal' makes no (overt) mention of conditionality, so why should its symbolic representation do so? This is one respect in which we may find the Fregean representation of natural language sentences surprising.

A second area of divergence between the surface appearance of natural language sentences and their formal representations in Fregean logic comes with respect to numerical claims like 'One girl is happy' or 'Fifteen men are mortal.' Frege's suggestion is that, for all such counting quantifier expressions (like 'one' or 'fifteen'), we use combinations of '$\forall$' and '$\exists$' claims to deliver the logical forms of sentences containing them. So, for example, we can use an instance of the existential quantifier '$\exists$' to symbolically represent that there is (at least) one thing satisfying the given predicate (e.g. '$(\exists x)(Man(x))$'). If we introduce another instance of the existential quantifier (e.g. '$(\exists y)(Man(y))$') and then state that the two existential claims are about distinct objects (e.g. by using a non-identity claim '$y \neq x$'), the final product, symbolized as '$(\exists x)(\exists y)((Man(x) \& Man(y)) \& y \neq x)$' can be used to symbolize the English sentence 'There are (at least) two men.'

Obviously, we can go on 'counting' indefinitely, simply by introducing more existential quantifiers and more non-identities to those already introduced. This might seem a rather laborious way of symbolically representing numerical claims, especially those involving large numbers (imagine the length of the logical representation of 'One hundred and one Dalmatians came home' on this model!); but the technique does allow the Fregean system to formulate many more quantificational claims than we might have envisaged at first, particularly given the limited base of '$\forall$' and '$\exists$'. Despite containing only two basic quantifier expressions, the Fregean system can express, and therefore, formalize, any natural language claim involving a counting quantifier. In short, though Frege's system may introduce more parsimony than the project of codifying logical inferences asks for or demands, if it achieves this end (i.e. if it captures all the inferences that need to be captured), it's hard to see what project is jeopardized by doing it with a minimum of logical symbols.

The conditional form of universal statements in predicate logic, and the complexity of statements involving count quantifiers, might be surprising to us but nothing so far said would require withdrawing our proposed symbolizations based on this sort of consideration. We have been assuming that symbolic representations function merely to codify logical properties and relations involving natural language sentences. If these symbolic representations contain elements not obvious in their natural language counterparts, why should it matter as long as the right inferences are captured in virtue of these assigned forms?

One reason for concern will be addressed in Section 2; for the moment we'll assume that the only self-evident constraint on an adequate symbolization is that it captures correct logical inferences. It should be obvious that this condition can serve to rule out certain suggestions about the logical form of natural language sentences – those which fail to preserve logical inferences will be ruled out. However, it may also turn out that this constraint is insufficient to chose between alternative logical renditions of a natural language sentence; and when this happens, we might, perhaps, expect there to be further constraints which come into play to help us choose. To see this, let's consider a particular example: for in the realm of definite descriptions we can see both the con-

straint to capture logical inferences and the potential need for an additional constraint in play. Questions about the appropriateness of any such additional constraint will then lead us, in Section 2, to consider how we should construe the relationship between natural language sentences and logical form.

Case study: Representing definite descriptions

The idea we will explore in this section is as follows: perhaps finding an adequate symbolic notation for natural language is difficult not because of an absence of *any* symbolic system which looks like it might be up to the job, but because of a *surplus*, each of which is *prima facie* promising. One thought might be that what we need, when faced with alternatives, is a way to choose amongst them. We can tie down the main point here with reference to a well-explored example, viz., definite descriptions. These are expressions of the form 'the F,' where 'F' is a complex or simple common noun, as in 'The woman' or 'The woman who lived in New Jersey.' These expressions have been a focus for philosophical logicians, in part because of divergent intuitions about their linguistic status, and accordingly, about which logical inferences they participate in. One can find in the literature a myriad of different accounts of the logical form of definite descriptions, but we'd like to explore just three which will help to demonstrate the constraints involved in a choice of symbolization.

The first proposal is Frege's, who treated definite descriptions as members of his class of referring terms. The details of his larger philosophy of language are inessential here; what matters is that, according to him, definite descriptions are akin both to names (like 'Bill Clinton' or 'Gottlob Frege') and indexical expressions (like 'I,' 'you' and 'today,' which depend on a context of utterance for a referent).[3] Each of these expressions is treated identically within his system: each is assigned a designator which appears in predicate assignments. 'I am happy,' 'Bill Clinton is happy' and 'The president of the US is happy' can all be symbolized in Frege's notation as 'Ha' (with 'H' symbolizing the predicate 'is happy,' and 'a' designating the object picked out by each referring term).

Famously, Russell, disputed Frege's analysis, arguing that definite descriptions are not proper names, but instead belong to an alternative logical category in Frege's system: viz., the class of quantifier expressions.[4] At first his suggestion might seem odd, for the Fregean quantifiers were explicitly introduced to play roles equivalent to 'all' and 'some,' and *prima facie*, whatever the role of 'the' in our language, it isn't playing either of these. However, Russell's contention is that we can symbolically represent definite descriptions as complex entities constructed out of these two primitive Fregean quantifiers. That is to say, he suggests that we can treat the definite article 'the' in a way analogous to the account Frege gave for counting quantifiers like 'two.'

Informally, a sentence of the form 'The F is G' is represented, according to Russell, as making a uniqueness claim, viz., there is one and only one F that is G.[5] So a sentence of the form 'The tallest man is happy' will be analysed as stating:

(3) There is a tallest man; and
(4) there is only one tallest man; and
(5) whoever he is he is happy.

Collectively these three claims are symbolized within Frege's logical system as (DD):

(DD) $(\exists x)(\text{Tallest man } (x) \ \& \ (\forall y)((\text{Tallest man } (y) \supset y = x) \& \text{Happy}(x)))$

Whereas a sentence containing a genuine referring term is represented by Frege with a simple formula (as in 'Fa'), for Russell, a sentence with a definite description requires a complex logical symbolization like (DD).

Finally, let's introduce a third relatively recent development in quantification theory, which gives us our last account of the logical form of descriptions. Many contemporary theories of quantification, such as the 'Generalized Quantifier' (hereinafter, GQ) theory of Higginbotham and May (1981), and Barwise and Cooper (1981), recommend altering the way the relationship between a quantifier expression and the rest of the sentence it appears in is handled, as well as acknowledging many more primitive quantifier expressions than the two inherited from Frege.[6] The first point relates to a feature of Fregean quantification we noted at the outset: viz., that it needs to introduce the logical connective '$\supset$' into the formal representations of sentences containing 'all'. The reason for this is that the Fregean quantifiers '$\forall$' and '$\exists$' are *unary* or 'free-standing' expressions: they act autonomously to bind a variable, which can then go on to appear in property assignments. It is this independent nature of the quantifier which leads to the need for a logical connective: we treat 'All men are mortal' as containing a free-standing quantifier – '$\forall(x)$' – and then say of the variable which appears next to (and hence which is bound by) the quantifier that: 'if it is a man, then it is mortal.'

However, advocates of a theory of quantification like GQ reject this autonomy for quantifier expressions; they maintain that quantifier expressions are ineliminably bound to the common noun they modify. That is to say, rather than treating 'all' and 'men' as separable units within the logical form of 'all men are mortal,' they suggest we should treat 'all men' as a single, indissoluble unit, which acts together to bind a variable which then appears in the predicate assignment 'is mortal.' In the GQ system of quantification, then, this kind of claim can be represented along the following lines: '[All (x): Man (x)] Mortal (x).' On this kind of model, quantifier expressions are said to be *binary* or *restricted*, requiring a common noun to act in tandem with a quantifier to bind a variable.

Unlike predicate logic, GQ is a second-order logical system: roughly, this means that the objects quantifiers are taken to range over are sets (of objects), rather than their constituents (i.e. the objects themselves). Logical rules for GQ quantifiers are given in terms of numerical relations between sets; for example a GQ quantifier might tell us about the number of objects in common between two sets (i.e. the set of objects in the intersection of two sets). The intuitive idea here is easy enough to see: for instance, the sentence 'All men are mortal' can be understood as telling us that there is no object which belongs to the first set, that is the set satisfying the general term 'men,' which does not also belong to the second set, that is the set of things satisfying the general term 'mortal.' In other words, the number of men that are non-mortal is zero. The GQ rule for 'all' captures this numerical claim: take X to be the set of 'F'-things and Y to be the set of 'G'-things, then a sentence of the form 'All F's are G' is true just in case there are zero objects left over when you take the set Y away from the set X (i.e. that everything in X is also in Y). Similarly, for a quantifier like 'some'; the GQ rule for 'some' is

that a sentence like 'Some man is mortal' is true just in case the number of objects in the intersection of the set of men and the set of mortal things is (greater than or equal to) one.

This leads us on to a second area of difference between GQ and predicate logic relevant for our concerns, for GQ theorists reject Frege's technique for handling counting quantifiers. Rather than analysing expressions like 'three' and 'nine' (which seem to play the grammatical and inferential roles of quantifiers), with complex combinations of '∀' and '∃' statements, GQ theory introduces symbols for them in the formal language. For instance, it represents the quantifier 'three' by requiring that (at least) three objects fall within the intersection of two sets X and Y in order for 'Three F's are G' to be true. The result is that GQ contains a logical element for each numerical expression in the natural language that can modify a count noun. Again, however, if GQ is capable of capturing all relevant inferential properties, no *a priori* reason exists for resisting introducing additional logical items (with their additional rules of inference – a technical topic we do not need to discuss here).

Advocates of GQ can, then, agree with Russell (in opposition to Frege) that definite descriptions are best represented as quantifier phrases, yet disagree that their best symbolic representation is given by anything like (DD). The definite article 'the' in GQ is symbolically represented by its own quantifier, which for ease of translation we might represent by the symbol '[The x]'. '[The x: Fx] Gx' is true just in case exactly one object lies in the intersection of the 'F' and 'G' sets. This end result is similar to Russell, for both systems treat phrases of the form 'The F is G' as being true just in case there is exactly one thing which is F and it is also G;[7] but the GQ theorist can obtain this same semantic result without treating the natural language phrase 'the' as possessing a complex, multiply quantified logical form.

To recap: we now have three distinct proposals for symbolizing sentences with definite descriptions: the Fregean treatment, in which they are handled as akin to sentences with proper names; the Russellian analysis where they are treated as combinations of universally and existentially quantified claims; and GQ, where the definite article is treated as a quantifier phrase, which requires a common noun to be complete, and which maps on to its own unique element in the formal language. The question now is: 'how do we decide between all these alternative accounts?'

Recall, first, our initial adequacy constraint on symbolic representations: viz., that they capture logically valid inferences involving the expression in question. One way of understanding the objections Russell leveled at Frege's account of definite descriptions, then, is that the latter's proposal fails this constraint (i.e. there are logically valid inferences Frege's notation fails to capture by virtue of symbolizing definite descriptions as singular terms). For instance, in 'Everyone wants John,' the quantifier expression 'everyone' is its subject, 'John' its object, and 'wants' its transitive verb. This sentence is unambiguous, having only one possible translation into the formal system of predicate logic. The sentence 'Everyone wants the winner,' on the Fregean assumption that 'the winner' is a referring term, ought then to be unambiguous as well. Both should be symbolically representable in predicate logic as 'Rab.' But the definite description sentence is ambiguous; it has two readings, one in which there is a particular person everyone wants, and another where everyone wants whoever has the property of being the winner, regardless of whom he or she turns out to be.[8]

The difference between these readings is sometimes indicated by saying that in one the description takes *wide scope* over the rest of the sentence, and in the other it takes *small scope*. This feature of definite descriptions – that they enter into what we might call 'scope ambiguities' – likens them more to quantifier expressions (since it is a hallmark of expressions containing 'all' or 'some' that they display scopal ambiguity) and less to singular referring terms. Indeed, if we symbolically represent them as singular referring terms (as Frege did), we have no way to explain this logical phenomenon.[9] In short, Frege's treatment of definite descriptions is flawed; to capture all the inferential properties of sentences with the expression 'the F' we need to assign it more structure than the Fregean analysis does. Thus a quantificational theory of descriptions is preferable over the Fregean approach; but what are we to say about the debate between the Russellian and GQ theorist? Since the two approaches *agree* about inferential properties expressions of the form 'The F is G' possess, their disagreement cannot emerge from the failure of either approach to accommodate such inferential properties. Instead, it seems the GQ theorist assumes that it is permissible to invoke wider features of our symbolization to decide between competing approaches. That is to say, GQ theorists object to the Russellian approach to definite descriptions on the grounds that Fregean logic is inadequate for formalizing natural language as a whole. To see why the advocate of GQ might think this, we need now to take a slight diversion through the analysis of quantifier phrases, before returning again to the issue of definite descriptions.

An initial point GQ theorists have pressed in their favor is that other expressions in natural language look intuitively to be playing the same logical role as 'all' or 'some' (or 'the'), but provably resist analysis in terms of the primitive Fregean quantifiers '$\forall$' and '$\exists$'. The problem is that, although any quantifier making a specific numerical claim can be logically captured by a complex construction of Fregean quantifiers, some quantificational elements in natural language make no such claims. Consider 'many,' 'most,' and 'few'. These quantifiers are like traditional Fregean quantifiers inasmuch as sentences like 'All men are mortal' and 'Most men are mortal' seem to share grammatical makeup, and convey general claims about the extension of certain properties, rather than making referential claims about a particular individual. Furthermore, both apparently display the same kinds of ambiguity in linguistic contexts when nested inside other quantifiers. 'Every boy loves many girls' is ambiguous between there being one single privileged set containing many girls which are loved by all boys, and it being the case that, for each boy, there are many girls he loves, though each boy may love a different set of girls. Since these expressions intuitively seem so much like those expressions that Frege originally chose to symbolically represent as quantifiers, why not treat them as such? But how can we accomplish this end armed only with '$\forall$' and '$\exists$'?

To see the problem that the Fregean system faces, let's run through its options for an expression like 'most'. First, we might try representing 'Most girls are happy' with either (6) or (7), thereby equating 'most' with one of the two existing quantifier phrases:

(6) $(\exists x)(\text{Girl}(x) \,\&\, \text{Happy}(x))$
(7) $(\forall x)(\text{Girl}(x) \supset \text{Happy}(x))$

(6) states only that some girl is happy, and (7) that all girls are happy, and neither of these is what we need. (6) doesn't even logically imply the 'most' statement, and the 'most' statement does not logically imply (7).

Alternatively, we might try representing 'most' as expressing a specific numerical claim, since we know that expressions making these sorts of claims can be captured by complex combinations of '∀' and '∃'. Perhaps 'most' tells us that some specific number of happy girls is greater than the number of unhappy girls; for example (8).

(8) $(\exists x)(\exists y)((Girl(x) \,\&\, Happy(x)) \,\&\, (Girl(y) \,\&\, Happy(y)) \,\&\, x \neq y) \,\&\, (\exists z)((Girl(z) \,\&\, \neg Happy(z)) \,\&\, (\forall w)((Girl(w) \,\&\, \neg Happy(w)) \supset w = z)))$

(8) states that there are at least two happy girls and only one unhappy girl; but intuitively, our original sentence does not logically imply (8). (8) provides a circumstance in which our original sentence would be true, but it does not adequately logically capture what the original sentence means (after all, 'Most girls are happy' would also be true if five girls were happy and one unhappy, and in countless other situations as well).

So, neither '∀', nor '∃', nor some combination of them, seems adequate for capturing 'most'; but now we are in a position to see that the problem lies not merely in our limited range of quantifiers, but in the very form that Fregean quantifiers take. The problem is that in order to logically represent 'most' correctly we need to see it as having an intimate connection to the common noun it appears concatenated with (i.e. 'girls' in 'most girls'). Unlike with 'all' and 'some,' we cannot simply 'hive off' the quantifier expression for analysis (as the Fregean system does) and see it as binding a variable which then appears in predicate assignments, tied together by one of our sentential connectives. We can see that this is so by allowing the advocate of predicate logic to introduce a brand new quantifier expression, to add to '∀' and '∃'.

Let's use the symbol 'Σ' and simply stipulate that it stands for 'most'. However, although we are extending the Fregean system by one new quantifier, we will retain the general picture of how quantifiers and predicates relate; that is to say, 'Σ', like '∀' and '∃', will be a unary (free standing) quantifier. So with 'Σ' we can construct the following kinds of formulae:

(9) $(\Sigma x)(Girl(x) \,\&\, Happy(x))$
(10) $(\Sigma x)(Girl(x) \supset Happy(x))$

The problem with this suggestion is, first, that (9) states that 'Most things (in the world?) are happy girls', a sentence which is false just in case girls are not the largest set of objects in the world; so (9) seems an incorrect analysis of our original sentence. Sentence (10), on the other hand, states 'Most things are, if girls, then happy', and the logical rule for conditional statements tells us that if its first claim (its antecedent) is false, then the whole 'if . . . then . . .' claim will be true (regardless of the truth or falsity of the second claim, the consequent). Yet the antecedent in (10) will be false on almost all occasions, for what it claims is that given most objects, they are girls. So, if the majority of objects are not girls, this is sufficient to falsify the antecedent claim, and this in turn is sufficient to make the whole conditional claim true. So (10) turns out to be true

just in case girls do not form the majority of objects in the domain; on this construal, 'Most girls are unhappy' turns out to be true as well!

The problem with both (9) and (10) is that they issue in claims of truth or falsehood based on considerations about the wrong sets of objects: (9) is false and (10) true just in case there are less girls than boys and boats and trains, etc., all combined. Yet we wanted a much more specific condition for the truth or falsehood of our original claim, viz., that more *girls* be happy than unhappy.

What the failure of (9) and (10) demonstrates is that we *cannot* symbolically represent 'Most girls are happy' as containing two acts of predication, bound together by a sentential truth-functional connective, and concerning a variable previously bound by a distinct quantifier. Instead, what we need is to represent one predicate as an ineliminable part of the quantifier expression itself. Suppose we treat 'most girls' as an indissoluble unit that binds a variable *then* available for the predicate assignments 'are happy.' Then we can formulate a sentence like 'Most girls are happy' as: [Most (x): Girls (x)] Happy(x)', which yields precisely the interpretation we were after – it tells us that, given the set of girls, the majority of this set are happy. However, to adopt this kind of proposal is precisely to reject the Fregean form of quantification for sentences involving 'most,' in favor of something like the GQ proposal which treats quantifiers as binary expressions (i.e. as requiring both a quantifier phrase, like 'most,' *and* a common noun to yield a complete expression).

Returning, finally, to the central debate about definite descriptions, we are ready to draw a moral for logically representing these expressions. Advocates of GQ argue that since English has expressions which logically play the same role as straightforward quantified noun phrases, and yet which *cannot* be successfully formalized using either '$\forall$' or '$\exists$', combined with various sentential connectives, we must reject the Fregean system of quantification as inadequate for capturing logical inferences in natural language. Since *some* intuitively quantified expressions in natural language require a non-Fregean system of quantification, the conclusion drawn is that *all* quantified expressions in natural language require a non-Fregean system.

In effect, the GQ theorist is assuming that our original constraint on an adequate formalization (viz., that it capture inferential properties of a sentence) is insufficient. In addition, the formalization must belong to a formal system adequate for symbolizing other natural language expressions of the same type. There remains the question of how to spell out the notion of 'same type,' but as a first approximation, we might appeal to similarity in grammatical distribution and inferential properties (such as whether or not the expression can be concatenated with a common noun to form a larger phrase, and whether or not the expression gives rise to scope ambiguities in suitably complex contexts, like those containing other quantifiers or intentional verbs). Because the Russellian symbolization of 'The F is G' uses a logical system inadequate for expressions of the same type, like 'Most Fs are G,' it is held to be inadequate *simplicities*, despite capturing all the logical inferences definite descriptions support in natural language. The GQ analysis of definite descriptions is therefore alleged to be preferable over its Russellian competitor, because GQ is judged preferable over the Fregean quantificational system which at most adequately treats '$\forall$' and '$\exists$'.

Note that, if we accept this line of argument, a traditional and persistent objection to Russell's theory of descriptions actually carries over to the formalization of all quan-

tified claims in predicate logic. This objection is that the Russellian theory 'butchers' surface grammatical form, seeing an apparently simple sentence like 'The F is G' as possessing a vastly complex underlying content involving two distinct acts of quantification, a conditional, a conjunction and an identity claim. Yet this apparently runs counter to our intuitions about the grammar and structure of the original sentence. The GQ analysis avoids this worry, positing a simple underlying logical form; but interestingly it also suggests that this traditional objection can be leveled against other logical form claims. The GQ theorist challenges us to explain why, since we cannot represent 'Most girls are happy' as containing a conditional or conjunctive element, we should represent 'All girls are happy' and 'Some girls are happy' as containing a conditional or conjunctive element.

In response to this kind of attack, advocates of the Fregean system might question the crucial GQ assumption: why should we accept that a symbolization in a logical system, L_1, for a sentence, s_1, of a natural language, N, is adequate only if L_1 is capable of symbolizing all sentences of the same type as s_1 in such a way that the inferential properties of those sentences are preserved? The Fregean who rejects this assumption might recommend that we hold on to the predicate logic analysis for 'all,' 'some,' 'the' and other counting quantifiers, and employ a GQ analysis only for 'non-standard' quantifiers like 'most' and 'few' – ones that are probably not definable in terms of the notation of the others.

One counter-response to this, of course, is that if we had *started* our formalization of quantifier phrases by concentrating on expressions like 'most' we would have needed a GQ-type analysis from the outset, and this would have rendered the Fregean treatment otiose, since we could have handled all quantifiers within a single system of notation. Thus, the proposal that we adopt two different systems of quantification to handle the class of quantifier phrases in natural language might seem to go against some quite general philosophical principle, such as posit only the minimum set of explanatory items needed to explain the data. However, the issue here is perhaps not as settled as the GQ theorist presumes: there are technical costs involved in moving from the Fregean system to the richer GQ system (which we cannot explore here), and there may still be reasons that the Fregean can bring to light to license special treatment for 'all' and 'some.'[10] At the very least, we should note that the general philosophical principle appealed to above cuts both ways – also telling in favor of the more austere two-quantifier Fregean system, as against the 'quantifier profligacy' of GQ.

So, which approach should we adopt here? Who has got the constraints on symbolic representations right? We have seen that a minimum condition on an adequate formalization for a natural language expression is that it capture all the inferential properties of that expression; a stronger condition is that the logical language be adequate for capturing all the inferential properties of expressions of that type; and a (perhaps) maximal condition is that the logical language be adequate for capturing all the inferential properties of all the expressions of that language. Which of these conditions of adequacy we choose to accept, and how we see them as playing out in practice, will help us decide which kind of logical representations we accept.[11] However, we might begin to think now that perhaps we can simply sidestep this entire debate: for why can't we simply allow that a natural language sentence like 'the F is G' has *multiple* adequate logical representations? Why should we presume that there must be, in the end, just

one single 'ideal' notational language which, in some sense, 'really' gives the logical form of natural language sentences? Considering these questions takes us on to the issue of how we should construe the relationship between logical form and natural language.

2 What is the Relationship between a Natural Language Sentence and its Formal Representation?

The debate in the previous section between Russell's theory of descriptions and a GQ analysis seemed so important because of an implicit assumption that one or other of these accounts (or, perhaps, neither) gave the unique correct logical form for the natural language expression. Indeed, this was precisely the assumption made by Russell, who held that the formal language of *Principia Mathematica* was the unique, ideal formal language within which to reveal the true underlying logic of our language.[12] It is because of this kind of assumption that the need to choose between formally equivalent representations (like the quantificational account of definite descriptions, given by Russell, and GQ) seemed so pressing. But perhaps the assumption is mistaken; indeed, it is explicitly rejected by the twentieth-century American philosopher and logician W. V. O. Quine.

Quine claims that the sole purpose in symbolically representing a natural language sentence in a regimented language is "to put the sentence into a form that admits most efficiently of logical calculation, or shows its implications and conceptual affinities most perspicuously, obviating fallacy and paradox" (Quine 1971: 452). There will be different ways of doing this even with the same system of representation – since any logically equivalent formulation will do, and there will be infinitely many such sentences. Consequently, talk of *the* logical form of a natural language sentence even within a single system of symbolic notation might be misguided. The American philosopher Donald Davidson, following Quine, sees logical form as relative to the logic of one's theory for a language (see Davidson 1984: 140).

This kind of approach avoids the central worry we have been pressing, for we have no compunction to treat 'All men are mortal' as in any sense *really* containing a conditional. It just so happens that one symbolic representation adequate for capturing inferential properties of this sentence treats it in this way. Yet a liberal approach to logical form faces its own problems; for there is one crucial aspect of our language that creates a serious worry for anyone who believes that the notation we adopt in logically regimenting our language is more a matter of taste than fact.

The productivity of natural language

One key aspect of natural language so far ignored but relevant to any question of admitting multiple logical forms (for a single sentence) is that natural languages have no upper bound on their number of non-synonymous expressions. This is because they abound with constructions that generate meaningful complex expressions out of simpler ones. Grammatical sentences can be formed in English by concatenating two

sentences with either 'and' or 'or'; for example (13) and (14) are concatenations of (11) and (12):

(11) John left.
(12) Mary stayed.
(13) John left *and* Mary stayed.
(14) John left *or* Mary stayed.

Our language also exploits relative clause construction to create complex expressions from simpler ones. For example, new definite descriptions can be devised from old ones by adding restrictive relative clauses on head nouns, as in (15)–(17).

(15) The man left.
(16) The man *whom* I met yesterday left.
(17) The man *whom* I met yesterday *who* was eating breakfast left.

Though we can list only *finitely* many members of any of these various classes of grammatical English constructions, a casual look should convince you that each is unbounded. New sentences are formed by conjoining old sentences; new descriptions by adding relative clauses on the head nouns of old ones. These are but a few of the devices that render our language limitless.

Obviously, after performing these operations several times, say, conjoining a few sentences or relativizing a few clauses, speakers inevitably fail to comprehend the products. This is not a feature of English, however, but merely of how our minds and memories are organized. Suppose that English speakers cannot comprehend sentences with more than seven relative clauses. Would it follow that sentences with eight relative clauses are ungrammatical? Not at all. If increased memories or processing powers allowed us to understand eight clause sentences, this would merely enhance an already intact linguistic competence.

The relevance of unbounded classes to our current debate is this: since members of each of these infinite sets of sentences stand in indefinitely many distinct logical relations to one another, no mere (finite) list can correctly or adequately complete the task of codifying the set of logical inferences of a natural language. There are too many inferences. Therefore, a theory must be devised about how to symbolically represent them, in order that various inferential relations are 'captured' correctly. Once we see how such theories are devised, we see why, though there may be indefinitely many logically equivalent formulations for any single natural language sentence, some are preferable over others – and not *just* on pragmatic grounds of overall simplicity.

First, note that it's no accident that both (18) and (19) are contained in (20).

(18) John left.
(19) Mary stayed.
(20) John left *and* Mary stayed.

Indeed, it's the non-accidental occurrences of (18) and (19) in (20) that accounts for their logical inter-relatedness. We see this in asking what is it about (20) in virtue of

which it logically implies (18) and (19)? The answer is that (20) is true as a matter of meaning alone in English just in case its components, (18) and (19), are true. Indeed, generally, sentences devised from other sentences with the word 'and' have this logical property as a matter of meaning alone. An excellent candidate that explains this logical relation between conjunctions of sentences and their simpler components is meaning rule (A).

(A) A conjunction is true just in case its conjuncts are true.

With enough such rules, we can show how every complex expression bears logical relations to simpler ones (and vice versa), relative of course to logical relations among their simpler expressions, say, down to primitives – where primitive expressions stand in no logical relation to one another.

As noted above, complex nouns can also be constructed out of simpler nouns and relative clauses. From a primitive expression like 'man,' a complex expression like 'man who loves a woman' can be formed, which in turn can be used to form a more complex expression like 'man who loves a woman who hates a dog' and so on. Rules are required to show how logical properties of such complex expressions are determined by those of simpler nouns and relativizations on these nouns.

To figure out what such a rule might be, consider the primitive 'man' and the more complex 'loves a woman.' The relative pronoun 'who' can grammatically conjoin these expressions. How is the logical role of the complex 'man who loves a woman' predicted from whatever logical roles its component parts might have? The complex expression 'man who loves a woman' is true of an individual just in case that individual has its simpler components 'man' *and* 'loves a woman' true of him as well. A perfectly fine rule, then, that enables us to project from the logical roles of simpler components to those of the complex expression built up by relativization is meaning rule (R).

(R) A construction of the form – X who Y – (where X is a noun and Y is the rest of the relative clause prefaced by 'who') is *true of* an individual just in case both X and Y are true of this same individual.

It is both interesting and surprising how much (R) resembles (A). Like (A), (R) also sees the components of complexes as making a conjunctive contribution. Conjunction by itself explains that complex relativizations are true of something just in case their components are as well, and this biconditional rule suffices to explain the logical relations between the complex relativization and its simpler components.

When the project of symbolically representing natural language sentences into a formal notation is seen from the perspective of the unboundedness of natural language, and therefore, the unboundedness of inferential relations among sentences of natural language, the idea of allowing all logically equivalent notations to be employed in representing the same sentence becomes harder to swallow. Codifying known inferences might be a project independent of any particular choice of formal notation, but our logical representations must *also* make the right projections and predictions for inferential relations. For this to be possible, we need a tighter connection between natural language sentences and their formal representations than mere codification. To have

any hope that a system of symbolic representation will work, we need to assume that natural language sentences *actually possess* some kind of formal structure on the basis of which we can project out to the explanations of the inferential properties of novel linguistic items.

So, it seems that the laissez faire approach to logical form, which sees it merely as a tool of codification, encounters difficulties when confronted with productivity. If we are to be able to account for the limitless nature of our language, it seems that we must posit structure inherent within natural language sentences, and the intriguing proposal is that these same structures can account for the logical properties of the sentences of the natural language. This realization reinstates the seriousness of the debate between opposing accounts of the logical form of definite descriptions which closed the first section of this chapter: it is not, it seems, sufficient for us simply to admit multiple adequate logical representations for a sentence of the form 'The F is G,' for we need to know which form captures its inherent logical form.[13] This in turn brings us back to the question of which constraints are correctly placed on our choice of logical form for a sentence and whether such constraints will guarantee a unique logical form for each sentence of natural language. Although we haven't answered these questions in this essay, we do hope to have shown why it is important to ask them. As noted earlier, the notion of logical form has become commonplace in the philosophical arena at the turn of the twenty-first century; however, if symbolic logic is really to advance our understanding of language, we need to be very clear from the outset about the relationship envisaged between the two realms.

Notes

1 Exactly how this replacement takes place will vary, however, depending on the logical system in play; for instance, propositional logic will replace each whole proposition with a schematic letter, whereas (as we will see below), a system like predicate logic will introduce schematic letters for sub-propositional elements.

2 This argument form is so common as to have a special name: *modus ponens*.

3 Referring terms were to be handled by Frege's notions of *sense* (the mode of presentation of an object) and *reference* (the object), see Frege (1879).

4 Russell (1905).

5 Of course, we need to take into account context as well. If someone says, 'The man left', what he said might be taken to be true in a context, say, where there are two women and only one man, even perhaps some small children.

6 Higginbotham and May (1981), and Barwise and Cooper (1981).

7 Not all logical representations of definite descriptions agree on the claim of uniqueness; cf. Szabo-Gendler (forthcoming).

8 We might think of the difference as turning on whether or not the property of 'being the winner' figures essentially in the agent's wanting.

9 We might also wonder how, on a Fregean analysis, we capture the valid inference from the truth of 'The man who broke the bank at Monte Carlo died' to the truth of 'Some man died.' (Replacing 'broke the bank at Monte Carlo' with any other meaningful complex noun will also underwrite the inference, and so it's an inference we want to accommodate in virtue of logical form.)

10 The technical issues concern the fact that certain properties of Frege's logical system (i.e. 'closure' and 'completeness') are lost in the move to GQ; these are quite complex technical properties that need not concern us here.

11 The latter point matters, for the advocate of GQ *could* argue that our first and second conditions collapse with respect to quantifier phrases, since sentences like 'any girl is happy' and 'some girl is happy' seem to be logically connected (the former apparently entailing the latter), and this fact could prove difficult, if not impossible, to accommodate given different methods of representation for the two sentences. Again, however, the details of this debate go beyond our present concerns.

12 A similar assumption was also made by the early Wittgenstein, though he remained agnostic on the choice of ideal language, merely holding that natural language items actually possessed some particular logical form, which would be revealed by a process of logical analysis.

13 Of course, the discussion about definite descriptions is meant as a paradigm example of the issues discussed here, not as the only case of them. For another particularly clear example, consider prepositional phrase modification and the apparently valid move from 'John buttered some toast in the kitchen' to 'John buttered some toast.' The question is: what kind of logical form might capture this inference and will the form suggested be acceptable as the genuine underlying logical form of the sentence? One suggestion for the logical form of such sentences is Davidson's 'event' approach (see Davidson 1967), which posits an 'extra place' in the logical form for an event variable; but a common objection to such an approach is that it diverges too far from the surface form of the original sentences. The debate is thus parallel to that had in the text concerning definite descriptions and the conditional representation of 'all' statements: we have approaches which predict the right inferential relations, but which may be deemed unsuitable as the 'real' logical form of the sentence in question due to divergences from surface form.

References

Barwise, J. and Cooper, R. (1981) Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4, 159–219.

Davidson, D. (1967) The logical form of action sentences. In his *Essays on Actions and Events*, 105–21. Oxford: Clarendon.

Davidson, D. (1984) *Inquiries into Truth and Interpretation*. Oxford: Clarendon.

Frege, G. (1879) *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle: Niemoyer.

Frege, G. (1892) Uber Sinn und Bedeutung, translated as 'On Sense and Reference' in P. Geach and M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell (1952), 56–78.

Higginbotham, J. and May, R. (1981) Questions, quantifiers and crossing. *Linguistic Review*, 1, 41–79.

Quine, W. V. O. (1971) Methodological reflections on current linguistic theory. In D. Davidson and G. Harman (eds.), *Semantics of Natural Language*. Boston, MA: Reidel, pp. 442–54.

Russell, B. (1905) On denoting. *Mind*, 104, 827–62.

Strawson, P. F. (1952) *Introduction to Logical Theory*. London: Methuen.

Wiggins, D. (1980) "Most" and "All": Some comments on a familiar programme, and on the logical form of quantified sentences. In M. Platts (ed.), *Reference, Truth and Reality*. London: Routledge, 1981, 318–46.

Whitehead, A. N. and Russell, B. (1927) *Principia Mathematica*. Cambridge: Cambridge University Press.

Further Reading

Davies, M. (1981) *Meaning, Quantification and Necessity*. London: Routledge.

Evans, G. (1976) Semantic structure and logical form. In Evans and J. McDowell (eds.), *Truth and Meaning*. 199–222, Oxford: Oxford University Press.

Evans, G. (1977) Pronouns, quantifiers and relative clauses (I). *Canadian Journal of Philosophy*, 7, 467–536.

Larson, R. and Segal, G. (1995) *Knowledge of Meaning* Cambridge, MA: MIT Press.

Mates, B. (1973) Descriptions and reference. *Foundations of Language*, 10, 409–18.

Neale, S. (1990) *Descriptions*. Cambridge, MA: MIT Press.

Peacocke, C. (1975) Proper names, reference, and rigid designation. In S. Blackburn (ed.), *Meaning, Reference and Necessity*. 109–32, Cambridge: Cambridge University Press.

Sainsbury, M. (1991) *Logical Forms*. Oxford: Blackwell.

Sainsbury, M. (1996) Philosophical logic. In Anthony Grayling (ed.), *Philosophy: A Guide through the Subject*. 61–122, Oxford: Oxford University Press.

This page intentionally left blank

Part III

PHILOSOPHICAL DIMENSIONS OF LOGICAL PARADOXES

This page intentionally left blank

7

Logical Paradoxes

JAMES CARGILE

Logical paradoxes centrally involve difficulties in determining truth values. But not all such difficulties are paradoxes and not all paradoxes are paradoxes of logic. Considerable trouble can be taken in trying to delineate the right subclass. But the questions about truth value are often more interesting. It is better to begin by trying to answer some notable problems of this sort, even at the risk of contributing to some subject other than logic.

One such case, is the Eubulidean Liar (UL), attributed to Eubulides of Megara, who is supposed to have said "What I am saying is false." This could be an unproblematic assertion made by a spy to an assistant as an aside about some item of misinformation he is in the process of sending out in a broadcast. But when understood with a certain kind of 'self-reference' the remark would be absurd. It was never offered as a sincere effort at communication, but as a way of presenting a problem for rationalistic philosophy. This problem may be better understood in the following version. The sentence

(A) The sentence A is not true,

is one which it seems could not be true. For if it were true, it would seem to follow that it is not true. But if we conclude from this that the sentence A is not true, then it seems this could not be right, being the very same words as the sentence A itself, which would seem to suggest that A is true after all. This has led to the suggestion that allowing truth to be attributed self-referentially, as in 'No proposition is both true and false' should be somehow avoided or restricted.

The most reasonable way to follow that idea would be to deny that there is any such property as truth. There is obviously an English predicate '. . . is true' and the grammar of A is unassailable. We can grammatically assert the sentence and grammatically attribute to it its predicate. But this is no guarantee of asserting a proposition and attributing a property. Formal logic is primarily concerned with sentences and predicates. But philosophical logic must be concerned with propositions and properties. Both propositions and sentences can be asserted or said and both predicates and properties can be predicated or attributed or said of. But propositions and properties are more important for philosophy, which makes for problems in the assimilation of formal logic and its impressive results. The great precision and secure consistency of some systems

of formal logic may seem to point in favor of nominalism. But we can continue to value this precision while keeping open the possibility of a consistent use of propositions and properties in philosophy.

The primary connection between asserting or saying, and predicating or saying of, is the property of truth. We have Rule R1: To assert a proposition is one and the same thing as to predicate truth of it, and to deny a proposition is the same as predicating nontruth, which in application to a proposition is the same thing as falsity. To call a thing a nontruth is not to call it false. But to call it a nontrue proposition is to call it false, and that is how it is natural to understand calling a proposition nontrue – as merely short for 'nontrue proposition.' This fundamental connection does not apply at all in the case of believing. To believe that a proposition is true is not the same as believing it. One may believe that what Bill will say tomorrow is true without believing what Bill will say tomorrow, but to assert that what Bill will say tomorrow is true is to assert that thing, whether or not you know what proposition it is. To assert a proposition is to take a certain unique kind of responsibility for its being true. This is often done knowingly, but that an agent asserts a proposition does not in general entail that the agent believes it or even has the ability to understand it or know what proposition it is.

The ruling that A does not express a proposition seems to be good reason to conclude it is not true. But then it seems that is a true judgment about A which is expressed by A. To avoid that problem it would be tempting to conclude there is no such property as truth. A better response is to assume that A expresses some proposition. Whatever it is, A says it. But A says that whatever it says is not true. So by the above R1, A both says and denies the same proposition, saying whatever it says and that whatever it says is not true. This gives us an adequate basis for saying that A is not true.

It will be objected that this is just what A says. On the contrary, we cannot adequately say in full what it is that A says. It is not that what A says is that A is not true and also that what A says is that it is not true that A is not true. It is just that these are equally good representations of what A says. That is good enough to show that whatever A says is contradictory. It is not required that those equally good representations, which are as good as we can get, are good enough to warrant either one being taken as making it clear to us what A says. A is one of many counterexamples to the idea that to identify the proposition expressed by a sentence S it suffices to write 'the proposition that . . .' followed by the sentence S. Thus it would be quite wrong to think that 'The sentence A is true as a sentence of English if and only if the sentence A is not true' is licensed by a correct rule for describing the content of a sentence of English.

This may be reinforced by considering

(A′) The sentence A′ is true.

We can say that what the sentence A′ says is that what it says is true. That would indeed be, by R1, to say whatever it is that A′ says. That does not tell what A′ says or offer any ready guide as to what its truth value would have to be. If we rule that A′ says nothing, then we should treat anyone who claims that A′ is true as speaking falsely. He might have used the same words as A′. So why not count A′ as false also?

If a man says 'What I am now saying is true' (when it is clear there is no other reference) then he cannot be serious and we are right to rule that he has not asserted any

proposition. Having so ruled, we must hold that one who says 'What he said then was true' (when the reference is clear) is speaking falsely. This can be explained by appeal to the thoughts expressed. When we treat sentences by themselves as doing the saying, matters cannot be clarified by that means. It might be suggested that we should not treat sentences in this way, but that is not a practical possibility. It is often important to ask, not what the author of certain words, such as a constitution, intended, but what has been said by them. Whether the founders of the USA intended their words to be incompatible with the institution of slavery or not, it is important that the words were not compatible with it.

Denying that A or A′ say anything seems to provide good reason to call A true and A′ false. This then leads to paradox. We do better to note that if A says anything, that thing is contradictory and thus false. Thus prepared, it is best to rule that A says something, albeit an obscure and worthless thing. If A′ says anything, we have no reason whatever to consider this thing to be false. To the extent that we have no reason either to consider it true, we must consider this a bad mark against the practice of taking sentences by themselves as saying things at all. But the consideration that A′ is doing absolutely nothing but endorsing whatever it is that it says may suffice as a reason for counting it trivially true. For every saying endorses itself, being equivalent to calling itself true. A′ may be taken to report this triviality about itself.

The Epimenidean Liar (EL) can be put as follows: we build a one room shed known as Building B, working in total silence. One of us then goes in and asserts

(B) Nothing true is asserted in Building B at any time

and nothing else. We then burn Building B to ashes. It seems that B cannot be true, since it was asserted in Building B and having something true asserted there would make B false. But if it is not true that no truth is asserted in B then it seems to follow (since something has been asserted) that some truth has been asserted in B. Since this cannot be B, and given the history of the building, we might then have to conclude that elves or similar beings slipped in and made at least one false assertion while the building was there to house such assertions. But this is hard to bear. We seem to have to avoid contradiction only by accepting a preposterous factual claim.

Here we appeal to the principle R2: that to assert that all Xs are Ys is to predicate being a Y of every X. Thus to assert B in the building is to predicate nontruth of everything asserted in the building. That is to assert that it is not true that everything asserted in B is nontrue in the very course of asserting that everything so asserted is nontrue. So the assertion of B in the building is false. Our assertion out of the building is more fortunate, since it is not among the assertions it is calling nontrue.

It would not be clear to say that our assertion of sentence B is not self-referential while the in-building assertion of it is, due to ambiguity concerning 'self-reference.' All assertions are self-referential in the sense that to assert any proposition P is to assert that everything whatsoever (including P) is such that P. (That (x)P is equivalent to P has been questioned for sentences of predicate logic in the 'empty domain' case, but this is an extremely eccentric system which should not influence our considerations.) In that sense, both the in-building assertion of B and our assertion of it are self-referential. On another interpretation of 'reference' we do not say that 'All Fs are Gs' is

about itself unless it is an F. The in-building assertion of B says that if it is an in-building assertion then it is not true. In that sense it is self-referential, while our judgment about B is not an in-building assertion, so that in this other sense of 'reference' it is not self-referring.

The present line, that sentences such as B serve to reject all assertions of a given kind, has been advocated by some thinkers who go further, to hold that such sentences do not convey any additional assertion beyond each of those denials. Similarly, it is said that to say 'Everything he says is true' is merely to endorse everything he says and not to say anything further. This idea has the consequence that there would be no difference between asserting B in the building or outside it. Similarly, if someone asserts 'Every Cretan assertion is false' it would be irrelevant to the assessment of this performance whether it was put forward by a Cretan. This is quite implausible. The paradoxes are best answered, not by economies about meaning, but by paying attention to the full meaning.

We may contrast problem cases superficially similar but involving belief rather than assertion. Suppose that we build a Building C on exactly the lines of B, except that the only person to go in does not know what building he is in. He believes (never mind how) that Building C has been so constructed that any carefully considered belief held therein is false. While waiting there to be called upon to leave, he reflectively believes (as opposed to his unarticulated assumptions that he is in a building, clothed, etc.) only that

(C) Any beliefs reflectively held in Building C are false.

If that were the only such belief in room C then we would appear to have a situation similar to the EL. But the principle R2 is obviously false for beliefs. To believe that all C beliefs are false is not to believe that it is false that all C beliefs are false. However, to believe that all C beliefs are false (as opposed to merely believing such a thing as that the sentence C expresses a truth) is to believe that you are not in building C. This is not a separate belief, but rather, part of what it is to believe that all C beliefs are false. To assert that all C beliefs are false is to predicate the falsity of every belief reflectively held in Building C. The sincerity of that assertion would mean believing, not everything that is thereby asserted, but only that all C beliefs are false. The assertor would unknowingly predicate falsity of the belief he expresses in making the assertion. An assertor could be fully informed as to what he is doing in asserting C, but then in order to persist he must be insincere. To believe that all C beliefs are false is to believe among other things that that belief is not one reflectively held in Building C. It is not possible to believe *directly* (a notion which cannot be clarified further here) that your very belief is false (or that it is true). There is no belief analogue to the Eubulidean Liar. The indirect cases, such as C, always involve more content in the belief than is assumed in the formulation of apparent conflicts about truth value.

Two innocent but logically acute persons might be conversing about C, one of them standing in the yard of Building C, the other speaking to him from inside that building, neither one knowing that building to be C. They could both believe that all C beliefs are false and have essentially the same belief. It would be a belief including the mistaken thought that the building housing a party to their conversation is of course not the

Building C which they are conversing about. But if they asserted their common belief, they would, unbeknownst to either, make different assertions.

Suppose that Bill declares that P, Q, R, S, and T, and Bob and Ben hear his declaration and agree that everything that Bill declared is true. Bob remembers well that Bill said that P, Q, R, S, and T, while Ben is unable to recall just what Bill said. Here we could have many degrees between having only a confidence in Bill and not knowing at all what he said, perhaps even disagreeing with those propositions while ignorant that they are what Bill said, to being in Bob's state of complete understanding. To say that Bob and Ben agree on the proposition that everything that Bill said is true badly underdescribes this situation. We might best describe Bob as believing that Bill's declaration was that P, Q, R, S, and T and that as a matter of fact, P, Q, R, S, and T. Bob does agree with Ben that 'everything Bill said is true.' But what this belief consists in would need to be worked out in dialogue between them, so that Bob's belief is reduced to Ben's or Ben's expanded to Bob's or to some intermediate compromise.

It is only in successful dialogue that we achieve a good understanding of what is believed. A proposition is essentially something which can be conveyed to others in successful dialogue, or a compound of such things. (It is the compounding that allows for unbelievable propositions of various kinds.) Logic generalizes about these things, and in thus stepping back from specific dialogue, loses track of the identity of the propositions and treats merely of sentences. In extreme cases we have sentences such as A or A′, which could never be used, in their logically problematic roles, in good dialogue, as expressing contributions to the dialogue, though they can of course be objects of dialectical discussion.

Besides the belief cases, there are puzzles in which someone fears that all his fears are unfounded or hopes that all his hopes are unfulfilled, etc. Believing, hoping, fearing, and asserting are all things done by people, but the former are attitudes, while asserting is not. Finding common 'propositional objects' and restricting 'self-reference' or rejecting a 'global truth predicate' or employing evaluation rules which do not assign truth values to all propositions makes possible a uniform treatment of these puzzles. But this is a costly uniformity which blurs important differences. It is important to note the distinction between cases which require talking with someone who either expresses an attitude or attributes one to someone else, and cases which involve just looking at the powers of a sentence by itself.

For example, when someone claims an odd belief, we need to talk with that person rather than making adjustments in logic. If a man claims to fear that all his fears are unfounded we need to know if he intends to express fear that, among other things, his fear that all his fears are unfounded is unfounded. If he does, then the problem is not for logic, but for those who think this could be sincere.

The Geach–Lob implication liar (IL) involves

(D) D materially implies that P.

It seems that if D were false, then it would have to be true (and P false). Since that is impossible, it seems that D is necessarily true, and thus that P is too. A suitable choice of P can bring out how bad this would be. This paradox for the material conditional is not essentially different from other paradoxes based on other truth functional connec-

tives. Here we appeal to the rule R3: that to assert that if P then Q and that P, in one assertion, is to assert that Q. Now the suitable choice of Q only yields a bad assertion, not a bad problem. When P = (2 + 2 = 4) that version of D is true. When P = (2 + 2 = 5) that version is false. Similar considerations apply to cases based on other connectives. (Our earlier case, A, could have been interpreted disjunctively, as saying that A either expresses no proposition, or expresses a false proposition, and that would require a principle for disjunction.)

The Grelling Liar (GL) involves the predicate 'heterological' defined as "a predicate which expresses [as a term in some systematic usage – expression cannot be analyzed here] a property of which it is not itself an instance." This self-reference seems to threaten both the saying that 'heterological' is heterological and the saying that it is not. The answer is that there is no such property as that of being a predicate which expresses a property of which it is not itself an instance, any more than there is such a property as being a property which is not an instance of itself. It is common for logicians to agree with this. But the crucial problem is to properly explain why it is so.

It is not that to say a term is heterological is to say nothing of it. It is rather, that it is not to say the same thing for every term. 'Heterological' expresses a property in application to 'obscene,' but the property is that of expressing the property of being obscene while not possessing it. In application to 'English' the property is that of expressing the property of being English while not possessing it. Even this much is often accepted. The question still remains as to why it is so. For one may of course assert, about a term whose meaning is unknown, that it is heterological. Why is this not merely to say that there is some property it expresses but does not possess? (It is widely held to be an important point of logic that to say that some F is a G is *not* to predicate being a G of any F.)

It is because R4: to assert that Some F is a G is to assert that if anything whatever is such that everything other than it is not an F which is a G, then that thing is an F which is a G. When we say that there is some property expressed by 'obscene' which is not possessed by it, we assert of each thing there is that if nothing other than it is expressed by 'obscene' then it is expressed by 'obscene' and not possessed by it. The one and only thing which satisfies the antecedent condition of this conditional predication is the property of being obscene, and so, for that reason, calling 'obscene' heterological is to predicate not being obscene of it. By contrast, the predicate 'heterological' fails to *uniformly* express any property. It always picks up its property from the term to which it is applied, so that when applied to itself, there is no property to pick up. For that reason it is not heterological – it does not possess a property it expresses, because it does not express a property. (If we define 'heterological' differently, as 'does not both express and possess a property' then it is heterological.)

Russell's Paradox involves the predicate (RP) 'class or set which is not a member of itself.' It seems that RP expresses a property, and since it is a truth of logic (call it the Abstraction Principle) that to every property there corresponds the class of all and only the things having that property, the Russell predicate, through expressing that property, determines a class which, it seems, can neither belong nor fail to belong to itself.

Cantor's paradox involves the same Abstraction Principle applied to the property of being a thing to yield (UC) the Universal Class. Cantor's Theorem says that every class is of lower cardinality than its power class (the class of all its subclasses) which implies that UC is of lower cardinality than its power class PUC. But this is incompatible with

the requirement that any member of PUC must of course, like everything else, belong to UC.

RP is logically similar to 'heterological.' To say (truly) that the class of men does not belong to itself is to say it is not a man. To say (falsely) that the class of classes is not a member of itself is to say that it is not a class. There is no such property as being a non-self-membered class and thus no such class. This does not at all impugn the Abstraction Principle. There are indeed other ways of forming classes than as the extensions of properties, but they are inadequate for the determination of classes on the scale of interest to mathematical study. Mental acts of attention can identify a class, but not a very big one. Some will appeal to the mental powers of God, but they are especially unsuited to the task of forming classes by mental attention, since God is equally and perfectly aware of absolutely everything. God distinguishes things not by paying special attention but by knowing what properties they have.

Versions of Cantor's Theorem in first order set theories are unassailable specimens of mathematical truth. But as a principle of philosophy, it is false. Its proof depends on an alleged class essentially the same as the Russell Class. It is assumed that there is a one-to-one mapping M between UC and PUC. Then it is held there would have to be a class URC of all elements of UC which have the property RUP of not belonging to their M-correlate from PUC. URC would have to be a member of PUC and have an M-correlate X in UC. Now X is a member of URC if and only if it is not a member. The situation is exactly like Russell's Paradox.

The answer should also be the same. There is no such property as RUP for the same reason that there is no such property as RP. If there is such a property as existing (which has, of course, been disputed), then the Abstraction Principle guarantees the existence (and self-membership) of UC. Among its peculiarities will be its isomorphism with its power class. This is more satisfactory philosophically than making it out to be a thing which does not itself belong to any class. It would be better to say that Cantor's Theorem only holds for 'sets' and that UC is thus not a set in that sense.

One modal liar, (ML) is

(E) The proposition E expresses is not a necessary truth.

It seems to be (contingently) true that the proposition that E expresses is that the proposition that E expresses is not a necessary truth. It seems that proposition could not fail to be true. For if it were not true that that proposition is not necessary, it would be necessary, which is incompatible with its not being true. But then, since it cannot fail to be true, it must be necessary. But that implies that it is false.

Here it is best to answer that what E, as a matter of contingent fact, says, is that what it says is nonnecessary. Since what it says is (at least) that it is nonnecessary, it must then say that it is nonnecessary that it is nonnecessary. But this is, in the broad sense, a contradiction, since it is necessarily true that if anything is nonnecessary, then it is necessarily true that it is nonnecessary. That is the characteristic axiom of S5. Its utility in this case is just one more indication of its truth.

Yablo's infinite liar (YL), in one version, involves the sentence form

(F) Every sign along The Path numbered n or greater expresses a falsehood,

where a sentence of that form is written on each of an infinite series of signs arranged in such a way that each one points in the same direction along The Path and is labeled with a number one less than the value of n in its sentence. (This has been held to have the consequence that not a one of the signs in this series is self-referential.) Now, if any one of the signs, X, were true, then every sign following it would be false. But then any sign Y following X would be such that every sign following Y was false, which would make Y true. Thus the truth of any sign X in the series entails a contradiction. So all the signs in the series would have to be false. But that seems to entail that each sign in the series would be true.

YL essentially involves the idea of a completed infinite series. If the signs were being produced one a day into the indefinite future, there would be no basis for trouble. A plain ordinary falsehood might turn up at any time, which would make the sentence immediately before it unproblematically true and thus in turn, all the sentences preceding it unproblematically false. This might seem implausible if the signs are being produced by a machine that merely puts up a duplicate sign a step down The Path each day, with nothing in the machine's repertoire to allow it to do anything else. But infinite time allows all sorts of things to happen. If it is added to the specifications that the machine is not going to break down, it will be a primary question whether this guarantee can be accommodated by a merely potential infinite. But if the complete infinite series could exist, the above argument for a paradoxical contradiction would apply.

This has been seen by some as showing that logical paradox of the Liar type does not depend on self-reference. As was observed above, this would depend on what is meant by self-reference. In any case, that question would not be important on the present approach, since this case of YL can be treated in the same way as EL. For any n, sign n attributes falsity to what is said by sign n + 1 and to what is said by sign n + 2. But n + 1 attributes falsity to what is said by n + 2. So n attributes falsity to what is said by n + 2 and also attributes falsity to that attribution of falsity. Thus sign n contradicts itself for each n. This is assuming the series is infinite. (If it is not, the result is different, but that need not be considered now.)

One Knower family paradox is (UK):

(G) No one knows that H is true.

This seems probably true by showing that the assumption to the contrary leads to a contradiction. And yet giving this proof somehow cannot qualify anyone as knowing H is true. Variations on this theme have been offered as *the* 'Surprise Test Paradox.' A teacher announces "There will be a test tomorrow and none of you know that this announcement is true." There are actually a number of candidates for paradox about 'surprise' events and some involve no announcement at all, just a known tradition of the 'teacher' being punished if he fails to spring a test which qualifies as a 'surprise,' and related arguments suggesting that he can (and cannot) succeed.

Paradoxes of the Knower family are not resolvable by the method used above for paradoxes of assertion and predication. Attributing having an unknown truth value is not like attributing truth or falsity or necessary truth or the like. These paradoxes need to be treated in a way similar to the belief case C above. They are cases requiring talking with alleged believers rather than cases which involve merely the logical powers of

sentences by themselves. Does the assertor of H really believe that he does not know that what he is saying is true? Consider

(G′) No one believes that G′ is true.

A foreigner could easily believe that G′ expresses a true proposition, thanks to being ignorant as to what proposition it does express. But is there a proposition G′ expresses to intelligent speakers of English, which none of them believe? If not, should we not then conclude that no one believes, with full understanding, that G′ is true? And then, does not that proposition appear to be, after all, what one with a full understanding of G′ would see it as expressing? And doesn't that put us in logical trouble? The mistake here is in thinking that a proposition that turns up in the course of a certain line of reflection on G′ could have been the one G′ by itself was expressing all along. No one can believe G′ in a certain self-referential way. One can use G′ to express this thought. But then, so used, G′ is being believed true. To take that as proof that G′ is false is to slip into a confused equivocation. It is not the property of truth that needs stratifying here, but the various thoughts that get associated with G′.

A more serious problem about the applicability of the present approach can be brought out by considering a case in which a dozen people are required by the law to make exactly one deposition in regard to a certain case, and each one of them deposes a token of

(H) Something deposed by one of the others is false.

This case can be presented without reference to people, in terms of sentences by themselves, so that our rules about assertion and predication should be the answer. We could arbitrarily stipulate that a certain one of these depositions is false. That would have the consequence that all the others are true, which works out nicely as far as consistency goes. But this arbitrariness is obviously unacceptable. Here the above rule R4, which is the basis for answering the Grelling paradox, is not adequate, because no one of the depositions is such that no other of the depositions is a false one.

R4 is also not applicable to a version of Yablo's paradox in terms of existential quantification, in which the signs read

(I) Some sign along The Path, numbered n or greater, expresses a falsehood,

with each sign, as before, labeled with a number one less than the value of n in its sentence. This is unproblematic if the series is finite, since the last sentence is then false, making its predecessors true. But the infinite series raises the problem that if any one of the signs were false, all its successors would have to be true, which is impossible, leading to a contradiction just as with G. However, the treatment which works for YL-G and EL does not work for YL-I. And R4 does not work either, for the same reason that it does not work for H.

It might be tempting to write off YL-I as just a paradox of the completed infinite. Whether such paradoxes are logical paradoxes is the sort of question set aside at the beginning of our discussion. It would depend on whether it is a truth of logic that there

are completed infinites. But we can continue to spare ourselves this question by noting the similarity between the problem of YL-I and that of the obviously finite case H. The inadequacy of R4 is the same for each.

Let us proceed directly to R5: To assert that some F is a G is to assert that if anything is such that nothing other than it is any better candidate for being an F that is a G than it is, then it is an F that is a G. R5 deals nicely with H. Every one of the deposers i has said of every other one of the depositions [{1,2, . . . 12} – i] that it is false, since every member of [{1,2, . . . ,12} – i] is equally qualified for being a deposition other than i which is false. But calling any one of these, j, false, is to attribute falsity to the claim that some one of [{1,2, . . . ,12} – j] is false which is to endorse all those claims, while calling all of them other than i false. Thus all the depositions are inconsistent and for that reason all are false. And that is not sufficient to make any of them true, because they have not claimed simply that one of them is false, as an external observer could simply claim.

The same goes for YL-I. For any sign n, all the subsequent signs qualify equally as candidates for a false subsequent sign and are thus, by R5, all called false by n. This makes each sign n contradictory just as in the case of YL-G. They do not claim simply that a subsequent is false, as an external observer could.

It will be objected that R5 does not sound at all like a logical rule, but more like something from ethics. The notion of being as good a candidate as there is, for being an F that is a G, will be held to be objectionably vague. It may well be vague in many cases. But in the two problem cases just considered it is perfectly clear. The problems in fact arose from the fact that it would be absurdly arbitrary to treat one candidate as a better case of an F that is a G than any among a set of others.

8

Semantical and Logical Paradox

KEITH SIMMONS

1 Introduction

Consider the following array:

	'monosyllabic'	'French'	'inanimate'	'infinite'	...
'monosyllabic'	f	t	f	f	...
'French'	f	f	f	f	...
'inanimate'	t	t	t	t	...
'infinite'	f	f	f	f	...
⋮	⋮	⋮	⋮	⋮	⋱

Down the side and along the top are the 1-place predicates of English, taken in the same order. In each box, we put t or f according to whether the predicate at the side is true of the predicate at the top. We obtain rows of ts and fs. For example, the row of values associated with 'monosyllabic' is: ftff. . . . Consider the diagonal of values from the top left towards the bottom right: fftf. . . . Observe that each f in this *diagonal sequence* corresponds to a predicate false of itself ('monosyllabic,' 'French' and 'infinite' are each false of themselves). Now form the *antidiagonal sequence* by changing each f in the diagonal sequence to a t, and each t to an f. We obtain the sequence ttft . . . , where now each t corresponds to a predicate false of itself. Notice that this antidiagonal sequence cannot occur as a row: it differs from the first row in the first place, from the second row in the second place, and, in general, from the nth row in the nth place. So there can be no predicate of English true of exactly those English predicates false of themselves – for if there were such a predicate, its associated row would be our antidiagonal sequence. But there *is* such a predicate – consider the predicate we've just used in the

previous sentence, namely 'English predicate false of itself,' or 'heterological' for short. We are landed in paradox.

Now make some changes to the array. Replace each predicate by its extension, and in each box put '$\in$' or '$\notin$' according to whether the extension at the side has for a member the extension at the top. On certain natural assumptions, we obtain this array:

	extension of 'monosyllabic'	extension of 'French'	extension of 'inanimate'	extension of 'infinite'	...
extension of 'monosyllabic'	$\notin$	$\notin$	$\notin$	$\notin$	...
extension of 'French'	$\notin$	$\notin$	$\notin$	$\notin$	...
extension of 'inanimate'	$\in$	$\in$	$\in$	$\in$	...
extension of 'infinite'	$\notin$	$\in$	$\in$	$\in$	...
⋮	⋮	⋮	⋮	⋮	⋱

(In particular, we assume there are finitely many monosyllabic words, and infinitely many French objects – consider the totality of French sentences. And we assume there are infinitely many inanimate things, and infinitely many infinite things – just consider the infinitely many infinite, and inanimate, extensions generated by the predicates 'natural number greater than 1,' 'natural number greater than 2,' and so on.) Again we can form the diagonal sequence $\notin \notin \in \in \ldots$. The antidiagonal sequence is $\in \in \notin \notin \ldots$, where each $\in$ corresponds to an extension that is not a member of itself (such as the extension of 'monosyllabic' and the extension of 'French'). Again, this antidiagonal sequence cannot occur as a row. So, on pain of a contradiction, there can be no English predicate whose extension is exactly the *non-self-membered extensions of predicates of English*. But the italicized predicate in the previous predicate *is* such a predicate, and we are landed in paradox again.

Our first paradox – the heterological paradox – is a member of the family of *Liar paradoxes*. The Liar takes many forms: for example, versions of the Liar are generated by the sentences "This sentence is false," "This sentence is not true," and "I am lying now." All forms of the Liar turn on the semantic notions of truth or falsity, and so they in turn are members of the family of *semantic paradoxes*. Other members of this extended family turn on the notion of reference or denotation – these are the so-called 'definability paradoxes' due to Richard, König and Berry (see Richard 1905; König 1905; and for the Berry, Russell 1908). Consider, for example, Berry's paradox. There are only a finite number of English expressions with fewer than 19 syllables, and some of these (like 'the square of 3') denote integers. But there are infinitely many integers. Let k be *the least integer not denoted by an English expression in fewer than 19 syllables*. This

italicized phrase denotes k, but it has fewer than 19 syllables – and we've reached a contradiction.

Our second paradox is a version of Russell's paradox, couched in terms of extensions. Russell's paradox also arises for sets – consider the set of exactly the non-self-membered sets, and ask whether or not it is a self-member. This version of Russell's paradox is one of several set-theoretical paradoxes discovered at the turn of the twentieth century. Among these are Burali-Forti's paradox, turning on the set of all ordinal numbers, and Cantor's paradox, concerning the universal set, the set of all sets.

Following Ramsey (1925), it has become standard to divide the paradoxes into two groups: the *semantical* paradoxes (such as the Liar and the definability paradoxes), and the *logical* paradoxes (such as Russell's, Burali-Forti's, and Cantor's). And the attempts to resolve the two kinds of paradoxes have tended to go their separate ways. There is something to this division: the semantical paradoxes arise from ordinary terms of English, like 'true' and 'denotes', while the set-theoretical paradoxes arise in the setting of a mathematical language, and turn on technical notions like *set* and *ordinal number*.

Nevertheless, we should be wary of the division, at least the way Ramsey draws it. As we have seen, the heterological paradox and Russell's paradox have a shared structure: each is generated by a *diagonal argument*. Diagonal arguments establish positive theorems – for example, Gödel's first incompleteness theorem, Tarski's indefinability theorem, and many theorems of recursion theory. But they also generate paradoxes. (For more on the diagonal argument, see Simmons 1993.) The shared diagonal structure of the heterological paradox and Russell's paradox encourages the search for a common resolution. Moreover, Russell's paradox for extensions is tied to predication – and that encourages the thought that it belongs in the category of semantical paradox, along with the heterological paradox.

So there may be a question about how best to classify the paradoxes. But there is no doubt about their tremendous significance: they have forced logicians and philosophers to rework the foundations of semantics and set theory.

2 Semantic Paradoxes: Some Proposals

The hierarchy

Think back to our first paradoxical array, where the top and the side were composed by *all* the 1-place predicates of English, *including* the problematic 'English predicate false of itself.' We can escape paradox if we restrict the array in some suitable way, so that this and other paradox-producing predicates are excluded. Suppose the restricted side and top is the collection D of semantically unproblematic 1-place predicates of English. In particular, the predicate 'English predicate in D false of itself' is excluded from D, on pain of paradox. Here is an application of Russell's Vicious Circle Principle: "Whatever involves *all* of a collection must not be one of the collection" (Russell 1908: 155). We might think of the predicate 'English predicate in D false of itself' as standing above the collection of predicates that it involves.

These ideas may lead us to a *hierarchical* account of truth and falsity. One such account runs as follows. At the first level of the hierarchy are the expressions of English

that do not contain any semantic terms (predicates like 'monosyllabic' and sentences like 'Aardvarks amble'). At the second level we find these first-level expressions together with semantical predicates, like 'true_1,' 'false_1,' 'true_1 of itself,' and 'false_1 of itself,' *which apply only to sentences and predicates of the first level*. We can think of the second-level language as a *metalanguage* for the object language of the first level – the metalanguage contains the semantical terms that apply to the object language. At the third level we find all the second-level expressions (including all the first-level expressions), together with semantical predicates, like 'true_2,' 'false_2,' 'true_2 of itself,' and 'false_2 of itself,' *which apply only to sentences and predicates of the second level*. And so on.

Now semantical paradox does not arise. For example, we can no longer generate the heterological paradox. There is no *absolute* predicate 'English predicate false of itself,' but rather *relativized* predicates of the form 'English predicate false_α of itself' for some ordinal α. This predicate is of level $\alpha + 1$, and applies only to predicates of level α. So it does not apply to itself – and a contradiction is no longer forthcoming.

Or consider the Liar sentence:

(L) (L) is not true.

Here we generate a contradiction by observing that

(1) "(L) is not true" is true if and only if (L) is not true.

This is an instance of Tarski's famous truth-schema:

(T) X is true if and only if p,

where 'p' abbreviates a sentence, and 'X' is a name of that sentence (see Tarski 1944: 15). Given (1), and given that '(L) is not true' just is the sentence (L), we may infer:

(2) (L) is true if and only if (L) is not true,

from which a contradiction immediately follows.

But if we adopt the hierarchical view, this derivation is blocked. Just as the truth (and falsity) predicates are always relativized to a level, so is the truth-schema. The occurrences of 'true' in (L) and in (T) are relativized to some level. So (L) is to be understood as '(L) is not true_β,' for some ordinal β. The T-schema associated with 'true_β' is:

(T_β) X is true_β if and only if p,

for some ordinal β, where 'p' abbreviates a sentence of level β, and 'X' names that sentence. Observe that, according to the hierarchical line, (L) is a sentence of level $\beta + 1$ and not of level β. So it may not be substituted for p in the schema (T_β), and this blocks the Liar reasoning.

How attractive is the hierarchical resolution of semantical paradox? It faces a number of serious difficulties. First, the splitting of 'true' and 'false' into an infinity of

distinct, stratified predicates seems to go against the spirit of a natural language like English. English doesn't seem to be stratified, and the predicate 'true' appears to be univocal. Before his eventual endorsement of the hierarchical approach, Russell himself described it as "harsh and highly artificial" (Russell 1903: 528).

Second, the stratification of 'true' (and 'false') involves massive restrictions on occurrences of 'true.' On a standard hierarchical line, Tim's utterance of "Aardvarks amble" is of level 1; Joanne's utterance of "'Aardvarks amble' is true" is of level 2; and so on, through the levels. Joanne's use of 'true' has in its extension all sentences of level 1 and no others. So all sentences of level 2 and beyond are excluded from the extension of Joanne's use of 'true' (and any use of 'true' in a sentence of level 2). Gödel remarked of Russell's type theory that "each concept is significant only . . . for an infinitely small portion of objects" (Gödel 1944: 149). A similar point can be made here about the hierarchical line: an ordinary use of 'true' will apply to only a fraction of all the truths.

Third, the hierarchical resolution invites a *revenge Liar* – a version of semantical paradox couched in the very terms of the resolution itself. Consider the sentence:

(3) This sentence is not true at any level of the hierarchy.

Suppose (3) is true – that is, on the hierarchical line, true at some level ρ. Then what (3) says is the case, and so (3) is not true at any level, and, in particular, not true at level ρ. Suppose, on the other hand, that (3) is not true at any level. But that is just what (3) says – so (3) is true (at some level). We obtain a contradiction either way: we have traded the old paradoxes for a new one.

Truth-value gaps

Given these difficulties, we might wonder if we can dispense with the hierarchy. In the terms of our first array, let us admit 'English predicate false of itself,' or 'heterological,' to the side and top, and make adjustments elsewhere. Now we should ask: what value can we put in the 'heterological'/'heterological' box in the leading diagonal? On pain of contradiction we cannot put 't' or 'f' in this box. So a natural thought is to appeal to *truth-value gaps*, and say that the predicate 'heterological' is neither true nor false of itself. And we can put 'u,' say, in the 'heterological'/'heterological' box. Now suppose we form the antidiagonal by converting each t to an f, each f to a t, and leaving each u unchanged. Observe that the antidiagonal is identical to the row associated with 'heterological,' and no contradiction arises. Contradiction arises if we assume heterological is true or false of itself – but if it is neither, we escape the paradox. Similarly for Liar sentences; for example, the sentence 'This sentence is false' only generates a contradiction if we assume it is either true or false.

The claim that Liar sentences are gappy seems natural enough – after all, the assumption that they are true or false leads to a contradiction. Moreover, one can motivate gaps independently of the Liar (e.g. by appeal to presupposition theory, or category considerations, or vagueness).

With gaps on board, we can allow the predicate 'English predicate false of itself' to belong to the collection of English predicates – we have no need to invoke Russell's

Vicious Circle Principle. More generally, Kripke (1975) has shown that, if we admit truth-value gaps, it is possible for a language to contain its own truth-predicate. By a fixed-point construction, Kripke obtains a language – call it L_σ – that contains the predicate 'true-in-L_σ,' the extension of which is exactly the true sentences of L_σ. And similarly for the predicate 'false-in-L_σ.' The language L_σ exhibits a striking degree of semantic closure. L_σ has the capacity to express its own concepts of truth and falsity; there is no need to ascend to a metalanguage.

So truth-value gaps are natural enough, and it might appear that they allow us to dispense with the hierarchy. But a moment's reflection shows that any such appearance is deceptive. The 'truth-value gap' approach to semantic paradox faces its own revenge Liar, couched in terms of gaps. Consider the sentence:

(4) This sentence is either false or gappy.

This Liar sentence generates a contradiction whether we assume it is true, false, *or gappy*. In particular, if (4) is gappy, then it is either false or gappy – but that's what (4) says, so it's true. A similar paradox is produced by the sentence:

(5) This sentence is not true,

as long as 'not true' is taken in a suitably wide sense, as coextensive with 'false or gappy' (and not with 'false'). This is a perfectly natural sense of 'not true.' False sentences are not true, of course, but so are gappy sentences – indeed, gappy sentences are, by definition, not true (and not false).

There is a revenge version of the heterological paradox too – just consider the predicate 'English predicate false or neither true nor false of itself' ('superheterological' for short). Or to put it in terms of our array: form the antidiagonal by changing each t to an f, each f to a t, and each u to a t. Now this antidiagonal cannot occur as a row of the array, on pain of contradiction. And yet this antidiagonal sequence just is the row associated with 'superheterological.'

Perhaps we must appeal to the Vicious Circle Principle again, and exclude 'superheterological' from the class of English predicates that it involves. And that would lead us back to the hierarchy. Similarly with Kripke's language L_σ. Although L_σ contains its own truth and falsity predicates, it does not contain 'neither true-in-L_σ nor false-in-L_σ,' or 'not true in L_σ' (in the appropriately wide sense). If we admit these predicates into L_σ, the revenge Liar returns. According to the truth-gap approach, Liar sentences are gappy, and they are not true; however, we cannot say so in L_σ, but only in a semantically richer metalanguage. The language in which we state the gap account, in which we express the notion of a truth-value gap, must be regarded as a metalanguage for L_σ (see Kripke 1975: 79–80, and fn. 34).

Return of the hierarchy?

The point here can be generalized. Suppose I offer a resolution of semantical paradox that makes no appeal to a hierarchy. Let $\mathcal{L}$ be the object language, the language that my semantical theory is a theory of. And let $\mathcal{L}_T$ be the language in which I state my

theory. We can ask: is $\mathcal{L}_T$ a metalanguage for $\mathcal{L}$, on pain of semantical paradox? This is a crucial question, for if the answer is affirmative, then I have not dispensed with the hierarchy, and I have not dealt with semantical paradox in all its forms.

It is a question we can raise not only for Kripke's theory, but for a wide variety of non-hierarchical theories of truth, such as the revision theory (see Gupta 1982; Gupta and Belnap 1993; Herzberger 1982), McGee's treatment of 'true' as a vague predicate (McGee 1990), and Feferman's type-free theory of truth (Feferman 1982). For example, a key notion of the revision theory is that of *stable truth*. The leading idea is that Liar sentences are unstable: if we ascribe truth to the Liar sentence (L), we must revise that ascription, declaring (L) untrue, and then in turn revise that ascription, declaring (L) true, and so on indefinitely. We can ask whether the notion of stable truth must be confined to a metalanguage, on pain of the revenge Liar generated by

(S) (S) is not stably true.

Parallel questions can be raised for McGee's notion of *definite truth*, and for the notion of 'not true' in Feferman's theory (where negation is classical). Observe that all these notions at issue – truth-value gaps, stable truth, definite truth, untruth – are natural enough, and so it is all the more urgent that a purported solution to the Liar come to grips with them. (For an extended discussion of these matters, see Simmons 1993.)

Dialetheism

We have seen that there are serious difficulties with the hierarchical approach. Now suppose we become convinced that nonhierarchical approaches cannot really avoid the hierarchy. In the face of this dilemma, we might seek more radical measures. According to *dialetheism*, Liar sentences are *both true and false* (see, e.g., Priest 1979, 1984). According to Priest, once we admit such truth-value 'gluts', we may dispense with the object language/metalanguage distinction altogether (see Priest 1984: 161). Of course, dialetheism requires that we abandon classical principles of semantics and logic – but only for a certain class of pathological cases, like the Liar family. Perhaps we can cordon off the paradoxical sentences, so that truth-value gluts will be the exception rather than the rule, and classical principles will hold good everywhere else.

But it may not be so clear that the dialetheist can prevent the spread of pathology. One dialetheist account of the truth conditions of 'A is true' and 'A is false' is summed up by these tables:

A	*A is true*
t	t
p	p
f	f

A	*A is false*
t	f
p	p
f	t

where 'p' abbreviates 'paradoxical' (i.e. 'true and false') (see Priest 1979). Let L be a Liar sentence. Then L is both true and false. By the truth tables,

L is true $\leftrightarrow$ L
and L is false $\leftrightarrow$ L.

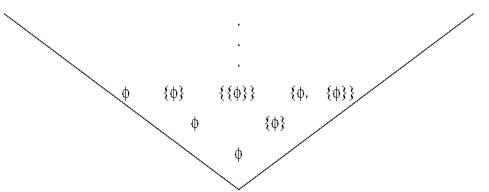
So 'L is true' and 'L is false' are paradoxical. Since, according to the present dialetheist account, the conjunction of two paradoxical sentences is paradoxical, 'L is true and L is false' – that is, 'L is paradoxical' – is paradoxical. So a defining claim of the dialetheist account, that L is paradoxical, is itself paradoxical. The theory itself is not immune from paradoxical assertions. And perhaps this should give us pause.

3 Sets and Extensions

Recall Russell's paradox for sets. Given the set R of exactly the non-self-membered sets, we obtain a contradiction if we assume it is self-membered, and if we assume it isn't. Nowadays, this paradox is no longer considered a real threat: it does not arise in the received set theory, Zermelo-Fraenkel set theory (ZF).

ZF set theory embodies the *combinatorial* or *iterative* conception of set (see Boolos 1971). Think of a set as formed this way: we start with some individuals, and collect them together to form a set. Suppose we start with individuals at the lowest level. At the next level, we form sets of all possible combinations of these individuals. And then we *iterate* this procedure: at the next level, we form all possible sets of sets and individuals from the first two levels. And so on.

In pure set theory we start with no individuals, just the empty set ϕ. Every pure set appears somewhere in this endless cumulative hierarchy:



Observe that no ZF set is a self-member. So if the Russell set R existed it would be the universal set. But there is no universal set, since there is no end to the hierarchy. In this way, Cantor's paradox is avoided. And since it follows that there is no set R, Russell's paradox for sets is also avoided. (Similarly with Burali-Forti's paradox: there is no set of *all* ordinal numbers.)

ZF does provide a consistent set-theoretical basis for mathematics. But there are costs. For one thing, we expect a well-defined predicate to have an extension. In particular, we expect the self-identity predicate to have an extension – but since there is no universal set, ZF does not provide an extension for 'x = x.' Or again, since ZF provides a clearcut concept of *set*, we expect the predicate 'set' to have an extension – and in ZF it doesn't. Note further that in ZF we quantify over sets, and so we need a domain of quantification; but again no set in the hierarchy can serve as this domain.

Such considerations have led some to explore the prospects of a set theory with a universal set (see, e.g., Quine 1937). Thus far those prospects do not seem very bright, at least if we are after a set theory that is plausible and intuitive. A more entrenched response has been to introduce another kind of collection: *classes* or *proper classes*. Proper classes are collections 'too big' to be sets; there is, for example, a proper class of all sets and a proper class of all ordinals. (Proper classes were first explicitly introduced in von Neumann 1925; for a recent discussion, see Maddy 1983.) Of course, a version of Russell's paradox threatens proper classes too. Von Neumann's way out placed a restriction on proper classes: they cannot themselves *be* members. This restriction is severe – we cannot even form the unit class of a proper class. There followed more liberal theories of classes (see, e.g., Levy et al. 1973), but in none of these theories can a proper class be a self-member, and so Russell's paradox does not arise.

However, the introduction of classes seems merely to push the problem back. Still there is no extension for 'x = x,' or for the predicates 'class' and 'proper class.' And no class can serve as the domain of quantification over classes.

A more promising strategy, it would seem, is to develop a theory of extensions from scratch. In my view, the notions of *extension* and *set* (or *class*) are independent and mutually irreducible. We cannot reduce sets or classes to extensions, for extensions are essentially tied to predication, and sets and classes are not. (Given some natural assumptions, there are strictly more sets in the ZF hierarchy – and more classes – than there are predicates in, say, English.) And we cannot reduce extensions to sets or classes. No set can serve as the extension of 'set,' and no class can serve as the extension of 'class'; and there are, as we have seen, self-membered extensions, but no self-membered classes or ZF sets.

If we do develop a theory of extensions directly, we must of course find a way out of Russell's paradox for extensions. We saw in Section 1 that this paradox is best viewed as a semantical paradox, and that it shares structural similarities with the heterological paradox. All the better, then, if we can find a unified solution to this version of Russell's paradox, the paradoxes of definability, and the Liar paradoxes.

4 Three Paradoxes

In search of such a unified account, consider three paradoxes. First, suppose that I am confused about my whereabouts (I think I am in room 102), and I write on the board in room 101 the following denoting expressions:

(A) the ratio of the circumference of a circle to its diameter.
(B) the successor of 5.
(C) the sum of the numbers denoted by expressions on the board in room 101.

It is clear what the denotation of (A) and (B) are. But what is the denotation of (C)? Suppose (C) denotes k. Then the sum of the numbers denoted by expressions on the board is $\pi + 6 + k$. So (C) denotes $\pi + 6 + k$. So $k = \pi + 6 + k$. We are landed in a contradiction.

So we should conclude:

(6) (C) is pathological, and does not denote a number.

Now we can reason that (A) and (B) are the only expressions on the board that denote numbers. So we may conclude that *the sum of the numbers denoted by expressions on the board in room 101* is $\pi + 6$. Observe that in the previous sentence there occurs a token of the same type as (C), call it (C*). Unlike (C), (C*) is not pathological, and it does have a denotation. We may conclude:

(7) (C*) denotes $\pi + 6$.

How can two expressions – composed of exactly the same words with the same linguistic meaning – differ so dramatically in their semantic status?

Suppose next that I write on the board in room 101 these two predicates:

(E) moon of the Earth
(F) unit extension of a predicate on the board in room 101.

The extension of predicate (E) is a unit extension, and so it is a member of the extension of (F). What about the extension of (F)? Suppose first that it is a self-member. Then the extension of (F) has two members, so it is not a unit extension – and so it is not a self-member. Suppose second that it is not a self-member. Then the extension of (F) has just one member, so it is a unit extension – and so it is a self-member. Either way we obtain a contradiction.

So we should conclude that (F) is a pathological predicate that fails to have an extension. But if (F) does not have an extension, then in particular it does not have a unit extension. So the only *unit extension of a predicate on the board in room 101* is the extension of (E). We've just produced a token of the same type as (F), call it (F*). But unlike (F), (F*) has a well-determined extension (whose only member is the extension of (E)). Again we can ask: how is that these two expressions – composed of the very same words – differ in their semantic status?

Finally, consider the case of truth. If I write on the board in room 101 the following sentence:

(L) The sentence written on the board in room 101 is not true,

then I have produced a Liar sentence. We are landed in a contradiction whether we assume (L) is true, or not true. So we can conclude that (L) is semantically pathological. As we have seen, semantic pathologicality may be cashed out in a variety of ways – for example, perhaps (L) is gappy or unstable. But if (L) is pathological, then it is not true. That is, we may conclude:

(L*) The sentence written on the board in room 101 is not true.

And while (L) is pathological, (L*) is true. Again, the two sentences differ in semantic status, yet they are tokens of the same type.

5 A Contextual Approach

How should we resolve these paradoxes? In each case, we have the same phenomenon: a change in semantic value (in denotation, extension, or truth-value) without a change in linguistic meaning. Such a change suggests some pragmatic difference.

Consider the case of denotation, though what we say about this case carries over to the others. There are a number of differences between the context of (C) and the context of (C*). Beyond the familiar contextual parameters of speaker, time, and place, there are differences in *discourse position, intentions*, and *relevant information* as well. (C*) is produced at a later stage of the discourse, *after* it has been established that (C) is pathological. At this later stage, we reason in the light of (C)'s pathology – we may say that the context in which we produce (C*) is *reflective* with respect to (C). Intentions shift too. At the later stage, our intention is to treat (C) as pathological and see where this leads us. But I have no such intention at the first stage – my intention in producing (C) is to refer to expressions on the board next door. There is a corresponding shift in information: the information that (C) is pathological is available throughout the later stage of the reasoning, but it is not available to me when I first produce (C). These contextual differences all contribute to a crucial contrast between the contexts of (C) and (C*): the former is *unreflective with respect to (C)*, and the latter is *reflective with respect to (C)*.

If we accept the appropriateness of a pragmatic explanation, then we should expect to find a term occurring in (C) and (C*), and in (1) and (2), that is context-sensitive. When we inspect the terms occurring in these expressions, there seems to be only one candidate: the predicate 'denotes.' Accordingly, let us represent (C) by

(C) the sum of the numbers denoted$_C$ by expressions on the board in room 101,

where the subscript indicates that the use of 'denotes' in (C) is tied to (C)'s unreflective context of utterance.

To determine the denotation of (C), then, we must determine the denotation$_C$ of expressions on the board – that is, the denotations$_C$ of (A), (B), and (C). The conditions under which an expression denotes$_C$ is given by a denotation schema (analogous to the truth-schema):

s denotes$_C$ n iff p = n,

where instances of the schema are obtained by substituting for 'p' any referring expression, for 's' any name of this expression, and for 'n' any name of an individual. When we apply this C-schema to (C), we obtain a contradiction, and this leads to the conclusion (8), represented by:

(8) (C) does not denote$_C$ a number.

We go on to reason that (A) and (B) are the only expressions on the board that denote$_C$ numbers, since (C) does not. So we infer that the sum of the numbers denoted$_C$

by expressions on the board in room 101 is $\pi + 6$. In producing (C*) here, we have in effect repeated (C). But we have repeated (C) in a new context, a context that is reflective with respect to (C). We no longer provide denotation conditions via the C-schema. In this new reflective context – call it R – denotations are determined *in the light of (C)'s pathology*. That is, denotations are determined by the R-schema:

s denotes$_R$ n iff p = n.

And (C*) *does* have a denotation$_R$. Consider the biconditional:

(C*) denotes$_R$ k iff the sum of the numbers denoted$_C$ by expressions on the board in room 101 at noon 7/1/99 is k.

The right-hand side is true for $k = \pi + 6$, since we have established that (C) does not denote$_C$. And so we infer

(9) (C*) denotes$_R$ $\pi + 6$.

(C) and (C*) are semantically indistinguishable – the difference between them is a purely pragmatic one. It is a matter of the denotation schemas by which (C) and (C*) are given denotation conditions. At the first stage of the reasoning, (C) is assessed via the unreflective C-schema; at the second stage, (C*) is assessed via the reflective R-schema. Notice that if we assess (C) via the R-schema, we find that (C), like (C*), denotes$_R$ $\pi + 6$; and if we assess (C*) via the C-schema, we find that (C*), like (C), does not denote$_C$ a number. So the tokens of 'denotes' in (8) and (9) have different extensions: (C) and (C*) are not in the extension of 'denotes$_C$,' but both are in the extension of 'denotes$_R$.' So 'denotes' is a context-sensitive term that may shift its extension with a change in context – as it does in the move from (8) to (9).

We can give exactly parallel analyses of the cases of extension and truth. We take 'extension' and 'true' to be context-sensitive terms, and explain the difference between (F) and (F*), and between (L) and (L*), in terms of a change in evaluating schemas.

6 A Singularity Proposal

The question naturally arises: what is the relation between the unreflective and reflective stages? A possible response here is a Tarskian one: when we move from the first stage of the reasoning to the second, we push up a level of language. (For contextual accounts of truth that appeal to a hierarchy, see Parsons 1974; Burge 1979; Barwise and Etchemendy 1987; Gaifman 1988, 1992.) So, for example, the terms 'denotes$_C$' and 'denotes$_R$' belong to distinct levels, and the extension of 'denotes$_R$' properly contains the extension of 'denotes$_C$.'

We have already seen the difficulties that hierarchical accounts face. But a unified hierarchical account of reference, extension, and truth faces a special difficulty. It is the case of extensions that presents the problem. Extensions can be self-membered; for example, as we saw in Section 1, the extension of the predicate 'infinite extension'

belongs to itself. According to the hierarchical approach, this predicate is of the form 'infinite extension$_\sigma$.' And the predicate itself is a predicate of $L_{\sigma+1}$ and not of L_σ. So the extension of this predicate is not a self-member – it contains extensions of predicates of L_σ only. The hierarchical account cannot accommodate self-membered extensions. A distinctive feature of extensions is regimented away.

But perhaps we can retain the contextual idea and jettison the hierarchy. This is the idea behind the *singularity theory* (see Simmons 1993, 1994). Occurrences of 'denotes,' 'extension,' and 'true' are to be *minimally* restricted, in accordance with the pragmatic principle of Minimality. Suppose, for example, you say "'The square of 1' denotes 1." Here, your use of 'denotes' is quite unproblematic. Should (C) be excluded from its extension? According to Minimality, the answer is no – because there is no need to exclude it. We have seen that (C) denotes$_R$ $\pi + 6$ because the sum of the numbers denoted$_C$ by expressions on the board *is* $\pi + 6$. And for the same reason, (C) denotes$_N$ $\pi + 6$, where N is the context of your utterance, a context neutral with respect to (C).

If we adopt Minimality we respect a basic intuition about predicates. In general, if an individual has the property picked out by the predicate φ, then we expect that individual to be in the extension of φ. The more restrictions we place on occurrences of 'denotes' (or 'extension' or 'true'), the more we are at odds with this intuition. Minimality keeps surprise to a minimum.

So the present proposal identifies *singularities* of the concepts *denotes*, *extension*, and *truth*. For example, (C) is a singularity of 'denotes$_C$,' because it cannot be given denotation$_C$ conditions. Notice that (C) is a singularity only in a context-relative way – it is not a singularity of 'denotes$_R$' or 'denotes$_N$.'

No occurrence of 'denotes' or 'extension' or 'true' is without singularities. For example, consider again (7):

(7) (C*) denotes $\pi + 6$.

Consider the following perverse addition to (7):

(10) And so the number denoted by (C*), plus the number denoted by 'the square of 1,' plus the sum of the numbers denoted by phrases in this sentence, is irrational.

Given the context, the occurrences of 'denotes' in our continuation will be represented by 'denotes$_R$.' Consider the final definite description token in our utterance (beginning 'the sum of') – call this token (D). (D) is a singularity of 'denotes$_R$' – the R-schema cannot provide it with denotation conditions.

The example of (D) brings out the anti-hierarchical nature of the singularity proposal. Observe that we can reflect on (D), just as we earlier reflected on (C). In a suitably reflective context, we can conclude that (D) denotes $(\pi + 6) + 1$ – since the only denoting phrases in (10) that denote$_R$ numbers are the first two phrases, and these phrases denote $\pi + 6$ and 1 respectively. And by Minimality, (D) will have this denotation when assessed by any schema other than the R-schema. In particular, the token *does* denote$_C$ – it is *not* a singularity of 'denotes$_C$.' The C-schema does determine a denotation for it. On a Tarskian account, the extension of 'denotes$_C$' will be a proper subset

of the extension of 'denotes$_R$.' According to the singularity proposal, neither extension includes the other.

Gödel once made the following tantalizing remark about the paradoxes:

> It might even turn out that it is possible to assume every concept to be significant everywhere except for certain 'singular points' or 'limiting points', so that the paradoxes would appear as something analogous to dividing by zero. Such a system would be most satisfying in the following respect: our logical intuitions would then remain correct up to certain minor corrections, i.e. they could then be considered to give an essentially correct, only somewhat 'blurred', picture of the real state of affairs. (Gödel 1944: 229)

I take the singularity proposal to be in the spirit of Gödel's suggestion. According to the present account, our intuitions about 'denotes' – and 'extension' and 'true' – are almost correct. It is only in pathological or paradoxical contexts that we may mistakenly suppose that certain phrases denote when they do not – and in such cases our applications of 'denotes' require only *minimal* corrections. We retain a single denotation predicate which undergoes minimal changes in its extension according to context. There is no wholesale revision of the notion of denotation; no division of 'denotes' into infinitely many distinct predicates; no splitting of everyday English into an infinite hierarchy of languages.

7 Universality

It is of course beyond the scope of this chapter to provide a formal theory of singularities (see Simmons (1993) for a singularity theory of truth). But suppose we had such a formal theory $\mathcal{L}_T$ for an object language $\mathcal{L}$ containing the context-sensitive predicate 'denotes,' or 'extension,' or 'true.' Won't we now face the familiar objection, that $\mathcal{L}_T$ is a metalanguage for $\mathcal{L}$, and the hierarchy is inevitable? Moreover, since we may regard $\mathcal{L}_T$ as a classical formal language, it will be subject to Tarski's theorem. So the semantical predicates for $\mathcal{L}_T$ will be contained in a further metalanguage, which in turn cannot contain *its* semantic predicates. From $\mathcal{L}_T$, then, a whole hierarchy of languages is generated.

But the singularity account is not without resources here. The context-sensitive predicate 'denotes' applies to any denoting phrase of $\mathcal{L}_T$ as long as that phrase is not identified as a singularity – and similarly for any denoting phrase at any level of the ensuing hierarchy. No language of the hierarchy is a metalanguage for $\mathcal{L}$ – 'denotes' applies to phrases of all levels. (Parallel remarks can be made about 'extension' and 'true.') The scope of 'denotes' is as close to universal as it can be.

According to Tarski, natural languages are "all-comprehensive" and "universal":

> The common language is universal and is intended to be so. It is supposed to provide adequate facilities for expressing everything that can be expressed at all, in any language whatsoever; it is continually expanding to satisfy this requirement. (Tarski 1969: 89)

It is the apparent universal character of natural language that both generates semantical paradoxes and makes them so difficult to solve. Any semantic account of 'denotes,'

'extension,' or 'true' will just be more English, and so the stage is set for a revenge Liar. Whether the singularity account, or some other, can do sufficient justice to this feature of natural language cannot be settled here. But the challenge remains. At root, semantical paradox and the problem of universality are one and the same.

References

Barwise, Jon and Etchemendy, John (1987) *The Liar*. Oxford: Oxford University Press.

Boolos, George (1971) The iterative conception of set. *Journal of Philosophy*, 68, 215–32; reprinted in Putnam and Benacerraf (1983) 486–502.

Burge, Tyler (1979) Semantical paradox. *Journal of Philosophy*, 76, 169–98; reprinted with a postscript in R. L. Martin (ed.) (1984) 83–117.

Feferman, Solomon (1982) Towards useful type-free theories, I. *Journal of Symbolic Logic*, 49, 75–111; reprinted in R. L. Martin (ed.) (1984) 237–87.

Gaifman, Haim (1988) Operational pointer semantics: solution to self-referential puzzles I. *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, (eds.) M. Vardi and Morgan Kaufman. Los Altos, California, 43–60.

Gaifman, Haim (1992) Pointers to truth. *Journal of Philosophy*, 223–61.

Gödel, Kurt (1944) Russell's mathematical logic. In P. A. Schilpp (1944) 123–53.

Gupta, Anil (1982) Truth and paradox. *Journal of Philosophical Logic*, 11, 1–60; reprinted in R. L. Martin (ed.) (1984) 175–235.

Gupta, Anil and Belnap, Nuel (1993) *The Revision Theory of Truth*. Cambridge, MA: MIT Press.

Herzberger, Hans (1982) Notes on naive semantics. *Journal of Philosophical Logic*, 11, 61–102; reprinted in R. L. Martin (ed.) (1984) 133–74.

König, Julius (1905) On the foundations of set theory and the continuum problem. In van Heijenoort (1967) 145–9.

Kripke, Saul (1975) Outline of a theory of truth. *Journal of Philosophy*, 72, 690–716; reprinted in R. L. Martin (ed.) (1984) 53–81.

Levy, A., Fraenkel, A. A. and Bar-Hillel, Y. (1973) The role of classes in set theory. In Muller (1976) 173–215; first published as Chapter II, Sec. 7 of *Foundations of Set Theory*. Amsterdam: North-Holland.

Maddy, Penelope (1983) Proper classes. *Journal of Symbolic Logic*, 48, 113–39.

Martin, Robert L. (ed.) (1984) *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford University Press.

McGee, Vann (1990) *Truth, Vagueness, and Paradox*. New York: Hackett.

Parsons, Charles (1974) The liar paradox. *Journal of Philosophical Logic*, 3, 381–412; reprinted with a postscript in R. L. Martin (1984) 9–45.

Priest, Graham (1979) The logic of paradox. *Journal of Philosophical Logic*, 8, 219–41.

Priest, Graham (1984) Logic of paradox revisited. *Journal of Philosophical Logic*, 13, 153–79.

Quine, W. V. (1937) New foundations for mathematical logic. In Quine (1953) 80–101.

Quine, W. V. (1953) *From a Logical Point of View*. New York: Harper Torchbooks.

Ramsey, Frank (1925) *The Foundations of Mathematics*. London: Routledge & Kegan Paul.

Richard, Jules (1905) Les principes des mathématiques et le problème des ensembles. In *Revue générale des sciences pures et appliquées*, 16, 541. Also in *Acta Mathematica*, 30 (1906), 295–6. (English translation in van Heijenoort (1967) 143–4.)

Russell, Bertrand (1903) *The Principles of Mathematics*. Cambridge: Cambridge University Press.

Russell, Bertrand (1908) Mathematical logic as based on the theory of types. *American Journal of Mathematics*, 30, 222–62; reprinted in van Heijenoort (1967) 150–82.

Schilpp, P. A. (ed.) (1944) *The Philosophy of Bertrand Russell*. La Salle, IL: Open Court.

Simmons, Keith (1993) *Universality and the Liar: An Essay on Truth and the Diagonal Argument*. Cambridge: Cambridge University Press.

Simmons, Keith (1994) Paradoxes of denotation. *Philosophical Studies*, 76, 71–104.

Tarski, Alfred (1944) The semantic conception of truth. *Philosophy and Phenomenological Research*, iv, 341–76; reprinted in Leonard Linsky (ed.), *Semantics and the Philosophy of Language*. Chicago: University of Illinois, 1952, 13–47.

Tarski, Alfred (1969) Truth and proof. *Scientific American*, 220, 63–77.

Van Heijenoort, Jean (ed.) (1967) *From Frege to Gödel: A Source Book in Mathematical Logic*. Cambridge, MA: Harvard University Press.

Von Neumann, John (1925) An axiomatization of set theory. In Van Heijenoort (1967) 394–413.

9

Philosophical Implications of Logical Paradoxes

ROY A. SORENSEN

Dr. Seuss' *On Beyond Zebra* opens with a young boy proudly writing on a blackboard. Conrad Cornelius o'Donald o'Dell, has demonstrated his exhaustive knowledge of the alphabet: A is for Ape, B is for Bear, . . . and Z is for Zebra. An older boy compliments Conrad. He breezily concedes to young Conrad that *most* people stop with Z. But *his* alphabet continues beyond Z. The extra letters let him spell new things. The older boy thus introduces Conrad to an otherwise inaccessible realm of exotic creatures. For instance, the Q-ish letter quan is for the vertically symmetric Quandary who lives on a shelf

> In a hole in the ocean alone by himself
> And he worries, each day, from the dawn's early light
> And he worries, just worries, far into the night.
> He just stands there and worries. He simply can't stop . . .
> Is his top-side bottom? Or bottom-side top?

A metaphysician who tags along on the tour given to Conrad will be reminded of other never-never lands.

1 Paradoxes Stimulate Theory Development

Why aren't statements that assert the existence of an entity trivially true? If 'Santa Claus exists' is false, then the sentence must be about Santa Claus. 'About' is a two-place relation, so the sentence is about Santa only if there is a Santa. But then 'Santa Claus exists' is true! Alexius Meinong challenged the transition from 'There is a Santa' to 'Santa exists.' Meinong believed there is a complex domain of nonexistent objects that have been neglected by metaphysicians much as astronomers long neglected dark matter. Meinong's metaphysics illustrates how a logical paradox can stimulate the development of a philosophical theory.

The excesses of such a theory provoke debunkers. After initial sympathy with Meinong, Bertrand Russell (1957) traced belief in nonexistent objects to an illusion about 'the.' 'The present king of France is bald' appears to refer to the present king of

France. Russell dismantled this appearance. His theory of descriptions underwrites the synonymy of 'The present king of France is bald' and 'There is exactly one king of France and whoever is king of France is bald.' The first conjunct of this sentence can be denied without referring to any particular entity.

The hypothesis that names are disguised definite descriptions subsumes 'Santa Claus does not exist' under Russell's theory. Subsequent philosophers supported Russell's linguistic thesis with increasingly sophisticated proposals as to what the disguised definite description is. Typically, deflationary accounts generate auxiliary linguistic theories and distinctions. After all, logic can be applied only with the help of assumptions about how key locutions operate. Even those who do not accept the dissolutions (whether through substantive doubts or indifference toward the motivating problem) have been impressed by some of these auxiliary theories. Linguists were especially quick to incorporate the great handmaiden of logic, H. P. Grice's theory of conversational implicature.

When the theory development takes place within logic itself, the result is a powerful constraint on all future theorizing. Any scientific result is a constraint in the sense that it constitutes grounds against any future theory that conflicts with the result. But most scientific results are domain specific and of a limited degree of necessity. For instance, it is physically impossible to make a left shoe into a right shoe by turning it over through a fourth dimension. But crystallographers, topologists and philosophers can coherently study this kind of mirror reversal. They are interested in a wider domain of possibility. Logic is at the limit of this scale of possibility. Consequently, logical impossibilities are maximally coercive. Even the skeptic is careful to keep his scenarios within the bounds of logical law.

2 An Analogy with Perceptual Illusions

Russell (1957: 47) tested his theory of descriptions by how it handled related paradoxes such as the surprising informativeness of identity statements. He advised logicians to keep a stock of logical puzzles on the grounds that these play the same role as experiments do for scientists.

Some logicians are unimpressed with Russell's analogy. They adopt the same attitude toward logical paradoxes that the perceptual psychologist J. J. Gibson took toward perceptual illusions. According to Gibson's ecological view, perception must be understood within the perceiver's natural environment, as an adaptation toward practical ends. Gibson dismissed perceptual illusions as largely irrelevant, confined to picture books and computer generated toy worlds.

After the entrenchment of Aristotle's logic, little distinction was made between sophistry and logical paradoxes. The liar paradox and the sorites paradox (which are in high esteem today) were regarded by almost all subsequent philosophers as isolated curiosities. Medieval logicians are the important exceptions. Their highly structured system of scholarly debate encouraged attention to logic and language. Contrary to stereotype, they approached many philosophical issues with an escape artist's mix of imagination and rigor. Pseudo-Scotus' paradox of validity and Jean Buridan's variants of the liar have found their way back into academic publications – and not just for their antiquarian value.

Currently, the dominant view in psychology is that perceptual illusions are more than an entertaining sideline. As Hermann von Helmholtz said a century ago, illusions are anomalies that provide clues as to how normal perception originates. Each sense is a package of rough and ready modules that evolved to achieve collective reliability, not individual reliability. Illusions arise when experimenters isolate systems under laboratory conditions or venture into environments and circumstances alien to our hunter-gatherer heritage. If the detection of validity is also a 'bag of tricks,' then fallacies should provide clues about normal reasoning. And indeed, psychologists have had much to say about 'cognitive illusions' (ignoring base rates, confirmation bias in the four card selection task, the Monty Hall problem). Yet they have had little to say about logical paradoxes.

The silence of the psychologists is an anomaly for those who believe that there is only a quantitative difference between brainteasers of recreational logic and deep logical paradoxes. Logical pedagogy reflects this gradualist sentiment. Raymond Smullyan's *What is the name of this book?* starts at the shallow end of the cognitive pool and then moves cheerfully and continuously into deeper waters. There is no sharp line between non-philosophical puzzles that have algorithmic solutions and philosophical problems in which we only have a hazy idea of what would even count as progress.

The gradualist might venture a random walk theory of profundity. If we are restricted to trial and error, it is statistically easier to solve an n step problem than an n + 1 step problem. The probability of solution is a geometrical rather than an arithmetic function of the number of steps needed for a solution. Consequently, a problem that is a few more steps from resolution will have a much lower probability of solution. These recalcitrant problems are apt to be perceived as qualitatively different. In addition to being more difficult to solve, the more complex problems will be less detectable. Lewis Carroll produced thousands of unmemorable logic exercises but discovered only a handful of logical paradoxes.

Gradualism breeds optimism about the ultimate solubility of philosophical problems. If profound problems differ only in degree from solvable brainteasers, then philosophical progress is probable. The appearance of stagnation would be best explained as a sampling illusion: as soon as philosophical problems get solved, they get exported to some other field.

However, this optimism about philosophical progress comes at the price of deflationism about philosophy. If philosophical problems are just highly complicated brainteasers, then why expect their solution to be more illuminating than the solution of a highly complex brainteaser? Lewis Carroll's immense corpus of puzzles contains counterexamples to the thesis that sheer complexity always generates an appearance of profundity. To divert his Victorian mind from unwelcome thoughts, the sleepless Carroll carried certain puzzle genres to awesome lengths. One genre involves inconsistent story telling in which a contradiction can be derived from n statements in the story but not any n – 1 of those statements. Carroll has stories in which n = 25 and even one in which n = 50. These look like clerical feats rather than deep thinking.

To constitute a paradox, a problem must be an apparent counterexample to an attractive principle. Arthur Prior's runaway inference ticket, 'tonk,' is paradoxical because it is a counterexample to the conventionalist thesis that the meaning of the logical connectives is dictated by the truth-tables. Nelson Goodman's new riddle of

induction is a paradox because it refutes the assumption that induction is topic neutral. Hilary Putnam contends that Heisenberg's uncertainty principle refutes the principle of distribution $(A \& (B \vee C) \supset (A \vee B) \& (A \vee C))$.

Anti-gradualists are apt to interpret the silence of the psychologists as a sign that there is a qualitative difference between logical paradoxes and the shallower fare of howlers, forehead slappers, and joke demonstrations that $1 = 0$. Perhaps shallow logical errors are just performance errors. The subjects tested by psychologists only have a short time to answer the test questions and are vulnerable to distraction, memory overload, etc. Logical paradoxes, in contrast, have withstood the scrutiny of motivated, leisurely study by experts. The error occurs at the level of theory rather than implementation. Or perhaps the deep logical paradoxes reflect some master cognitive flaw – something akin to the transcendental illusion (of applying phenomenal categories to noumena) that Immanuel Kant postulated to explain the antinomies of space and time.

Possibly, the logical paradoxes will help us discover a single grand truth that explains the mix of anomalies. The random walk theory makes the opposite prediction that the paradoxes will only have the coincidental patterns that one normally finds in a well-shuffled deck of cards.

3 Do Logical Paradoxes Exist?

Kant believed there were no logical paradoxes. This is evident from his preface to the second edition of the *Critique of Pure Reason*. In Kant's opinion, Aristotle had successfully grasped the basic truths of logic in his theory of the syllogism just as Euclid had grasped the basic truths of geometry with his axiomization. There are geometrical sophisms and questions of application. But there are no anomalies within the theory itself. After Aristotle, logic "has not had to retrace a single step, unless we choose to consider as improvements the removal of some unnecessary subtleties or the clearer exposition of its doctrine, both of which refer to the elegance rather than to the solidity of the science. It is remarkable also, that to the present day it has not been able to advance a step and is thus to all appearance complete and perfect."

The history of logic refutes Kant. I mean to include the history that preceded Kant (especially the medieval era). Plus the era to which he belonged. But most of all, Kant is refuted by the history that followed him. Logic made great strides after the nineteenth century, often under the stimulus of paradox.

Despite the historical record, there remain strangely rich grounds for doubting the existence of logical paradoxes. For instance, most theories of belief imply that no one can believe a contradiction (Sorensen 1996). Since the paradigm cases of logical paradoxes involve belief in contradictions, the very existence of logical paradoxes is itself paradoxical.

Another difficulty is taxonomic. Paradoxes are classified in terms of the propositions they contain. Olber's paradox of why the night sky is dark is an astronomical paradox because its constituent propositions are astronomical. Presumably, a logical paradox contains logical propositions. However, there are conceptions of 'proposition' and 'logic,' which preclude the existence of logical propositions. In the *Tractatus*, Wittgenstein reserves 'proposition' for statements reporting contingent states of affairs.

Gilbert Ryle regarded logical laws as rules of inference. Since rules *prescribe* how we ought to behave rather than describe how things are, logical laws are neither true nor false. Logic can only be relevant to paradoxes as a means of relating propositions to each other. Adjusting a rule of inference might resolve the paradox in the sense of dissolving the appearance of inconsistency. But there are no logical propositions that can serve as members of a paradox. Or so one might infer.

The sharp distinction between inference rules and premises appears to undermine the notion of a logical paradox. Ironically, this distinction was itself drawn in response to a logical paradox. In 1895, Lewis Carroll published a dialogue in *Mind* between Achilles and Tortoise. The Tortoise will grant Achilles any premise he wishes. But the Tortoise insists that Achilles link the premises to the conclusion via a further premise, which states that if the premises are true, then the conclusion is true. Since this extra conditional is itself a premise, adding it to the premise set forces the addition of a new linking premise. The common solution to this paradox is to deny that any extra premises are needed to link the premises and the conclusion. They are instead linked by an inference rule.

Lewis Carroll's puzzle about Achilles and Tortoise does dramatize the need to distinguish between premises and inference rules. However, it does not refute the basic interchangeability of premises and inference rules. The axiom that p can be considered as an inference rule lets us introduce p without any premises. Carroll's puzzle does show that a system that contains just axioms cannot have any deductions. However, natural deduction systems are practical examples of how a system may have deductions without any axioms.

Once we agree that there are logical paradoxes, there remains the question of which propositions are logical propositions. The sentence-level answer appeals to logical words. Certain statements are guaranteed to have a truth-value by virtue of vocabulary found in logical theories. This vocabulary uncontroversially includes the following: and, or, not, all, some, is. These words are topic neutral, appearing across all domains of discourse – not just in physics or tennis or algebra.

This does little to relieve the surprising amount of disagreement over what qualifies as a logical word. There is consensus that all the vocabulary of first order predicate logic with identity qualifies as logical. There is a debate over whether the introduction of predicate variables (to obtain second order logic) is just disguised set theory. Other marginal examples of logical words tend to be diplomatically treated 'as if' they were logical words. The modal logician declares he will treat 'necessary' as a logical word and thereby obtains a supplemental logic. The same is done for temporal logic (earlier than), mereology (part of), deontic logic (permissible), epistemic logic (know), etc. The longer the list of logical words, the greater the number and variety of logical paradoxes.

It is more natural to characterize logical paradoxes at the theory-level. Given that logic is the *theory* of what follows from what, there will be propositions about propositions. These meta-propositions about the consequence relation will sometimes be individually plausible and yet jointly inconsistent. This conception of a logical paradox accommodates the tendency to include meta-logical paradoxes as logical paradoxes. The Lowenheim-Skolem paradox makes essential use of words that are about logic but which are not logical words.

The sentence-level conception of paradox accommodates the inclusion of puzzles that inadvertently involve a logical truth or logical falsehood. The doctrine of the trinity is sometimes described as a logical paradox because it is a paradox (to many Christians) that involves violations of the law of identity.

Theory-level paradoxes arise out of logical doctrines and intuitions. For instance, we believe logic must handle every possible state of affairs and hence it cannot imply the existence of anything. We also believe in logical laws such as the principle that everything is identical to itself. But since the quantifiers in standard logic have existential import $(x)(x = x)$ entails $(\exists y)(y = y)$. Thus the empty universe is excluded as *logically* impossible. So at least one of these propositions must be false. Which? Metaphysical intuitions have little standing against science or mathematics. Why should logic be any more deferential to metaphysics? Good bookkeeping requires the rejection of empty universe.

4 Imagination Overflows Logical Possibility

Logic is unnervingly forthcoming with respect to the philosophical question "Why is there something rather than nothing?" But logic has a way of building up your nerve. Intuitions and imposing theories about what is possible have both been challenged on logical grounds.

Russell's theory of definite descriptions shows how, in Ludwig Wittgenstein's words, "A cloud of philosophy is condensed into a drop of grammar." Philosophy has no monopoly on fog. Wittgenstein would see an analogy between the realm opened by Dr. Seuss's trans-Z letters and Georg Cantor's 'paradise' of transfinite numbers. The theologian-mathematician Cantor was trying to solve mathematical paradoxes involving counting. On the one hand, there seem to be more natural numbers than even numbers because the even numbers are properly included amongst the naturals. Yet it is possible to put the natural numbers into a one to one correspondence with the even numbers. This mapping indicates the number of even numbers *equals* the number of natural numbers. Instead of dismissing this correspondence as revealing that there is something wonky in the notion of infinity, Richard Dedekind boldly defined 'infinite set' in terms of this paradoxical property of having a proper subset that is as large as itself. Cantor took set theory much further. His innovative diagonal argument showed that the set of real numbers is larger than the set of natural numbers. The argument generalizes to reveal a hierarchy of infinities that obey strange but elegant laws of addition, subtraction, and so forth. Most of those who become familiar with this transfinite arithmetic emerge with Russell's conviction that Zeno's paradoxes now have a mathematical solution. Set theory was speedily erected into a grand unifying theory of mathematics.

What is the difference between 'Cantor's paradise' and the realm the older boy offers Conrad Cornelius o'Donald o'Dell? There cannot be any letters beyond Z because 'The letters from A to Z exhaust the alphabet' is an analytic truth. Of course, one could invent another alphabet in which Z is not the last letter. But then the older boy would not be *correcting* Conrad's initial impression that Z is the last letter. For young Conrad was talking about the standard English alphabet. The suggestion that young Conrad is

representationally deprived rests on an equivocation between the established alphabet and a pseudo-alternative. Similarly, Wittgenstein balked at the suggestion that Cantor had discovered numbers that had been previously overlooked by conceptually immature predecessors.

Philosophers have shrunk from Wittgenstein's intimation that Cantor's paradise is as mythical as Meinong's slum of nonexistents. They dwell on the ways in which paradoxes expand our horizons. Russell writes warmly of how Cantor's theory helps us distinguish between contradictions and possibilities that merely contradict our prejudice for thinking in finite terms. W. V. Quine advises us to abandon the quest to translate names into definite descriptions and to instead think directly in terms of Russell's logical notation for definite descriptions. Like any fluent speaker, we no longer need to translate back into our native tongue. Russell has enriched our minds with a new tool of thought. This kind of conceptual advance occurs in all fields. The scale of the effect varies with the centrality of the subject-matter. Since logic is at the center of the web of belief, the implications are wide indeed.

Although one may resist the Wittgensteinian assimilation of Cantor's transfinite numbers to Dr. Seuss's trans-Z letters, all must concede the general point that a cognitive advance often takes the form of an eliminated possibility. Children search their drawers for lost dogs. They scuttle toward the opposite end of the bathtub to avoid being sucked down the drain. This kind of open-mindedness needlessly alarms them and slows their searches. As children mature, their conception of what is possible more closely aligns with what is genuinely possible.

Some paradoxes rest on bogus possibilities. Consider the barber who shaves all and only those who do not shave themselves. Does the barber shave himself? If he shaves himself, then he is amongst those he does not shave. But if he does not shave himself, then he is amongst those he shaves. Contradiction. The universally accepted solution is that we should not assume that it is possible for there is to be a barber who shaves all and only those he does not shave. We need to rein in our imagination.

And not just about barbers. Our imaginations systematically run afoul of J. F. Thomson's theorem:

> Let S be any set and R any relation defined at least on S. Then no element of S has R to all and only those S-elements which do not R to themselves. (Thomson 1962: 104)

If we let S be the collection of men, then this set contains no man who bears the relation of shaving all and only those men who not shave themselves. That dissolves the barber paradox.

Thomson goes on to show how his 'small theorem' is at the root of Kurt Grelling's paradox about 'heterological.' The lesson is that there is no predicate that applies to all only those predicates that do not apply to themselves. This reveals a sobering limit to stipulative definitions. We cannot make the heterological predicate exist by fiat. Thomson could have gone up from predicates to larger linguistic units. In particular, the liar paradox can be seen as a logically impossible sentence (or proposition or thought).

Anyone who takes these rebuffs to intuition seriously will be more disposed to accept a logician's curt answer to "Why is there something rather than nothing?" They will

be inclined to assimilate the possibility of an empty universe to the possibility of a barber who shaves all and only those who do not shave themselves. Under this analogy, revising logic to save the possibility of an empty universe is like revising logic to spare the possibility of Russell's barber.

Logically conservative responses to other paradoxes are repressive in some respects and liberating in others. The mix of liberation and suppression can be subtly accomplished at the level of notation. Bertrand Russell's notation in *Principia Mathematica* was intended to explain the possibility of mathematical knowledge by reducing mathematics to logic and set theory (which Russell regarded as a branch of logic). Russell tends to dwell on the doors opened by this notation. But he also gleefully observed that the ontological argument for the existence of God cannot even be formulated in *Principia* notation. This double-edged effect is natural because theories need to show us which possibilities are genuine and which are bogus. A theory is expressively incomplete only when it stops us from saying what we *want* to say.

5 Paradoxes Evoke Logical Analogies

The theme of repression and liberation can also be extended to styles of reasoning. The arbitrary individuals which populate pre-twentieth-century proofs were long known to have conflicts with laws of logic. For instance, an arbitrary number is neither odd nor even and yet an arbitrary number has the disjunctive property of being either odd or even! The anomalies were tolerated for lack of a better alternative. But once Gottlob Frege developed an adequate (though more complicated) quantification theory, arbitrary individuals were unceremoniously jettisoned.

However, the dominant trend has been in the direction of liberation. The paradox's potential for innovation is pregnant in the common definition of a paradox as an argument from incontestable premises to an unacceptable conclusion via an impeccable rule of inference. In Quine's (1966) terminology, some paradoxes are "veridical": their conclusions are true – just surprisingly so. These arguments have promising futures as instructive proofs.

A promising future is not destiny. The true conclusion of the veridical paradox does not guarantee that the argument is sound. Quine neglects the historical point that many veridical paradoxes are fallacious.

The Pythagoreans argued that the earth was a revolving, rotating sphere. Their conclusion is true and was as absurd to their contemporaries as Nicholas Copernicus' conclusion was to his contemporaries. But unlike Copernicus, the Pythagoreans argued fallaciously for their surprising truth. Typically, the brilliant argument for the initially absurd conclusion is only the beginning of a successful proof. The valuable part of the argument is its broad outlines, not its details. In other cases, the brilliant proof is only accidentally correct and is of no lasting value whatsoever.

Even a sound veridical paradox may have flaws. Some are circular. Others are vulnerable to refutation by logical analogy. The basic argument that all identities are necessary truths was regarded as sophistry before Saul Kripke championed it in *Naming and Necessity*. Almost all philosophers believed that physicists had established numerous contingent identities (such as Water = H_2O) and that the curious argument just par-

alleled Frege's deliberately absurd arguments against unbelieved identities. Kripke rehabilitated this ignored argument that all identities are necessary by offering an attractive alternative interpretation of scientific identities and raising doubts about the logical analogy.

Quine's distinction between veridical and falsidical paradoxes is also non-exhaustive. Consider Frank Ramsey's proof that there are exactly two Londoners who have exactly the same number of hairs on their heads. Ramsey notes that there are fewer than a million hairs on any one's head and there are more than a million Londoners. The conclusion follows by the pigeonhole principle: if there are more pigeons than pigeonholes, then at least two pigeons must share a hole. The existence of like-haired Londoners is not surprising. Even before hearing Ramsey's argument, Londoners agree that there is a high chance that two Londoners have the same number of hairs. What is paradoxical about Ramsey's proof is the *connection* between the premises and the conclusion. Londoners who are unfamiliar with the pigeonhole principle accept the premises and the conclusion but deny that the conclusion is entailed by the premises.

Given the logical interchangeability of propositions and inference rules, one could convert any inferential paradox into a propositional paradox. The paradoxical proposition is the conditional whose antecedent is the conjunction of the premises and whose consequent is the conclusion. Or one could do the reverse, turning propositional surprises into inference surprises. Perhaps, on the model of natural deduction systems, one could turn all propositional paradoxes into inference paradoxes. But since any system that allows deductions must have inference rules, there is an extra obstacle to a universal reduction to propositional paradoxes. In practice, we use systems that will ensure that some paradoxes are propositional while others are inferential.

Some of the most interesting paradoxes are both propositionally and inferentially paradoxical. The epistemicist argument that vague predicates have sharp thresholds has an intrinsically surprising conclusion and a further surprise that there could be a connection with such trivial facts such as 'Bald men are logically possible.'

Fallacious paradoxes are often instructive disasters. They suggest analogous arguments that avoid a critical mis-step while retaining some of the power of the original paradox.

The liar paradox has been especially fertile. Kurt Gödel's incompleteness theorems are self-conscious, delicately re-moldings of the Richard paradox. Alan Turing's first example of an uncomputable function, the halting problem, was based on the liar. Gregory Chaitin's (1986) theorem that a computer cannot fully predict its own performance was based on Berry's paradox. The liar paradox contains a powerful style of reasoning that does not inevitably ignite into contradiction. Like engineers using dangerous explosives to safely demolish buildings, meticulous thinkers gingerly titrate the paradoxical reasoning in their refutations of completeness or computability or predictability.

When Russell (1917) was calculating how many things are in the universe, he was led to a set that included everything. The number of things in this set must be the largest number because there is nothing further to add! Russell therefore accused Cantor of committing some subtle fallacy in his proof that there is no largest number.

A resemblance gave Russell second thoughts. The self-referential aspect of the universal set evokes a liar paradoxical set – a set that includes all and only those sets that

do not include themselves as members. If this set contains itself as a member, then it does not contain itself as a member. But if it does not contain itself as a member, then it does include itself as a member. Contradiction. The set cannot exist! Accordingly, Russell repudiated his objection to Cantor's proof.

Contradictions hurt. Russell sent news of the paradox to Gottlob Frege just as Frege's magnum opus on arithmetic was going to press. Frege hastily inserted a patch-up appendix. After this debacle, Frege never contributed anything of significance. Frege thought that we had infallible access to logical truths by intuition. Russell's paradox shows that we can have a clear intuition that something is possible even though it is demonstrably impossible.

Russell's paradox shows that naive set theory must be revised in a way that restricts the formation of sets. Accordingly, mathematicians have developed powerful set theories that unintuitively restrict the formation of sets. In particular, Zermelo-Fraenkel set theory has achieved the main objectives envisaged by the founders of set theory. But it has been a stop and go exploration. Each spurt ahead is accompanied by a look-about for unexpected trouble.

6 An Implication about the Nature of Paradox

An alternative definition of 'paradox' is as a set of individually plausible but jointly inconsistent propositions. This definition needs size limits to avoid counting Lewis Carroll's clerical inconsistencies as paradoxes. The immense scale of belief systems guarantees many such inconsistencies. As the number of propositions in a set increases, the number of conjunctions that can be formed from those propositions grows exponentially. This ensures that consistency checking is an NP-complete problem. Consequently, even a futuristic computer must eventually be overwhelmed and fail to detect many inconsistencies.

The infeasibility of the consistency check may explain why people tolerate large-scale inconsistency. However, their tolerance may also issue from their use of acceptance rules. People believe propositions to which they assign a negligibly small chance of falsehood. Small chances of error accumulate so the same people also believe the negation of the conjunction of their beliefs. Henry Kyburg's lottery paradox crisply formulates this anti-agglomerative pattern of belief formation in his lottery paradox. Large-scale inconsistency will also be precipitated by meta-beliefs. Meta-beliefs are a distinct but closely related source of inconsistency. Given that I really have first order beliefs, my belief that some of my beliefs are false is enough to ensure that not all of my beliefs can be true. For if all my first order beliefs are true, then my second order belief is not true.

People find small-scale inconsistencies painful – the smaller the set, the more intense the pain. Consequently, most paradoxes are formulated as a set of between three and five propositions. More propositions may be involved but only as lemmas leading up to the key members of the paradox. The inverse relationship between size and pain also explains why the best known arguments have so few premises. The argument-based definition of 'paradox' requires a small size constraint for the same reasons required by the set-based definition. A set of n jointly inconsistent propositions can be

turned into n valid arguments by using the remaining n – 1 members of the set as premises.

The set-based definition is often read as having individually consistent members. Indeed, they tend to be pictured as having the stronger property that each member is compatible with any non-exhaustive conjunction of the remaining members (like the inconsistent Carroll stories). The idea is that the victim of the paradox has exactly n ways to regain consistency corresponding to the n ways of rejecting a member of the paradox. Various '-isms' correspond to each solution (Rescher 1985). Logic has a role in structuring this menu of solutions. But it cannot dictate which member of paradox should be rejected.

Pierre Duhem gained fame for a similar thesis in science. Logic may dictate that we cannot believe both the theory and the conclusion based on an experiment. But it cannot tell us whether we should abandon the theory or the experiment. The physicist must instead rely on his 'good sense.'

Such thoughts provide a congenial environment for Gilbert Harman's (1986) distinction between proof and reasoning. Only reasoning concerns revision of one's beliefs and plans. Someone who believes 'If p then q' and then learns p need not conclude q. He could instead revise his belief that 'If p then q.'

The assumption of piecemeal consistency undergirds the hope that human inconsistency can be understood with a divide and conquer strategy. The divide and conqueror says that the inconsistency of a self-deceived person is the result of believing and disbelieving the same proposition in different ways (implicitly vs. explicitly, intuitively vs. theoretically, etc.). Thus the self-deceived widower *unconsciously* believes he is too old to marry his 18-year-old nanny but *consciously* believes he is not too old to marry his 18-year-old nanny. Another divide and conquer strategy is to analyze inconsistency as disagreement between parts of a person. This hope is not restricted to philosophy. When modular psychologists attribute inconsistency, they assume that there is a disagreement between self-consistent homunculi.

The divide and conquer strategy systematically fails for logical paradoxes. The belief that there is a barber who shaves all and only those who do not shave themselves is a logical contradiction that is not a conjunction of opposed propositions. Ditto for the massive family of paradoxes that involve violations of Thomson's theorem.

Many logical contradictions at the level of sentence logic are divisible. Human beings are comfortable with conjunction and negation, and so tend to couch propositions in a form amenable to the divide and conquer strategy. Since all sentential truth functions can be expressed in terms of conjunction and negation, one might hope to reduce all sentential contradictions to divisible contradictions. This seems psychologically unrealistic for belief in ostensibly non-conjunctive contradictions such as $\sim(P \supset P)$. The anthrocentricism of the reduction is also disturbing. Consider a Neanderthal who comfortably wields the Sheffer dagger function but can only fumble along with negation and conjunction. He can reduce all the contradictions of sentence logic to ones involving the dagger function. The Neanderthal's contradictions are not amenable to the divide and conquer strategy. Thus a human reduction of sentence contradictions to ones involving negation and conjunction would not show anything universal about the nature of contradictory belief.

A graver objection is that some logical contradictions are at the level of predicate logic. Many of these are clearly indivisible. When a contradiction has only variables and a single quantifier binds those variables, the contradiction is indivisible. Three illustrations: $(\exists x)(Fx \ \& \ \sim Fx)$, $(x)(x \neq x)$, $(\exists x)(y)(Cxy \ \& \ \sim Cxy)$. Any paradox that contains a logical falsehood as a member (or premise) is a logical paradox in the sentence-level sense that it contains a logical proposition.

Logical paradoxes are unique counterexamples to the principle that logic alone never implies a solution to a paradox. When one of the members of the paradox is a logical falsehood, logic *does* dictate what must be rejected. Since the inference to a logical truth is premiseless, the conclusion cannot be avoided by rejecting a premise.

Can logic itself be rejected? In *Beyond the Limits of Thought*, Graham Priest (1995) contends that the liar paradox shows that some contradictions are both true and false. He bridles against the limits of thought by rejecting standard inference rules.

If Priest is right, Duhem is wrong. For Duhem believed that standard logic structures the issues by specifying all the responses that have at least a bare chance of being true. If Priest is correct, then Duhem overlooked further true alternatives that a rational scientist might adopt. Thus Priest offers the scientist more freedom than Duhem. But is this the enhanced intellectual sweep of man who has dropped a false presupposition? Or is it the pseudo-liberty offered to Conrad Cornelius o'Donald o'Dell?

References

Carroll, Lewis (1895) What the tortoise said to Achilles. *Mind* 4, 278–80.

Chaitin, Gregory (1986) Information-theoretic computational complexity *and* Godel's theorem and information: In *New Directions in the Philosophy of Mathematics*, ed. Thomas Tymoczko. Boston, MA: Birkhauser.

Gibson, J. J. (1966) *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.

Harman, Gilbert (1986) *Change in View*. Cambridge, MA: MIT Press.

Priest, Graham (1995) *Beyond the Limits of Thought*. Cambridge University Press.

Quine, W. V. (1966) The ways of paradox. In W. V. Quine's *The Ways of Paradox*. New York: Random House, 1–18.

Rescher, Nicholas (1985) *The Strife of Systems*. Pittsburgh, PA: University of Pittsburgh Press.

Russell, Bertrand (1917) Mathematics and the metaphysicians. In his *Mysticism and Logic and other essays*. Watford, UK: Taylor, Garnet, and Evans, 74–96.

Russell, Bertrand (1957) On denoting. In R. C. Marsh (ed.), *Logic and Knowledge*. London: Allen & Unwin.

Sorensen, Roy (1996) Modal bloopers: why believable impossibilities are necessary. *American Philosophical Quarterly*, 33/1, 247–61. Reprinted in *The Philosopher's Annual 1996*, ed. Patrick Grim, Kenneth Baynes and Gary Mar. Atascardero, CA: Ridgeview, 1998, vol. XIX.

Thomson, J. F. (1962) On some paradoxes. In *Analytical Philosophy*, ed. R. J. Butler. New York: Barnes & Noble, 104–19.

Part IV

TRUTH AND DEFINITE DESCRIPTION IN SEMANTIC ANALYSIS

This page intentionally left blank

10

Truth, the Liar, and Tarski's Semantics

GILA SHER

1 Tarski's Theory of Truth

The most influential (and arguably, the most important) development in the modern study of truth was Tarski's 1933 essay "The Concept of Truth in Formalized Languages." The theory formulated in this essay distinguished itself from earlier theories in a number of ways: (1) it was a formal, that is mathematical (or quasi-mathematical) theory; (2) it offered a detailed, precise, and rigorous definition of truth; (3) it confronted, and removed, a serious threat to the viability of theories of truth, namely, the Liar Paradox (and other semantic paradoxes); (4) it made substantial contributions to modern logic and scientific methodology; (5) it distanced itself from traditional philosophical controversies; and (6) it raised a spectrum of new philosophical issues and suggested new approaches to philosophical problems.

Historically, we may distinguish two goals of Tarski's theory: a *philosophical* goal and a (so-called) *metamathematical* goal. Tarski's philosophical goal was to provide a definition of the ordinary notion of truth, that is the notion of truth commonly used in science, mathematics, and everyday discourse. Tarski identified this notion with the classical, *correspondence* notion of truth, according to which *the truth of a sentence consists in its correspondence with reality.* Taking Aristotle's formulation as his starting point – "To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true" (Aristotle: $1011^{b}25$) – Tarski sought to construct a definition of truth that would capture, and give precise content to, Aristotle's conception.

Tarski's second goal had to do with logical methodology or, as it was called at the time, metamathematics. Metamathematics is the discipline which investigates the formal properties of theories (especially mathematical theories) formulated within the framework of modern logic (first- and higher-order mathematical logic) as well as properties of the logical framework itself. Today we commonly call this discipline 'metalogic.' The notion of truth plays a crucial, if implicit, role in metalogic (e.g. in Gödel's completeness and incompleteness theorems), yet this notion was known to have generated paradox. Tarski's second goal was to demonstrate that 'truth' could be used in metalogic in a consistent manner (see Vaught 1974).

2 Tarski's Solution to the Liar Paradox

One of the main challenges facing the theorist of truth is the Liar Paradox. There are many versions of the paradox. (In antiquity, it was formulated in terms of 'lie,' whence its name, 'the liar paradox.') Tarski formulates the paradox as follows:

Let *c* abbreviate the expression 'the sentence printed on line 10 of the present page'. Consider the sentence:

c is not true.

It is clear that:

(1) *c* = '*c* is not true',
(2) '*c* is not true' is true iff (if and only if) *c* is not true.

Using the laws of classical logic, we derive a contradiction from (1) and (2):

(3) *c* is true iff *c* is not true.

What is the source of the paradox? Tarski's premises appear innocuous: (1) is an easily verified empirical statement, and (2) is an instance of an uncontroversial schema, namely, the *Equivalence Schema,*

(E) x is true iff *p*,

where '*p*' represents a sentence and 'x' a name of this sentence. (A simple instance of this schema is *'Snow is white' is true iff snow is white.*) Assuming the laws of classical logic are not the source of the paradox, it is natural to look for its source in *c*. One special feature of *c* is its predicating a property involving truth of itself. Tarski identifies this feature as responsible for the paradox. A language which contains its own truth predicate as well as names of all its sentences Tarski calls *semantically closed*. (More generally, any language which has the resources for describing its own syntax and contains its own semantic predicates (see below) is semantically closed.) Provided that such a language has a reasonable logical apparatus, it generates paradoxical sentences. Tarski concludes that semantically closed languages are inconsistent, that is they generate sentences that cannot be consistently given either the value True or the value False. In particular, the notion of truth (and other semantic notions) cannot be consistently defined for such languages. This conclusion is far from trivial: Natural languages are *universal* in the sense that anything that can be said by a speaker in any language can be said by him/her in his/her natural language. As such, natural languages are (generally) semantically closed, and truth (and other semantic notions) cannot be defined for such languages.

Not all languages, however, are semantically closed. Most mathematical and scientific languages are not. Such languages Tarski calls *semantically open*. Tarski's solution to the Liar Paradox is to restrict the definition of truth to open languages. This solution

requires that we think of languages as placed in a *hierarchy*: To define truth for a given open language L (our 'target language' or, in Tarski's terminology, 'object language'), we ascend to a higher (open) language, ML or meta-L, which has the resources for referring to all expressions (in particular, sentences) of L, and we formulate our definition of truth for L in ML. Truth for ML is defined in a third open language, MML, still higher in the hierarchy, and so on. This solution to the Liar Paradox is commonly called the *hierarchical solution.*

Tarski directs his attention to a particular family of open languages, namely, languages formalized within the framework of modern mathematical logic. Each such language includes (1) a set of logical constants containing a complete collection of truth-functional connectives (classically interpreted), the existential and/or universal quantifier, and possibly identity; (2) an infinite set of variables; and (3) a set (possibly empty) of nonlogical constants: individual constants, functional constants, and predicates. (Note: If L is a Tarskian language of order n, then for each $1 \leq i \leq n$, L has an infinite set of variables of order i, and the number of its symbols and well-formed expressions of order i is countable, that is it does not exceed the number of positive integers.) Since only interpreted sentences can be said to be true or false, Tarski restricts his attention to *interpreted* languages, that is languages whose primitive constants (logical and nonlogical) are fully interpreted. Such languages are naturally viewed as formalizations of scientific and mathematical languages as well as of open segments of natural languages. Tarski refers to such languages as "formalized languages" (or "formalized languages of the deductive sciences"). His goal is to construct a general method for defining truth for formalized languages.

3 Tarski's Method of Defining Truth for Formalized Languages

General principles

Given a formalized language L, the definition of truth for L is formulated in a metalanguage of L, ML. To define truth for L in ML we introduce an uninterpreted 1-place predicate, 'T,' into ML, and define it as a truth predicate for L, that is as a predicate satisfied by all and only true sentences of L. The definition of T is required to satisfy two conditions: (1) it has to be formally correct, that is avoid paradox, and (2) it has to be materially adequate, that is capture the idea that truth is correspondence with reality.

Formal correctness

To define T in a formally correct manner we follow the usual procedures for formally correct definitions, and in particular we make sure that the circumstance responsible for the Liar Paradox, namely, the truth for L being defined in L itself, does not arise. To this end we construct ML as an essentially stronger language than L, that is ML has expressions which are not translatable to L. In particular, the definition of T in ML is not translatable to L.

Material adequacy

To ensure that the definition of T is materially adequate, we require that it satisfy the following criterion ("convention," in Tarski's terminology):

Criterion (T)

A definition of T (in ML) is a materially adequate definition of truth for L iff it implies, for every sentence σ of L, an ML-sentence of the form

T(s) iff p,

where 's' stands for an ML name of σ and 'p' for an ML sentence with the same content as σ (a translation of σ to ML).

The idea is that given a sentence σ of L, an adequate definition of truth for L implies that σ has the property T just in case things in the world are as σ says. For example, if σ is the sentence 'Snow is white,' the definition of T implies that σ has the property T iff the stuff snow has (in reality) the property of being white. To satisfy this requirement, ML is required to contain, for each sentence σ of L, a sentence with the same content as σ. Using the notational conventions that 'snow is white' is an ML-name of the L-sentence 'Snow is white,' and 'snow is white' is an ML sentence with the same content as 'Snow is white,' the definition of T implies the ML-sentence:

T(snow is white) iff snow is white.

In constructing a definition of truth for L in ML we have to take into account the fact that the number of sentences in any language formalized within the framework of modern logic is infinite. A definition like

T(s) iff (s = snow is white and snow is white, or s = grass is red and grass is red, or . . .),

will not do, since such a definition would be infinitely long. To avoid this difficulty Tarski uses the *recursive* method. The recursive method enables us to define predicates ranging over infinitely many objects in a finite manner, provided certain conditions are satisfied. Such definitions are finitely long and they determine whether a given object falls under a given predicate in finitely many steps. I will not specify the conditions for recursive definitions here (for a good account see Enderton 1972, section 1.4), but the idea is that if every sentence of L is uniquely generated from finitely many atomic sentences (of L) by finitely many logical operations, and if the atomic sentences and logical operators of L are finitely specifiable, then truth for L can be recursively defined. Such a definition determines the truth value of each sentence of L based on (1) the truth values of its atomic constituents, and (2) its logical structure. For example, if the only logical constants (operators) of L are Negation and Disjunction, then truth for L is definable by specifying (1) the truth values of the atomic sentences of L, (2) a rule for determining the truth value of a Negation given the truth value of the negated sentence, and (3) a rule for determining the truth value of a Disjunction given the truth values of its disjuncts.

If L contains quantifiers, however, truth for L cannot be defined in this way. Sentences involving quantifiers are generated not from atomic sentences but from atomic formulas, including formulas with free variables (variables which are not in the

scope of any quantifier), and such formulas do not have a truth value. (For example, '$(\forall x)Px$' is generated from the atomic formula 'Px' which, having a free variable, has no truth value.) But truth for L can be recursively defined via an auxiliary notion, *satisfaction,* applicable to formulas. The notion of satisfaction is an intuitive notion: The atomic formula 'x is even' is satisfied (in the domain of the natural numbers) by 0, 2, 4, More generally, '$Rx_1, \ldots x_n$' is satisfied by an n-tuple of objects, $\langle a_1, \ldots, a_n \rangle$, iff a_1, $\ldots$, a_n (in that order) stand in the relation $\underline{R}$ (the relation referred to by 'R'). The definition of truth for L proceeds in two steps: (1) a recursive definition of satisfaction for L, and (2) a (nonrecursive) definition of truth for L based on (1).

Tarski's example

Tarski explained his method through an example. Using contemporary terminology, his example can be concisely described as follows.

Object language: L_C

The target language is the language of the calculus of classes (an interpretation of the language of Boolean algebra). I will refer to it as 'L_C.' L_C is an interpreted first-order language whose primitive vocabulary consists of the logical constants '$\sim$' (negation), '$\vee$' (disjunction) and '$\forall$' (the universal quantifier), the nonlogical constant '$\subseteq$' (a 2-place predicate interpreted as class inclusion), and variables, 'x_1', 'x_2', 'x_3', . . . , ranging over all objects in the domain, D_C, of L_C. D_C is a set of classes.

Meta-language: ML_C

Truth for L_C is defined in a meta-language, ML_C. ML_C relates to L_C in the way described above. In particular: (1) the syntax of L_C is describable in ML_C; (2) each constant of L_C has both a name and a translation (a constant with the same meaning) in ML_C; (3) ML_C has an undefined 1-place predicate, 'T,' designated as the truth predicate of L_C, as well as other predicates definable as semantic predicates of L_C; and (4) ML_C has variables of a higher-order than those of L_C (or a set-theoretical apparatus richer than that of L_C).

Definitions (in ML_C*)*

Notation: Let 'v_i' and 'v_j' be schematic symbols representing arbitrary variables, x_i and x_j, of L_C, and let 'Φ,' 'Ψ' and 'σ' be schematic symbols representing arbitrary expressions of L_C. Let '$\ulcorner$' and '$\urcorner$' be square quotes, where '$\ulcorner \Phi \vee \Psi \urcorner$' stands for 'the result of concatenating the formula Φ, the symbol '$\vee$' and the formula Ψ, in that order' (see Quine 1951). For each primitive constant c of L_C, let $\underline{c}$ be a name of c in ML_C and $\underline{\underline{c}}$ a translation of c to ML_C.

Formula (of L_C*) – Inductive Definition*

1. $\ulcorner v_i \subseteq v_j \urcorner$ is a formula.
2. If Φ is a formula, $\ulcorner \sim \Phi \urcorner$ is a formula.
3. If Φ and Ψ are formulas, $\ulcorner \Phi \vee \Psi \urcorner$ is a formula.

4. If Φ is a formula, $\ulcorner \underline{\forall} v_i \Phi \urcorner$ is a formula.
5. Only expressions obtained by 1–4 are formulas.

Sentence (of L_C)

σ is sentence iff σ is a formula with no free occurrences of variables.

Let g be any function which assigns to each variable of L_C an object in the domain, D_C, of L_C. We will call g 'an assignment function for L' and refer to $g(v_i)$ as 'g_i'.

Satisfaction (of a Formula of L_C by g) – Recursive Definition

1. g satisfies $\ulcorner v_i \subseteq v_j \urcorner$ iff $g_i \subseteq g_j$
2. g satisfies $\ulcorner \sim \Phi \urcorner$ iff $\sim$ (g satisfies Φ).
3. g satisfies $\ulcorner \Phi \vee \Psi \urcorner$ iff [(g satisfies Φ) $\vee$ (g satisfies Ψ)].
4. g satisfies $\ulcorner \underline{\forall} v_i \Phi \urcorner$ iff $\underline{\forall}$g′ (if g′ differs from g at most in g_i, then g′ satisfies Φ).

T (Truth of a Sentence of L_C)

T(σ) iff: (1) σ is a sentence, and (2) $\underline{\forall}$g(g satisfies σ).

4 Tarskian Semantics

Semantics and correspondence

Truth, for Tarski, is (as we have seen above) a correspondence notion. But truth is not the only correspondence notion. The discipline which studies correspondence notions in general Tarski calls 'semantics':

> We shall understand by semantics the totality of considerations concerning those concepts which, roughly speaking, express certain connexions between the expressions of a language and the objects and states of affairs referred to by these expressions. (Tarski 1936a: 401)

Some semantic notions express correspondence directly: *reference, satisfaction,* and *definition* are such notions: the name 'Mount Everest' *refers* to a mountain in the Himalayas; the formula 'x was assassinated' is *satisfied* by John Kennedy; the expression 'x^2' (where 'x' ranges over the natural numbers) *defines* the set $\{0,1,4,9,16, \ldots\}$. Other semantic notions, for example 'truth', express correspondence indirectly. Truth is a property of sentences rather than a relation between sentences and objects, but truth holds of a given sentence only if the *objects* referred to by this sentence possess the *properties* (*relations*) attributed to them by it. (To apply this principle to sentences containing logical constants we either construe the logical constants as referential constants – that is Identity as referring to the identity relation, Negation as referring to complementation, the Existential quantifier as referring to the higher-order property of nonemptiness, etc. – or we construe statements containing logical constants as *reducible* to statements (or formulas) satisfying the correspondence principle.)

Correspondence and disquotation

Some philosophers regard semantic notions as *disquotational* notions: a sentence enclosed in quotation marks has the property of being true iff this sentence, its quotation marks removed, holds (Ramsey 1927). Tarski, however, views the two analyses as equivalent:

> A characteristic feature of the semantical concepts is that they give expression to certain relations between the expressions of language and the objects about which these relations speak, or that by means of such relations they characterize certain classes of expressions or other objects. We could also say (making use of the *suppositio materialis*) that these concepts serve to set up the correlation between the names of expressions and the expressions themselves. (Tarski 1933: 252)

We can explain Tarski's view as follows: There are two modes of speech, an *objectual mode* and a *linguistic mode* ('material' mode, in Medieval terminology). The correspondence idea can be expressed in both modes. It is expressed by

> 'Snow is white' is true iff snow is white,

as well as by

> '"Snow is white" is true' is equivalent to 'Snow is white.'

In the objectual mode we say that a sentence attributing the (physical) property of whiteness to the (physical) stuff snow is true iff the (physical) stuff snow has the (physical) property of whiteness; in the linguistic mode we say that a sentence attributing (the semantic property of) truth to a sentence attributing whiteness to snow is equivalent to a sentence attributing whiteness to snow.

Logical semantics

One of the most important achievements of Tarskian semantics is its contribution to the definition of meta-logical notions ('logical consequence,' 'logical truth,' 'logical consistency,' etc.). Shortly after completing his work on truth, Tarski turned his attention to the notion of *logical consequence*. Prior to Tarski, 'logical consequence' was defined in terms of proof (the sentence σ is a logical consequence of the set of sentences Γ iff there is a logical proof of σ from some sentences of Γ). Gödel's incompleteness theorem showed, however, that the proof-theoretic definition of 'logical consequence' is inadequate: Not all theories formulated within the framework of modern logic can be axiomatized in such a way that all their true sentences are provable from their axioms. Using the resources of semantics on the one hand and set theory on the other, Tarski developed a general method for defining 'logical consequence' for formalized languages:

Semantic Definition of 'logical consequence'

σ is a logical consequence of Γ (in a formalized language L)

iff

there is no model (for L) in which all the sentences of Γ are true and σ is false. (Tarski 1936b)

This definition (which can easily be converted to a semantic definition of other metalogical notions – 'logical truth,' 'logical consistency,' etc.) played a critical role in turning *logical semantics*, or *model theory*, into one of the two main branches of contemporary (meta-)logic.

5 Three Criticisms of Tarski's Theory

While Tarski's theory of truth is widely viewed as one of the prime achievements of twentieth-century analytic philosophy, its philosophical significance has been repeatedly questioned. Among the main criticisms of Tarski's theory are: (A) Tarski's hierarchical solution to the Liar Paradox is applicable to artificial languages but not to "natural" languages; (B) Tarski's theory relativizes truth to language; (C) Tarski's definitions of truth are trivial.

Limitations of the hierarchical solution

Many philosophers find Tarski's solution to the Liar Paradox unsatisfactory on the ground that it does not enable us to define truth for natural languages. These philosophers are not dissuaded by Tarski's claims that: (1) it is impossible to define truth for natural languages, since being universal, such languages are inconsistent (Tarski 1933: 164–5), and (2) the hierarchical solution accounts for, and legitimizes, the use of 'true' in many segments of natural language, namely, all segments which are open and can be represented by artificial languages whose structure is precisely specified. In particular, truth can be defined for all segments used in the formulation of scientific theories (Tarski 1944: 347; 1969: 68). Soames (1999), for example, rejects the claim that natural languages are inconsistent. Others point out that Tarski's solution is too strict: it eliminates not only paradoxical uses of 'true' and related notions (e.g. 'false') in discourse, but also legitimate uses of these notions. One example, due to Kripke (1975), is the following: Consider two sentences, the one uttered by Dean and the other by Nixon during the Watergate crisis:

(4) All of Nixon's utterances about Watergate are false,

and

(5) Everything Dean says about Watergate is false.

This pair of sentences is perfectly consistent, yet there is no room for it in Tarski's hierarchy: According to Tarski's principles, (4) must belong to a language higher in the hierarchy than the language to which (5) belongs, and (5) must belong to a language higher in the hierarchy than the language to which (4) belongs. But this is impossible.

Triviality and relativity to language

It is common to interpret Tarski's theory as a *reductionist* theory or, more specifically, a theory whose goal is to *reduce* the notion of truth for a given language to the satisfaction conditions of the atomic formulas (the denotation conditions of the nonlogical constants) of this language. (To simplify the discussion I will ignore the case of atomic sentences containing logical constants, i.e. Identity). Given a language L, we determine the truth value of sentences of L by first listing the denotations of the primitive nonlogical constants of L, and then applying the recursive 'instructions' in the definition of truth for L to these lists. For example, if L is a language with two primitive nonlogical constants, an individual constant, '*a*,' and a 1-place predicate, 'P,' whose denotations are the number 1 and the set of all even natural numbers, respectively, we first prepare a denotation list for L, <'*a*,'1>, <'P,'{0,2,4,6, . . . }>, and then we calculate the truth value of sentences of L by applying the recursive rules in the definition of truth to this list: 'P*a*' is true (in L) iff $1 \in \{0,2,4,6, \ldots\}$, '~P*a*' is true (in L) iff 'P*a*' is false (in L), that is iff $1 \notin \{0,2,4,6, \ldots\}$, etc.

Two influential criticisms, based on this analysis, are: (1) Tarski's notion of truth is trivial; (2) Tarski's notion of truth is relative to language.

The triviality criticism

Tarski's definition of truth for a language L reduces the truth of sentences of L to the satisfaction of atomic formulas of L. But its treatment of atomic satisfaction is utterly uninformative. Instead of identifying a feature (or features) in virtue of which an object (an n-tuple of objects) satisfies a given atomic formula, it says that an object satisfies an atomic formula iff it belongs to a certain list. (In the above example, an object satisfies 'Px' iff it belongs to the list 0,2,4, ...) But a definition of this kind is a definition by *enumeration* ('x is a P iff x is 0 or is 2 or x is 4 or . . .'), and as such it lacks informative value.

This criticism is forcefully articulated in Field (1972). Field likens Tarski's definition of satisfaction to a definition by enumeration of a scientific concept. Consider, for example, a definition by enumeration of the concept *valence:*

$(\forall x)\{\text{Valence}(x) = n$
$\equiv [(x = \text{potassium} \;\&\; n = +1) \vee \ldots \vee (x = \text{sulfur} \;\&\; n = -2)]\}.$

The valence of a chemical element is an integer which represents the sort of chemical combinations the element will enter into based on its physical properties. A definition associating valences with physical properties of elements would be highly informative; a definition by enumeration, on the other hand, would be utterly trivial. (Expanding the definition from chemical elements to configurations of chemical elements by using

recursive entries will not change the situation: if the 'base' is trivial, the definition as a whole is trivial.)

Although Field is particularly concerned with one aspect of the Tarskian project, namely its success in reducing semantic notions to nonsemantic (specifically, physicalistic) notions, his criticism is not restricted to this aspect. The standards used in philosophy, Field says, should not be lower than those used in other sciences, and a method for defining truth by enumeration "has no philosophical interest whatsoever" (Field 1972: 102).

The relativity criticism

Another criticism of Tarski's theory (based on the above interpretation) concerns its relativization of truth to language. The argument can be summed up as follows: Tarski's method generates definitions of truth for particular languages, where (as we have seen before) the notion of truth for a given language is based on a list of denotations specific to that language (i.e. a list which cannot serve as a basis of a definition of truth for any other language). For that reason, Tarski's notion of truth is *relative to language.* Blackburn (1984: 267) compares Tarski's definitions of 'true in L_1,' 'true in L_2,' . . . , to definitions of 'well-grounded verdict on Monday,' 'well-grounded verdict on Tuesday,' . . . In the same way that the latter would not amount to a definition of the *absolute* notion 'well-grounded verdict,' so Tarski's definitions do not amount to a definition of the *absolute* notion 'true'. Just as there is no philosophical interest in the *relative* jurisprudential notion 'well-grounded verdict on day X,' so there is no philosophical interest in the *relative* semantic notion 'true in L.'

While the criticisms of Tarski's hierarchical solution to the Liar Paradox have motivated philosophers to construct new, nonhierarchical solutions to that paradox, the triviality and relativity criticisms have led many philosophers to give up hope of an informative theory of truth. Below I will describe a nonhierarchical solution to the Liar Paradox, due to Kripke, and I will offer a new interpretation of Tarski's theory as an informative theory, immune to the relativity and triviality criticisms.

6 Kripke's Solution to the Liar Paradox

In a 1975 paper, "An outline of a Theory of Truth," Kripke offered a new, nonhierarchical solution to the Liar Paradox. The idea underlying Kripke's proposal is this: Instead of defining truth for an infinite hierarchy of languages that do not contain their own truth predicate, we can define truth for a single language that does contain its own truth predicate in an infinite number of stages. In Tarski's method we start with a language L_0 which does not contain its own truth predicate, and construct stronger and stronger languages, L_1, L_2, . . . , each containing a truth predicate, T_1, T_2, T_3, . . . , for the previous language in the hierarchy. In Kripke's method we have a single language, L, which contains its own unique truth predicate, T, and we define the extension of T (i.e. the set of all sentences of L satisfying 'Tx') in stages: S_0, S_1, S_2, S_3, . . .

The definition of T proceeds by constructing two sets: Σ_1 – the extension of T, and Σ_2 – the counter-extension of T. Σ_1 is the set of all true sentences of L in the domain D of L, Σ_2 is the set of all false sentences of L in D plus all objects in D which are not sen-

tences of L. (D may contain codes of sentences of L instead of sentences of L, but for the sake of simplicity I will assume it contains (only) the latter.) Let us think of L as a union of a Tarskian hierarchy, $\cup\{L_0, L_1, L_2, \ldots\}$, where '$T_1$', '$T_2$', '$T_3$', ... represent partial applications of T. Σ_1 and Σ_2 are constructed in stages as follows:

Stage 0: $\Sigma_1 = \varnothing$

$\Sigma_2 = \{a \in D$: a is not a sentence of L$\}$

Stage 1: $\Sigma_1 = \{a \in D$: a is a true sentence of L_0 **or** a is a true sentence of L whose truth value is logically determined based on the truth value of sentences of $L_0\}$

$\Sigma_2 = \{a \in D$: a is a false sentence of L_0 **or** a is a false sentence of L whose truth-value is logically determined based on the truth-value of sentences of L_0 **or** a is not a sentence of L$\}$

Stage 2: $\Sigma_1 = \{a \in D$: a is a true sentence of L_0 or L_1, **or** a is a true sentence of L whose truth-value is logically determined based on the truth-value of sentences of L_0 or $L_1\}$

$\Sigma_2 = \{a \in D$: a is a false sentence of L_0 or L_1, **or** a is a false sentence of L whose truth-value is logically determined based on the truth-value of sentences of L_0 or L_1, **or** a is not a sentence of L$\}$

Thus, if 'Snow is white' and 'Snow is green' are sentences of L, then since both belong to the L_0 part of L, in stage 0 neither belongs to Σ_1 or Σ_2. In stage 1, 'Snow is white' and '~ Snow is green' are among the sentences added to Σ_1, and 'Snow is green' and '~ Snow is white' are among the sentences added to Σ_2. In stage 2, 'T "Snow is white"' and 'T "~ Snow is green"' are among the sentences added to Σ_1, and 'T "~ Snow is white"' and 'T "Snow is green"' are among the sentences added to Σ_2. And so on. The list of stages can be extended into the transfinite, using standard set theoretic methods. Thus we can have transfinite stages ω, $\omega + 1$, $\omega + 2$, ..., (where ω is the smallest infinite ordinal), including higher limit ordinals. (The details of the transfinite stages can be omitted.)

Throughout the finite stages, Σ_1 and Σ_2 are continuously extended and their extensions are *forced* by (1) the rules for the nonlogical, nonsemantic primitive constants of L (i.e. the rules determining the denotations of these constants and the truth/satisfaction of sentences/formulas composed of these constants (and, possibly, variables) – eventually, facts about what constant denotes what object, property or relation, what object has what nonlogical property and/or what objects stand in what nonlogical relation); (2) the rules for the logical constants of L; and (3) the rules for the semantic constants of L. (See Rules I–III below.) Thus, 'Snow is white' and 'NOT snow is green' must be added to Σ_1 in Stage 1 (due to facts concerning the denotations of 'snow,' 'white,' and 'green' and the color of snow, as well as the semantic rule for 'NOT'), 'True "Snow is white"' must be added to Σ_1 in Stage 2 (due to the semantic rule for 'true' and the fact that 'Snow is white' belongs to Σ_1 in stage 1), 'True "True 'Snow is white' "' must be added to Σ_1 in Stage 3 (due to the rule for 'true' and the fact that 'True "Snow is white"' belongs to Σ_1 in Stage 2), etc. And similarly for Σ_2. We say that all the sentences placed in Σ_1 and Σ_2 in the finite stages are *grounded*. However, since no sentence of L contains infinitely many occurrences of 'T,' and in particular, infinitely many embedded occurrences of 'T' (or other semantic predicates), eventually we arrive at a stage in

which neither Σ_1 nor Σ_2 is properly extended. We call such a stage a *fixed point.* It is important to note that not all sentences of L belong to either Σ_1 or Σ_2 in the *least fixed point.* For example, Liar sentences as well as sentences like

(10) T(10)

do not. How does Kripke deal with such sentences?

To deal with paradoxical sentences Kripke constructs T as a *partial* truth-predicate and L as a language with *truth-value gaps*: some sentences of L are either in the extension of T or in its anti-extension, but other sentences are in neither; some sentences of L have a truth value, others do not. All paradoxical sentences are truth-valueless in Kripke's semantics, but sentences like (10) can either be assigned a truth value (True or False) in later stages, or remain truth-valueless.

I will not formulate Kripke's semantics for L in detail here. But the following are its main principles:

I *Rules for determining the denotation, satisfaction and truth-value of expressions of L_0 (the L_0 part of L)*
Same as in Tarski's semantics.

II *Rules for determining the truth-value and satisfaction of sentences and formulas of L governed by logical constants*
Based on Kleene's strong 3-valued semantics. (Coincides with Tarski's semantics in the bivalent part of L, in particular, in the L_0 part of L.)

Let σ_1 and σ_2 be sentences of L. Then:

$\ulcorner\neg\sigma_1\urcorner$ is
- true if σ_1 is false
- false if σ_1 is true
- undefined otherwise

$\ulcorner\sigma_1 \vee \sigma_2\urcorner$ is
- true if at least one of σ_1 and σ_2 is true
- false if both σ_1 and σ_2 are false
- undefined otherwise

Let Φ be a formula of L, let g be an assignment function (as in Section 3), and let us use 'Φ is true under g' for 'g satisfies Φ'. Then:

$\ulcorner\forall v_i\Phi\urcorner$ is
- true: Φ is true under every g′ which differs from g at most in g_i
- false under g if: Φ is false under some g′ which differs from g at most in g_i
- undefined otherwise

III *Semantic rule for sentences governed by the truth predicate, T, of L (Kripke's version of Criterion (T)):*

Let σ be a sentence of L and s a name of σ in L. Then:

$$\ulcorner T(s) \urcorner \text{ is } \begin{cases} \text{true} & \text{iff} \quad \sigma \text{ is true} \\ \text{false} & \text{iff} \quad \sigma \text{ is false} \end{cases}$$

The definition of T can be viewed as completed in any of the fixed-points. If we view it as completed in the least fixed-point, then only grounded sentences are in the extension of T. If we see it as completed in later fixed-points, some ungrounded sentences (e.g. (10)) may also be in the extension of T. Paradoxical sentences are never in the extension of T.

Two noteworthy features of Kripke's method are: (1) it does not uniquely determine the truth predicate of a given closed language; and (2) it allows empirical circumstances to determine whether a sentence is paradoxical in a given language. The first point should be clear by now: the semantic status of some sentences (i.e. being true, false, or truth-valueless) is 'forced' by the semantic rules, that of others is a matter of choice or convention. Grounded and paradoxical sentences fall under the first category, ungrounded and unparadoxical sentences fall under the second.

The role of empirical circumstances

One important intuition captured by Kripke's proposal is that semantic properties of sentences (being true, false, ungrounded, paradoxical, etc.) are often determined by empirical circumstances. Consider, for example, the sentence

(11) $(\forall x)(Px \supset Tx)$

of a Kripkean language L. If P is an empirical predicate satisfied by exactly one object, *a*, then: if *a* = 'Snow is white,' (11) is true; if *a* = 'Snow is green,' (11) is false; if *a* = (11), (11) is ungrounded; and so on. And these semantic features hold or do not hold of (11) empirically. The same applies to

(12) $(\forall x)(Px \supset \sim Tx)$.

If the only object satisfying 'Px' is 'Snow is green,' (12) is true; if it is 'Snow is white,' (12) is false; if it is (12) itself, (12) is paradoxical. And the truth, falsity, or paradoxicality of (12) are due to empirical circumstances. In making statements, Kripke observes, we often take a risk. Under certain circumstances a sentence is grounded and true, under others – ungrounded and paradoxical.

This feature of Kripke's theory enables it to assign a truth value to sentences which (in the specific circumstances of their utterance) are not paradoxical, yet are regarded by Tarski as illegitimate. Let us go back to (4) and (5). If at least one statement made by Dean about Watergate is true and all Nixon's statements about Watergate other than (5) are false, then (4) is true and (5) is false.

The ghost of Tarski

While Kripke's method provides a semantics for languages containing their own truth predicate, the account itself is carried out in a Tarskian meta-language. Furthermore, some truths about sentences of a given Kripkean language L are, though expressible in L, true only in its meta-language, ML. Thus, if σ is a Liar sentence of L, the statements 'σ is not true,' 'σ is ungrounded' and 'σ is paradoxical' are true in ML but lack a truth value in L. In Kripke's words: "The ghost of the Tarski hierarchy is still with us" (Kripke 1975: 714).

Kripke's relegation of certain truths to the meta-language is not accidental. It is the means by which he avoids the so-called *strengthened Liar paradox*. The strengthened Liar paradox arises in languages with truth-value gaps as follows: Let

(13) ~T(13)

be a sentence of a 3-valued language L and let T be a truth predicate of L satisfying Kripke's version of Criterion T. Then: T((13)) *iff* (13) *iff* ~T(13).

Kripke avoids the strengthened Liar paradox by rendering (13) undefined but its meta-linguistic correlate, 'the sentence (13) of L is not true,' true. This means that Kripke's method falls short of providing a complete semantics for natural languages which, being universal, have no richer meta-languages.

Kripke's solution to the Liar Paradox is not the only alternative to Tarski's solution. For other alternatives see Martin (1984), Gupta and Belnap (1993), and others.

7 A Reinterpretation of Tarski's Theory

The deflationist approach to truth

The view that the base entries in Tarski's definitions render them uninformative has led some philosophers to search for an informative base for Tarski's definitions. Field (1972) suggested that instead of using lists of reference as a basis for a definition of truth, we use a general, informative theory of reference as such a basis, and pointed to Kripke's (1972) outline of a causal theory of reference as a promising starting point. But the slow progress and difficulties involved in the development of an informative and general theory of reference led Field (1986) and others to adopt a so-called *deflationist* or *minimalist* attitude towards truth.

The deflationist attitude is reflected by such statements as:

> [T]ruth is entirely captured by the initial triviality [that each proposition specifies its own condition for being true (e.g. the proposition *that snow is white* is true if and only if *snow is white*)]. (Horwich, 1990: xi)
>
> Unlike most other properties, *being true* is insusceptible to conceptual or scientific analysis. (*Ibid.*: 6)

> [The theory of truth] contains no more than what is expressed by the uncontroversial instances of the equivalence schema,
>
> (E) It is true *that p* if and only if *p*. (*Ibid.*: 6–7)

While deflationists differ on many issues, most agree that a theory of truth need not be more informative than Tarski's theory. Some would like to extend Tarski's definitions to a greater variety of linguistic structures: indexicals, adverbs, propositional attitudes, modal operators, etc., but none requires a more substantive analysis. According to deflationists, "the traditional attempt to discern the *essence* of truth – to analyze that special quality which truths supposedly have in common – is just a pseudo-problem". (Horwich, 1990: 6) There is no substantive common denominator of all truths, and therefore there is no substantive theory of truth. The task of a theory of truth is to generate a list of all instances of the Equivalence schema, and regardless of how this list is generated, the theory of truth is still a collection of trivialities.

Critique of the deflationist approach

The deflationist approach is based on a traditional conception of theories: A theory of a concept X is a theory of the common denominator of all objects falling under X. If the common denominator of all these objects is trivial, X is trivial and a theory of X is a collection of trivialities. This conception of a philosophical theory is, however, based on an unfounded assumption: namely, that the content of a given concept X is the common denominator of all instances of X. It is quite clear that the content of some concepts is not exhausted, or even close to being exhausted, by the common denominator of their instances. The concept of *game* is a case in point (Wittgenstein, 1958). Yet if 'game' is not a common-denominator concept, it is clearly not an empty or a trivial concept. And neither is a theory of games empty or trivial. A theory of games may not be able to condense all there is to say about games into a single principle, expressible by a single formula, but it could identify a number of significant principles governing games and describe their nature, workings, interrelations, and consequences in a general and informative manner.

The question arises as to whether Tarski's theory of truth is – or can be made to be – substantive in this (non-traditional) sense.

What does Tarski's theory actually accomplish?

One thing that both defenders and critics of Tarski's theory agree about is its substantial contribution to logic (see above). Now, it is striking that Tarski's theory does not make similar contributions to other disciplines. While Tarski's definition of truth for a language L yields, all by itself, a definition of *logical consequence* for L (assuming ML has a sufficiently rich set-theoretical apparatus), it does not yield (all by itself) definitions of *epistemic*, *modal*, *physical*, or *biological consequence* for L. (Examples of the latter kinds of consequence are: '*a* knows that P; therefore, *a* believes that P,' 'Necessarily P; therefore Possibly P,' 'The force exerted on body *a* at time *t* is zero; therefore the acceleration of *a* at *t* is zero,' '*a* is a human female; therefore *a* does not have a Y chromosome,' etc.)

Why does Tarski's theory yield an account of *logical* consequence, but not of other types of consequence? What features should a theory of truth have in order to yield a concept of consequence of type X?

The answer to this question is quite clear. A consequence relation is a relation of preservation (or transmission) of truth: If C stands in a consequence relation R to a set of sentences, Γ, and all the sentences of Γ are true, then their truth is preserved through R (or is transmitted to C through R). If R is a relation of consequence of type X, the preservation (or transmission) of truth is due to the *X-structure* of the sentences of Γ and X, that is due to the *content* and *organization* of constants of type X in these sentences (where for non-X constants, only their identities and differences, but not their content or interrelations, play a role). Thus, if C stands to Γ in a relation of *logical* consequence, this is due (except in the trivial case of $C \in \Gamma$) to the *logical* structure of the sentences involved; if C stands to Γ in the relation of modal, epistemic, physical, or biological consequence, this is due to the modal, epistemic, physical, or biological structure of those sentences. To yield a definition of consequence of type X for a language L, a definition of truth for L has to specify the contribution of X-structure to the truth value of sentences of L. Tarski's definition of truth for a language L is tuned to the *logical* structure of sentences of L; therefore, it gives rise to the notion of *logical* consequence for L. (Note that due to the generality of logic, it is common to conceive of non-logical consequences of type X as based not only on the content and interrelations of the X vocabulary, but also on the interrelations of the X vocabulary and the logical vocabulary. Yet what renders these consequences X-consequences is the role played by the X-vocabulary.)

These observations suggest that what Tarski's theory actually accomplishes is an account of the *contribution of logical structure to truth*. Tarski's theory tells us how the logical structure of a given sentence affects its truth value, not how other types of structure (modal, physical, . . .) do. Tarski's theory, on this interpretation, is a theory of a specific, albeit basic and general constituent of truth, namely, its logical constituent. Its goal is to describe, in an exhaustive, systematic and informative manner, that part of the truth-conditions of sentences which is due to their logical structure. This interpretation explains why Tarski's theory of *truth* is so important and fruitful in *logic*. Furthermore, it shields Tarski's theory from the relativity and triviality criticisms.

Relativity

While the role played by nonlogical constituents of sentences in determining their truth conditions is relative to language (in Tarski's theory), the role of the logical constituents is not. The denotation lists for the nonlogical constants vary from one Tarskian language to another, but the semantic rules for the logical constants are fixed across languages. The difference between Tarski's treatment of logically-structured and nonlogically-structured formulas of a given language is a difference between *rule* and *applications*. To calculate the truth value of a sentence – say, 'John loves Mary and John loves Jane' – of a Tarskian language L we take the *fixed* truth condition associated with 'and' in Tarski's method and apply it to the truth conditions of 'John loves Mary' and 'John loves Jane' in L. We may say that the *principles* governing the contribution of logical structure to truth are *absolute*; their *instances* (*applications*) – *relative* to language.

But this is the case with any theory: the rule of, say, addition, is the same in all applications of arithmetic, but in biology this rule operates on sets (quantities) of biological entities, while in theoretical physics it operates on sets (quantities) of abstract physical entities.

Triviality

The triviality criticism, like the relativity criticism, is directed at Tarski's treatment of the nonlogical constituents of truth. Considering Tarski's definition of truth for a given language L, the claim is that the satisfaction and denotation conditions for formulas and terms with no logical constants of L are given by enumeration (i.e. based on lists), and as such they trivialize the entire definition. While this criticism is warranted with respect to the first interpretation of Tarski's theory, it is unwarranted with respect to the second. On the first interpretation, Tarski's theory is a *reductionist* theory. Its task is to reduce the notion of truth for a given language to the satisfaction and denotation conditions of its nonlogically-structured formulas and its nonlogical constants. As such, the burden of informativeness falls on its *nonlogical* entries. Since these are trivial, the definition as a whole is trivial. But on the second, *logical* interpretation, the burden of informativeness falls on the *logical* entries. (The nonlogical entries play a merely auxiliary role.) So long, and to the extent that, the logical entries are informative, the definitions of truth are informative.

Are the logical entries in Tarski's definitions informative? To be informative, the logical entries have to describe the truth conditions associated with different logical structures based on principles, rather than by enumeration. Now, on a first reading, the logical entries in Tarski's definitions are not very informative. Take the logical connectives. The entries for Negation and Disjunction essentially say that ⌜**not** σ⌝ is true iff σ is **not** true, and that ⌜σ **or** ς⌝ is true iff σ is true **or** ς is true. These entries do not *explain* the satisfaction conditions of 'not' and 'or'; they take them as given. ('Not' in the definiens merely repeats 'not' in the definiendum.) But on a less literal and more charitable interpretation we may view the entries for the logical connectives as implicitly referring to the highly informative Boolean, or truth-functional, account of these connectives. The Boolean account provides (1) an informative a criterion of logicality for connectives, and (2) a systematic characterization of the satisfaction conditions of each logical connective based on this criterion. According to this characterization, Negation is characterized by a 1-place Boolean function, f_, defined by: f_(T) = F and f_(F) = T, Disjunction is characterized by 2-place function f_V, defined by: f_V(T,T) = f_V(T,F) = f_V(F,T) = T and f_V(F,F) = F, and these definitions are precise and informative. In 1933 there did not exist an analogous criterion for logical predicates and quantifiers, but in later years such a criterion, and a systematic characterization of the satisfaction conditions of individual logical predicates and quantifiers based on it, have been developed. (See Mostowski 1957; Lindström 1966; Tarski 1966; Sher 1991 and others.) Today, therefore, it is possible to avoid the triviality criticism altogether by expanding Tarski's definitions to languages containing any logical constant satisfying this criterion and constructing (interpreting) the satisfaction entries for the logical constants as referring to the informative characterizations of these constants based on this criterion. (For further details and examples see Sher 1999b, Sections 6, 7, and 9).

8 Truth Beyond Logic

Aside from its direct contributions to pure logic, Tarski's work on truth has indirectly contributed to other fields as well. Kripke (1963) developed a semantics for modal logic which incorporates elements from Tarski's logical semantics; Hintikka (1962) and others developed a semantics for epistemic statements based on Tarski's semantics; Davidson (1980, 1984) has begun an influential project of developing a general theory of meaning for natural languages based on Tarski's method; etc. How far Tarski's theory can be extended beyond logic without losing its informativeness is an open question.

References

Aristotle (1941) *Metaphysics. The Basic Works of Aristotle*, ed. R. McKeon. New York: Random House.

Blackburn, S. (1984) *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Oxford University Press.

Davidson, D. (1980) *Actions and Events*. Oxford: Oxford University Press.

Davidson, D. (1984) *Truth and Interpretation*. Oxford: Oxford University Press.

Devitt, M. (1984) *Realism and Truth*. Oxford: Blackwell.

Enderton, H. B. (1972) *A Mathematical Introduction to Logic*. San Diego: Academic Press.

Field, H. (1972) Tarski's theory of truth. *Journal of Philosophy*, 69, 347–75.

Field, H. (1986) The deflationary conception of truth. *Fact, Science, and Modality*, eds. G. MacDonald and C. Wright. Oxford: Blackwell, 55–117.

Gupta, A. and Belnap, N. (1993) *The Revision Theory of Truth*. Cambridge, MA: MIT.

Hintikka, J. (1962) *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell.

Kripke, S. (1963) Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16, 83–94.

Kripke, S. (1972) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kripke, S. (1975) Outline of a theory of truth. *Journal of Philosophy*, 72, 690–716.

Lindström, P. (1966) First order predicate logic with generalized quantifiers. *Theoria*, 32, 186–95.

Martin, R. L. (ed.) (1984) *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford University Press.

Mostowski, A. (1957) On a generalization of quantifiers. *Fundamenta Mathematicae*, 44, 12–36.

Quine, W. V. (1951) *Mathematical Logic*, revised edn. Cambridge, MA: Harvard University Press.

Ramsey, F. (1927) Facts and propositions. *The Foundations of Mathematics*. Paterson, NJ: Littlefield, Adams, 1960, 138–55.

Sher, G. (1991) *The Bounds of Logic: A Generalized Viewpoint*. Cambridge, MA: MIT.

Sher, G. (1999a) On the possibility of a substantive theory of truth. *Synthese*, 117, 133–72.

Sher, G. (1999b) Is logic a theory of the obvious? *European Review of Philosophy*, 4, 207–38.

Soames, S. (1999) *Understanding Truth*. New York: Oxford University Press.

Tarski, A. (1933) The concept of truth in formalized languages. In Tarski (1983) 152–278.

Tarski, A. (1936a) The establishment of scientific semantics. In Tarski (1983) 401–8.

Tarski, A. (1936b) On the concept of logical consequence. In Tarski (1983) 409–20.

Tarski, A. (1944) The semantic conception of truth. *Philosophy and Phenomenological Research*, 4, 341–76.

Tarski, A. (1966) What are logical notions? *History and Philosophy of Logic*, 7, 143–54.
Tarski, A. (1969) Truth and proof. *Scientific American*, 220, 63–77.
Tarski, A. (1983) *Logic, Semantics, Metamathematics*. 2nd edn. Indianapolis, IN: Hackett.
Vaught, R. L. (1974) Model theory before 1945. *Proceedings of the Tarski Symposium*, eds. L. Henkin *et al.* Providence, RI: American Mathematical Society.
Wittgenstein, L. (1958) *Philosophical Investigations*. 2nd edn. Oxford: Basil Blackwell.

11

Truth, the Liar, and Tarskian Truth Definition

GREG RAY

Alfred Tarski's work on truth has become a touchstone for a great deal of philosophical work on truth. A good grasp of it is critical for understanding the contemporary literature on truth and semantics. In this chapter, I will present a fresh interpretation of Tarski's view, one which aims to draw it out more fully in areas of philosophical interest. This has required extrapolation (e.g. drawing explicit implications for concepts and properties) and reverse engineering (e.g. introducing the notion of full conceptual warrant) for which I will not offer textual justification here. My purpose is to introduce Tarski's central ideas briefly and in the most tenacious way I can. It is my hope that this brief study will prove useful as a basis for further investigation.

1 Truth

Our topic of philosophical concern is truth, but we will be almost entirely concerned with *the concept of sentential truth*, that is the relational concept of (something's being) *a true sentence of* (some language). We will also have to deal with certain language-specific truth concepts, such as the concept of *a true sentence of English*. In addition to sentential truth, one can speak of *doxic truth* and *propositional truth* – these being the concepts of *a true belief* and *a true proposition*, respectively. Some think the concept of propositional truth more fundamental than others. We need not join this debate here, if we are careful to draw our conclusions with appropriate care.

The T-strategy

Underlying Tarski's work is a basic observation about sentential truth, namely that claims such as

'snow is white' is a true sentence of English if and only if snow is white
'neige est blanche' is a true sentence of French if and only if snow is white

seem quite obvious and unexceptionable, and they also state necessary and sufficient conditions for the truth of the sentences mentioned in them without appeal to any

further semantic notions. Tarski's starting point is the idea that such statements might be helpful in characterizing sentential truth. If *one* of these sentences gives necessary and sufficient conditions for the application of 'is a true sentence of French' to one sentence, then together a full set of such sentences could give necessary and sufficient conditions for the application of 'is a true sentence of French' to any sentence of French. Thus, if some finite way of expressing this infinite set of conditions could be found, we would have the makings of a definition of 'is a true sentence of French,' and this would be a start. The sense in which such a definition might characterize or 'capture' a truth concept will be an important question to take up later. Let us call the strategy of definition we've outlined, *the T-strategy*. The sentences we are concerned with are known as *T-sentences*.

DEFINITION A *T-sentence in English for* a language L, is any sentence which may be obtained from the *T-schema*,

s is a true sentence of L if and only if p,

by substituting for 's' a syntactic description in English of a sentence of L, and substituting for 'p' a translation into English of that same sentence of L.

I have characterized T-sentences expressed *in* English. There are obvious correlates for other meaningful languages. The T-strategy suggests that, if I wanted to characterize sentential truth for the Xanadic language (my *object language*) and I wanted to express myself in Polish (my *meta-language*), I should find some finite way of expressing the set of conditions given by the class of T-sentences in Polish for Xanadic. More generally, to make a definition realize the T-strategy for a language L in a language M, we will insist the definition satisfy the following condition of adequacy.

CONVENTION T All the T-sentences in M for L are theorems of that theory which consists of our definition statement (plus, perhaps, some axioms about syntax and sequences).

The problem of generality

There is no obvious generalization of the T-sentences – one which would be equivalent to stating that infinitude of sentences. So, the technical challenge of the T-strategy is finding a *finite* way of expressing what the relevant T-sentences express. Call this the *problem of generality*.

Tarski was not the first to see that something like T-sentences might be used to characterize a truth concept nor the first to tackle the problem of generality. Tarski seemed to think of certain familiar pronouncements about truth as unsatisfactory attempts at solving the problem. One such was Aristotle's famous dictum: *to say that what is, is not or that what is not, is false, while to say that what is, is or that what is not, is not, is true.* F. P. Ramsey pursued something like a T-strategy for doxic truth and expressed clearly the challenge of achieving generality.

> Suppose a man believes that the earth is round; then his belief is true because the earth *is* round; or generalizing this, if he believes that A is B his belief will be true if A is B and false

> otherwise. It is, I think, clear that in this last sentence we have the meaning of truth explained, and that the only difficulty is to formulate this explanation strictly as a definition. If we try to do this, the obstacle we encounter is that we cannot describe all beliefs as beliefs that A is B since the propositional reference of a belief may have any number of different more complicated forms. A man may be believing that all A are not B, or that if all A are B, either all C are D or some E are F, or something still more complicated. We cannot, in fact, assign any limit to the number of forms which may occur, and must therefore be comprehended in a definition of truth; so that if we try to make a definition to cover them all it will have to go on forever, since we must say that a belief is true, if supposing it to be a belief that A is B, A is B, or if supposing it to be a belief that A is not B, A is not B, or if supposing it to be a belief that either A is B or C is D, either A is B or C is D, and so on *ad infinitum*. (Ramsey 1929: 9)

Tarski thought that he had an elegant solution to the generality problem, but recognized two significant obstacles – natural language and the Liar Paradox. We will discuss these in turn.

Conceptual status of T-sentences

Before we proceed, however, let us consider more closely the conceptual status of T-sentences. What one might hope to get out of a definition of 'is a true sentence of' (or 'is a true sentence of French') using the T-strategy depends on the status of these T-sentences. So long as the mentioned sentence is indexical-free, tenseless, and not vague or ambiguous, each T-sentence gives a *materially* necessary and sufficient condition for the application of the truth predicate in question to the mentioned sentence. What more can be said? We should not rush to claim that T-sentences express conceptual truths or are analytically true. There is, nonetheless, some interesting conceptual linkage between the concept of sentential truth and the T-sentences – a linkage which it is our current task to elucidate. Consider again a T-sentence such as

> 'neige est blanche' is a true sentence of French if and only if snow is white.

Such a sentence recommends itself to us, because it seems, roughly, that one who has the proper linguistic understanding knows it to be true. There is something to this idea. Let us begin with a case simpler than T-sentences.

DEFINITION For language M and sentence, s, of M, we shall say that s has *simple conceptual warrant in* M iff one who understands s (as a sentence of M), is in a position to know (on non-truth-functional grounds) that if (1) each predicate of s is subserved in M by the concept it expresses in M, and (2) each singular referring term of s refers in M, then s is a true sentence of M.

Understanding s as a sentence of M, as used here, is meant to imply of the agent that he or she,

1. for each predicate, p, of s, grasps the concept, c, expressed in M by p and knows of c that it is expressed in M by p, and

2. associates with each singular referring term of s a condition and knows that it uniquely picks out the referent in M of the term if such there be, and nothing otherwise.

To explain what is intended by saying that a predicate is subserved by a concept, I will avail myself of the useful fiction that concepts come supplied with explicit *application rules* which say what sorts of things are supposed to be included or excluded by the concept.

DEFINITION A concept, c, *subserves* a predicate, p, of language M iff for all x in the domain of discourse of M, (1) if the application rules for c imply that x falls under c, then p applies in M to x, and (2) if the application rules for c imply that x fails to fall under c, then p fails to apply in M to x.

Ordinarily, of course, if a predicate expresses a concept, that concept subserves the predicate. However, a concept could subserve a predicate which did not express it, and, just possibly, a predicate could express a concept that did not subserve it.

Sentences with simple conceptual warrant evidently include (1) the analytically true; (2) sentences free logics treat specially, such as 'if Vulcan is green, then Vulcan is green'; as well as (3) some more interesting cases involving vacuous names, such as, 'if Vulcan is a planet, then Vulcan is a heavenly body.' These sentences have exceptional conceptual credentials, though not all are guaranteed to be true. We note in passing that they are all of a sort that we would be entitled to rely on for the purposes of scientific theorizing – at least until such time that it became known that 'Vulcan' fails to refer.

T-sentences do not have simple conceptual warrant, but an extension of the same idea applies to them.

DEFINITION For sentences M and L, and for t, a T-sentence in M for some language, L, (where t has form $\ulcorner \delta$ is a true sentence of λ iff $\Gamma\urcorner$), we shall say that t has *subtle conceptual warrant in* M just in case one who

(1) understands t as a sentence of M,
(2) *recognizes that the sentence denoted in M by* δ *is a sentence of the language denoted in M by* λ,
(3) *understands the sentence denoted in M by* δ *as a sentence of the language denoted in M by* λ.

is in a position to know (on non-truth-functional grounds) that

> if each predicate of t is subserved in M by the concept it expresses in M, and each singular referring term of t refers, then t is a true sentence of M.

Our T-sentences do have subtle conceptual warrant. The notion of subtle conceptual warrant aims to capture the special sense in which these sentences are conceptually underwritten. It grounds our feeling that example T-sentences are 'iron clad.' Let us say of a sentence of a language M that it has *full conceptual warrant* in M just in case it has either simple or subtle conceptual warrant in M.

Exactly specified languages

Tarski did not think that there was a well-defined class of T-sentences for natural languages like English. He thought that it was not clearly determined what was the basic vocabulary of English. Surely this is correct, since it is vague whether a new term in use, for example 'za,' should be thought of now as a term of English. Tarski also thought indexicality and tense presented difficulties for working directly with a natural language. These days, these are not seen as serious obstacles, because of the work of Donald Davidson (1967: 34) and others.

In the face of these obstacles, Tarski chose to pursue the T-strategy with regard only to languages with what he called an *exactly specified structure*. This would ensure that a language under examination had a well-defined primitive vocabulary and grammar, and this would help ensure that there could be a well-defined set of T-sentences for that language. Simplifying Tarski somewhat, let us say that to *exactly specify* a language, one must specify: a basic vocabulary, grammatical formation rules, the class of sentences, a set of axioms, and inference rules. If a language is exactly specified in purely syntactic terms, then it is said to be *formalized*.

The last two items on the specification list may seem objectionable. We do not think of languages as coming equipped with inference rules and axioms. The worry subsides, however, once we see that the inference rules and axioms in question are indeed determined (albeit not uniquely) by a meaningful language. First, logical relations between meaningful sentences obtain in virtue of what those sentences mean, and a set of inference rules is a way of codifying logical relations. One might also think of the set of inference rules as a way of identifying and specifying the meanings of the logical terms of the language. Either way, what we represent by including inference rules in a specification is determined by the language itself, not super-added. Second, it is a constraint on the axioms Tarski has in mind that they axiomatize the *conceptually assertible sentences* of the language – and these are, I propose, just the sentences of the language with full conceptual warrant. Thus, the axioms of an exact specification are also clearly determined by the language.

Using exactly specified languages makes Tarski's technical project more sure-footed, but makes our philosophical job harder. Since we have good reason to believe that there is no well-defined class of meaningful sentences for a natural language like English, there can be no exact specification of such a language. For this reason, care and reflection is necessary in considering any results we may obtain.

One further notion which we will make use of in the sequel is that of an *empirically assertible sentence*. Conceptually assertible sentences are ones which have exceptional credentials in virtue of which, special knowledge to the contrary, they may be 'treated as true' for the purposes of scientific and logical work. In an empirical language (e.g. a language suitable for expressing physical theory as opposed to the language of arithmetic) some sentences may be treated as true for the purposes of scientific theorizing not in virtue of their conceptual standing, but in virtue of being empirically confirmed. Keeping things as simple as possible, we will say that an *empirically assertible sentence* of a language is one which has met a certain (unspecified) standard of confirmation, and an *assertible sentence* of a language is one which is either conceptually or empirically assertible.

2 The Liar

Suppose, then, that we restrict further attention to languages with an exactly specified structure. The next obstacle to pursuing the T-strategy is more grievous. Considerations based on the Liar Paradox suggest that the T-strategy will lead us into inconsistency.

The Liar Argument

For the sake of argument, let us assume that the definite description 'the sentence with feature f' uniquely denotes the sentence which is *quoted in* sentence (a) below. Our argument will be given in (a fragment of) English. Also, for simplicity, we will suppose that 'L' refers to a language which looks and is structured just like a fragment of English and has no false cognates (so translation into English is transparent). Consider the following Liar Argument which begins with a T-sentence.

(a) 'The sentence with feature f is not true in L' is true in L iff the sentence with feature f is not true in L.
(b) 'The sentence with feature f is not true in L' is identical to the sentence with feature f.
(c) So, 'The sentence with feature f is not true in L' is true in L iff 'The sentence with feature f is not true in L' is not true in L.

It is worth stating carefully how this argument (sequence of sentences) poses a threat to reason. First, suppose you think that (a) and (b) represent beliefs that you hold. Then, certainly, (c) represents something that could be validly inferred from things you believe. But (c) is logically self-contradictory, and this suggests that your beliefs are in a sorry state indeed. You would be rationally compelled to conclude that you had a false belief. It is hard to see how (b) could be the culprit, so suspicion falls on (a). However, (a) could not represent a false belief you had, because we can prove (a) is not false:

> After all, a claim [like (a) which is of the form] ⌜A iff B⌝ can be false only if (i) A is true and B is false or (ii) A is false and B is true. Where A is ⌜'S' is true⌝ and B is S, these combinations cannot occur, for (i) if S is false, then the claim that it is true cannot be true and (ii) if S is true, then the claim that it is true cannot be false. (Soames 1999: 51)

The Inconsistency Argument

Thus, the Liar Argument presents us with an intolerable situation – a genuine affront to reason. The Tarskian analysis of this situation is based on the following Inconsistency Argument. Let M be a fragment of English sufficient for giving the Liar Argument.

(1) Sentence (a) is a conceptually assertible sentence of M. (Premise)
(2) Sentence (b) is an empirically assertible sentence of M. (Premise)
(3) The ordinary rules of logic apply in M (i.e. the rules of inference of M underwrite the usual deductive moves). (Premise)

(4) Thus, the deductively inconsistent sentence, (c), is derivable from (a) and (b) by the rules of inference of M.

(5) It follows that *the language M is inconsistent* in the sense that a deductively inconsistent sentence is derivable by the rules of inference of M from the assertible sentences of M.

This argument is not a problematical argument and its premises are ones that we have no reason at all to reject. Moreover, there *are* exactly specifiable languages for which these premises evidently hold, such as that fragment of English used in giving the Liar Argument earlier. For this reason Tarski held that an exactly specified language as much like English as possible would be inconsistent – a claim that has been a source of consternation and a subject of misinterpretation, for example Soames (1999).

Incoherence of the concept

Examination of the Inconsistency Argument reveals that one of the sentences, (a) or (b), must be *assertible but not true in M*. Again, suspicion falls only on (a). A simple argument showed that (a) cannot be false, so it is immediate that (a) must lack a truth value (and so on some understandings of belief you would surely have been mistaken to think it represented any belief you held).

Now, (a) is a T-sentence, and it is assertible because it has subtle conceptual warrant. Since it is certainly possible for an agent to satisfy the antecedent epistemological conditions for subtle conceptual warrant with respect to (a), we know by this that someone could be in a position to know of (a) that

> if each predicate of (a) is subserved in M by the concept it expresses in M, and each singular referring term of (a) refers in M, then (a) is a true sentence of M.

We know that (a) is *not* a true sentence of M and it is evident that there is no reason to think that any singular term of (a) fails of reference. From these we infer that some predicate of (a) is not subserved in M by the concept it expresses in M. The only candidate is the predicative expression, 'is a true sentence of.' Thus, we are led to conclude that *'is a true sentence of' is not subserved in M by the concept, c, that it expresses in M, i.e. the concept of sentential truth*. How could it possibly happen that we have made a predicate express some concept, and yet, in spite of our intentions, that concept does not subserve it? The only conceivable way this could happen is if it were *strictly impossible* for the concept to subserve it. Such an impossibility is guaranteed if

> the application rules for c imply that the pair ⟨'The sentence with feature f is not true in L', L⟩ *falls under* c, *and* the application rules for c imply that ⟨'The sentence with feature f is not true in L', L⟩ *fails to fall under* c.

The concept of sentential truth is, in a word, incoherent.

3 Tarskian Truth Definition

Evidently the T-strategy invites inconsistency since it is a T-sentence that sets up the Liar Argument. Nonetheless, Tarski has the idea that the strategy still might be usefully carried out by further restricting attention to exactly specified languages for which not all the assumptions of the Inconsistency Argument hold. Specifically, Tarski proposed that we can do this if we only consider object languages, L, which are *not* semantically closed. Where a *semantically closed* language, L, is characterized loosely as one which (1) has the resources to denote its own expressions, and (2) has the resources to predicate truth in L of those expressions. The crux of the matter is that a semantically closed language is one in which a liar sentence (like 'the sentence with feature f is not true in L') can be formed, and this is a sort of thing we are now aiming to avoid. A complete set of T-sentences *for* such a language must include a T-sentence for that liar sentence, and thus, any language, M, in which we could pursue the T-strategy would be one in which the T-sentence for the liar sentence was an assertible one, that is premise (1) of the Inconsistency Argument would hold. So long as we stick with languages that are not semantically closed, however, we effectively avoid this.

I am simplifying. The existence of sentences that form *Liar chains*, means that there are variants on the Liar and Inconsistency Arguments which will make the task of identifying the languages suited for Tarski's definitional project trickier yet (cf. Kripke 1975: 54–5; Yablo 1993).

Truth definitions

We have now (let us suppose) identified a class of exactly specifiable languages for which we might still hope to carry out the T-strategy. As stated earlier, Tarski's insight was that the problem of generality could be solved by employing the (now very familiar) technique of recursive definition. Tarski proceeds by example, showing how to give a recursive definition meeting Convention T for the language of the calculus of classes (a quantified language ranging over sets and having a single predicate term expressing the subset relation). Note, the language in which the definition is expressed is, perforce, *expressively richer* than the object language, since the former has sentences that translate all those of the object language, *as well as* the resources to denote the expressions of the object language.

To give an example definition here, we will use a language, L, which has a two-place predicate, '$\subseteq$', for the subset relation, plus logical terms for negation, conjunction, and quantification, and some individual variables.

DEFINITION Let an *L-sequence*, f, be a function from the variables of L into the domain of discourse of L.

DEFINITION For a variable, α, of L, let an *α-variant* of an L-sequence f be any L-sequence, f′, which is just like f except possibly for the value f′ assigns to α.

DEFINITION For all L-sequences, f, and every formula, σ, of L, f *L-satisfies* σ iff

if σ is of the form $\ulcorner \alpha \subseteq \beta \urcorner$ for some variables α and β, then f(α) is a subset of f(β),

if σ is of the form $\ulcorner \psi \,\&\, \theta \urcorner$ for some formulas ψ and θ, then f L-satisfies ψ and f L-satisfies θ,

if σ is of the form $\ulcorner \neg \psi \urcorner$ for some formula ψ, then f does not L-satisfy ψ, and

if σ is of the form $\ulcorner \forall \alpha\, \psi \urcorner$ for some variable α and formula ψ, then every α-variant of f L-satisfies ψ.

DEFINITION For all sentences φ of L, φ *is a true sentence of* L iff φ is L-satisfied by every L-sequence.

This is not the definition we are looking for, but it is a simple matter to transform our recursive definition of L-satisfaction into an explicit definition and combine it with our truth definition to yield the following explicit truth definition.

DEFINITION (explicit): For all sentences φ of L, φ *is a true sentence of* L iff for all L-sequences, g, $\langle g, \varphi \rangle$ is a member of the least set, X, such that for all L-sequences, f, and L-formulas, σ,

if σ is of the form $\ulcorner \alpha \subseteq \beta \urcorner$ for some variables α and β, then $\langle f, \sigma \rangle \in X$ iff $f(\alpha)$ is a subset of $f(\beta)$,

if σ is of the form $\ulcorner \psi \,\&\, \theta \urcorner$ for some formulas ψ and θ, then $\langle f, \sigma \rangle \in X$ iff $\langle f, \psi \rangle \in X$ and $\langle f, \theta \rangle \in X$,

if σ is of the form $\ulcorner \neg \psi \urcorner$ for some formula ψ, then $\langle f, \sigma \rangle \in X$ iff $\langle f, \psi \rangle \notin X$, and

if σ is of the form $\ulcorner \forall \alpha\, \psi \urcorner$, then $\langle f, \sigma \rangle \in X$ iff for every α-variant, f', of f, $\langle f', \psi \rangle \in X$.

It is widely accepted that these sorts of definitions do satisfy Convention T, and so we surely have a definition suitable for a predicate which expresses the concept of *a true sentence of L*.

Translingual truth predicates

Tarski succeeded in following the T-strategy to its completion, solving the problem of generality by showing how to give a recursive definition, and this rises to the technical challenge we identified at the outset. However, the sample definition Tarski gave is apt only for a predicate expressing the monolingual concept of *a true sentence of the language of the calculus of classes.* What about the relational concept of sentential truth? It has often been said in criticism of Tarski that he showed "how to define 'is a true sentence of L' for fixed L," but failed to show us how to define the relational 'is a true sentence of' – the implication being that it is only the latter that expresses a concept in which philosophers are *really* interested.

Yet, Tarski himself evidently thought that his technique could be generalized. Indeed, there would seem to be no barrier, in principle, to the construction of a definition suitable for a two-place predicate, like 'is a true sentence of' *provided that* the language, M, in which the definition is to be given not be one for which the premises of the Inconsistency Argument hold. This indirectly imposes a constraint on the object lan-

guages over which the target truth predicate can range. There may be real difficulties posed if the range of object languages is infinite, but such difficulties do not show that there are no suitable metalanguages, M.

The complaint we are considering is really misplaced. What is true is that the language in which a relational truth predicate were defined in the Tarskian way would have to be a language with a restricted domain of discourse. However, this is no more than the demand for consistency requires.

4 Discussion

The question of analysis

Does a Tarskian definition like the one just described provide a conceptual analysis of the concept of sentential truth? Certainly not. One of Tarski's main aims in giving a definition is to ensure consistency with empirical facts. But the upshot of the Inconsistency Argument is that the concept of sentential truth is incoherent in such a way that anything that might pass for an *analytical* definition would surely not be consistent in this way. So, it is very plainly not on the Tarskian agenda to provide an analytical definition. In fact, Tarski understands himself to be *defining the set* which is the extension, not a predicate at all, so the definitional part of his project is not about giving meanings (Coffa 1991: 293–6).

Still, it may come as a surprise that even a Tarskian definition for a semantically open language like Quadling could not at least give us an analysis of the humble concept of *a true sentence of Quadling*. Nonetheless, the denial of such analytic status is implied by a family of arguments promulgated in the literature. If successful, these arguments would show that if you *introduced a new predicate* using a Tarskian definition, this predicate would not mean the same thing as an antecedently meaningful truth predicate that expressed one of our truth concepts. These arguments proceed by comparing an example T-sentence to the result of performing definitional substitution on that T-sentence using a Tarskian truth definition. The arguments seek to impugn the Tarskian definition by finding telltale differences between the two sentences that point to differences in meaning. Philosophers have claimed differences in logical status, modal status, subject matter, and informativeness. Arguments of this sort can be found in Putnam (1985: 63–4); Soames (1995: 253–4). John Etchmendy (1988: 56–7) seeks to use this sort of argument to draw a further conclusion, namely that Tarskian definitions do not give any information about the semantics of their target language, appearances notwithstanding. If this were correct, Tarski (1936) would not have contributed to theoretical semantics in the way he is widely thought to have. For critical discussion of these arguments, see Davidson (1990: 288–95); Garcia-Carpintero (1996); Heck (1997).

Even if these arguments are correct, it would be a mistake to conclude as some do that Tarskian definitions are merely extensionally correct or that Tarski thought only as much. The appropriateness of the *T-strategy* depends on there being a significant conceptual connection between the concept in question and the T-sentences, but Tarski's

notion of *definition* does not itself require 'giving the concept.' Yet, since Tarskian definitions realize the T-strategy, they inherit something of that conceptual connection. *If one satisfied the knowledge conditions set forth in the definition of subtle conceptual warrant, and one knew that the language M was so chosen that the truth predicate would be subserved by the concept it expresses, then one would be in a position to know that each of these T-sentences is true in M.* Ditto for the Tarskian truth definition statement. For this reason, we might with some justice say that a Tarskian truth definition for L 'captures' the concept of *a true sentence of L*, even though the definition is not a concept-giving one.

Deflationism

The Tarskian view is often associated with deflationist views of truth. According to some, the central tenet of *deflationism* is that there is no property of truth. If this is what is meant by deflationism, then the Tarskian view of truth that we have outlined is most certainly deflationist. In fact, Tarski was in possession of the best possible reason for endorsing this deflationist thesis. Tarski's view of truth delivers a simple argument to the conclusion that there is no *property (relation) of sentential truth*. The argument I have in mind relies on two general principles:

1. For any binary relation, R, necessarily, for any pair, x and y, either x is R-related to y or x fails to be R-related to y.
2. For any R,p,M, if R is the relation of sentential truth, and p is a predicate which expresses in M the concept of sentential truth, then for every x and y in the domain of discourse of M, p applies in M to $\langle x,y \rangle$ iff x is R-related to y.

Item (1) states a conceptual truth about properties; that is just the sort of thing that a property was supposed to be. (2) articulates how the concept of sentential truth and the property thereof would be related to a predicate that expressed that concept. Simply put, a predicate is supposed to be 'underwritten' by the property (if any) associated with the concept the predicate expresses.

Now, we reason as follows. By way of contradiction, suppose there is a property of sentential truth, R. Then it is easy to see that (1) and (2) enable us to infer that any predicate expressing the concept of sentential truth will have a proper extension, but we know this is not so. Another way of seeing the point is to see that our supposition, together with (1) and (2) ensure that premise (a) of the Liar Argument is true in M, and we know that cannot be so. Therefore, by reductio, there is no property of sentential truth. This reasoning can be repeated for monolingual truth properties such as *being a true sentence of French*.

Thus, it is not hard to muster a deflationist conclusion from the Tarskian view as we have developed it here. However, sometimes 'deflationism' is associated with the rather more nebulous idea that truth is not a philosophically significant notion. Nothing we have said suggests that one who held Tarski's view should be a deflationist in *this* sense. Indeed, it looks like the view may be committed in quite the opposite way. Tarskian definitions don't give an analysis of our concept of sentential truth nor even of its sub-

sidiary monolingual notions. This *suggests* that there may be something to these concepts other than what the T-sentences codify. Working out a cogent *disquotationalist* version of deflationism could be seen as the attempt to resist this suggestion. Space prohibits further discussion here, but a *locus classicus* of the debate is Field (1987), and a useful collection in this area is Blackburn and Simmons (1999). For an extended discussion of the disquotationalist proposal, see David (1994).

Making truth safe for science

Even though his analysis of the Liar pointed to the incoherence of the concept of sentential truth, Tarski nonetheless saw value in carrying out his definitional project in a restricted context. By doing so, he showed that one *can* appeal to the notion of sentential truth in the conduct of inquiry without fear of introducing inconsistency. In so far as the restricted class of languages in which this applies is broad enough for the conduct of scientific inquiry (is it?).

Tarski succeeds in *making truth safe for science.* Note that success in this does not require that Tarskian definitions be analytical, nor, *contra* Field (1972), that they offer any sort of physicalistic reduction of semantics.

5 Conclusion

We have presupposed throughout our discussion that there *is* a concept of sentential truth and that, by dint of certain linguistic intentions, terms like 'is a true sentence of' express it. But we should now step back and examine this presupposition. To tell the story we have, we tacitly employed a conception of concept according to which, by making the term 'true' express a certain concept, we would make it the case that the word 'true' is *supposed to work a certain way,* it is *supposed to* apply to certain things and it is *supposed to* fail to apply to certain other things, *whether the term for whatever reason actually succeeds in applying (failing to apply) to those things or not.* We appealed earlier to the fiction that concepts were things that came equipped with explicit application rules to make this idea concrete. On this conception of concept, there is nothing funny about speaking, as we have, of an incoherent concept.

However, philosophers have commitments about these things and there is a history to the use of the word 'concept.' Some will hold that concepts are more akin to properties as I have characterized them. That is to say, some will hold that a concept is a thing that partitions the things of the universe into two classes – those that fall under the concept and those excluded by it – and if if there isn't such a partition, you don't have a concept. Earlier reasoning showed, however, that there can be no such partition when it comes to sentential truth. Thus, if we use 'concept' in this narrower sense, the Tarskian view will certainly force us to say that there is not even a concept of sentential truth. Instead, we should have to say, there is only a kind of predicate which, by dint of our linguistic intentions, is *supposed* to work this way and is *supposed* to work that way, when in fact *nothing could possibly work* this way *and* that way.

With such dramatic philosophical conclusions as these in the offing, it is easy to see how Tarski's work could become a centerpiece of the discourse on truth.

References

Coffa, J. A. (1991) *Semantic Tradition from Kant to Carnap: to the Vienna station*. Cambridge: Cambridge University Press.

David, M. (1994) *Correspondence and Disquotation*. New York: Oxford University Press.

Davidson, D. (1967) Truth and meaning. Reprinted in Davidson, D. (1984) *Inquiries into Truth and Interpretation* (pp. 17–36). Oxford: Clarendon Press.

Davidson, D. (1990) The structure and content of truth. *Journal of Philosophy*, 87, 279–328.

Etchemendy, J. (1988) Tarski on truth and logical consequence. *Journal of Symbolic Logic*, 53, 51–79.

Field, H. (1972) Tarski's theory of truth. *Journal of Philosophy*, 69, 347–75.

Field, H. (1987) The deflationary conception of truth. In G. MacDonald and C. Wright (eds.), *Fact, Science and Morality* (pp. 55–117). Oxford: Blackwell Publishers.

Garcia-Carpintero, M. (1996) What is a Tarskian definition of truth? *Philosophical Studies*, 82, 113–44.

Heck, R. Jr. (1997) Tarski, truth, and semantics. *Philosophical Review*, 106, 533–54.

Kripke, S. (1975) Outline of a theory of truth. Reprinted in R. L. Martin (ed). (1984), *Recent Essays on Truth and the Liar Paradox* (pp. 53–81). Oxford: Oxford University Press.

Putnam, H. (1985) A comparison of something with something else. *New Literary History*, 17, 61–79.

Ramsey, F. P. (1929) On truth. Published posthumously in N. Rescher and U. Majer (eds.) (1991), *Episteme* (vol. 16). Dordrecht: Kluwer Academic Publishers.

Soames, S. (1995) T-sentences. In W. Sinnott-Armstrong, D. Raffman and N. Asher (eds.), *Modality, Morality, and Belief: Essays in Honor of Ruth Barcan Marcus* (pp. 250–70). Cambridge: Cambridge University Press.

Soames, S. (1999) *Understanding Truth*. New York: Oxford University Press.

Tarski, A. (1936) The establishment of scientific semantics. Reprinted in translation in J. Corcoran (ed.) (1983), *Logic, Semantics, Metamathematics*. 2nd edn. (pp. 401–8). Indianapolis: Hackett.

Yablo, S. (1993) Paradox without self-reference. *Analysis*, 53, 251–2.

Further Reading

Blackburn, S. and Simmons, K. (eds.) (1999) *Truth*. New York: Oxford University Press.

Chihara, C. (1979) Semantic paradoxes: a diagnostic investigation. *Philosophical Review*, 88, 590–618.

Horwich, P. (ed.). (1994) *Theories of Truth*. Aldershot: Dartmouth.

Kirkham, R. L. (1992) *Theories of Truth: A Critical Introduction*. Cambridge, MA: MIT Press.

Tarski, A (1933) On the concept of truth in formalized languages. Reprinted in translation in J. Corcoran (ed.) (1983), *Logic, Semantics, Metamathematics*. 2nd edn (pp. 152–278). Indianapolis: Hackett.

Tarski, A. (1944) The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, 4, 341–75.

12

Descriptions and Logical Form

GARY OSTERTAG

According to a tradition associated with Frege, Russell, and Tarski, logical form is that aspect of sentence structure relevant to inference, semantical evaluation, and ontological commitment. More recently, Chomskian linguistics has given birth to an alternative conception, taking Logical Form (henceforth, 'LF') to refer to the level of grammatical description at which scope precedence among quantifiers and related expressions is made explicit. Although syntactically motivated, LF is an attractive and powerful medium for representing logical form in the sense associated with Frege, Russell, and Tarski. The present chapter considers in detail Russell's proposal concerning the logical form (in the traditional sense) of sentences exemplifying the surface grammar *the F is G* – namely, the Theory of Descriptions – and asks to what extent it can be accommodated in a picture of logical form inspired by LF.

Due to considerations of space, an alternative approach to the logical form of description sentences is not given emphasis equal to Russell's, although I believe it to be viable and worthy of consideration. This is the approach that takes descriptions to be referential singular terms. My reasons for preferring Russell's theory are given in the subsection of Section 1, "Descriptions as singular terms". In addition, the challenge to Russell posed by the referential use of definite descriptions is not addressed in what follows. I am assuming – what is perhaps not obvious – that the Russellian theory can accommodate such usage.

1 Preliminaries

Formal aspects of Russell's theory of descriptions

Russell's rendering of a sentence exemplifying the surface form *the F is G* is as follows:

(1) $G(\iota x)Fx$

(1) corresponds more precisely to the quasi-formal English sentence: the x such that x is F *is* G. (1) is defined in terms of (2):

(2) $\exists x(\forall y(Fy \equiv y = x) \wedge Gx)$

That is to say: something is both G and uniquely F. It will be useful to keep the English language paraphrase in mind, since much in subsequent sections turns on it. Note that, in (1), the variable 'x' as it occurs in 'Fx' is bound by the iota operator, '(ιx).' In its capacity to bind a free variable, the iota operator is similar to the first order quantifiers '$\exists x$' and '$\forall x$.' Yet, whereas these operators, appended to 'Fx,' will produce a closed formula, or sentence, the iota operator will produce an expression that functions, *syntactically*, as a term. As we shall see, the qualification is important.

It will be noted that, as an analysis of (1), (2) is not quite right as it stands, since it fails to tell us what to do when a description is embedded in a complex context – for example (3):

(3) $G(\iota x)Fx \supset p$.

This sentence has the following two readings:

(3a) $\exists x(\forall y(Fy \equiv y = x) \wedge (Gx \supset p))$
(3b) $\exists x(\forall y(Fy \equiv y = x) \wedge Gx) \supset p$

(3a) claims that something is both G-only-if-p and uniquely F, whereas (3b) claims that something is both G and uniquely F, only if p. That these readings are truth-conditionally distinct is easy to see: if nothing is F, then (3a) will be false, whereas (3b) will be true. Intuitively, the respective readings correspond to what we take the scope of '$(\iota x)Fx$' to be in (3). If the scope of '$(\iota x)Fx$' is the entire sentence (3), then (3a) is the correct reading; if the scope of '$(\iota x)Fx$' is restricted to '$G(\iota x)Fx$,' then (3b) is the correct reading.

Russell introduced a somewhat cumbersome notational device –'$(\iota x)Fx$' enclosed in square brackets – to mark off the scope of '$(\iota x)Fx$.' Assuming Σ is a context in which '$G(\iota x)Fx$' can occur as a subformula, Russell's idea is as follows: If the scope indicator is prefixed *directly* to the formula in which the contained description occurs as an immediate constituent (here, '$G(\ )$'), then the description takes narrowest scope. This situation is represented thus: $\Sigma([(\iota x)Fx]G(\iota x)Fx)$. On the other hand, if the scope indicator is prefixed to Σ, then the description has wide scope with respect to Σ (equivalently, the scope of the description is said to be Σ itself). The latter situation is represented in the following manner: $[(\iota x)Fx]\ \Sigma(G(\iota x)Fx)$. For example, in (3c), the description takes narrow scope, indicated by the fact that '$[(\iota x)Fx]$' is affixed directly to the simplest subformula containing '$(\iota x)Fx$.' In (3d) the description takes wide scope, as it is prefixed to the entire formula:

(3c) $[(\iota x)Fx]\ G(\iota x)Fx \supset p$
(3d) $[(\iota x)Fx]\ (G(\iota x)Fx \supset p)$

We are now ready for the official definition of '$(\iota x)\Phi x$' as stated in *Principia Mathematica*:

(*14.01) $[(\iota x)\Phi x]\ \Psi(\iota x)\Phi x =_{df} \exists x(\forall y(\Phi y \equiv y = x) \wedge \Psi x)$

(*14.01) does not define '$(\iota x)\Phi x$' directly, but rather provides a procedure for eliminating it from any context in which it may occur. That is to say, it provides a *contextual definition* of '$(\iota x)\Phi x$.' While Russell's views on definition are complex and cannot be adequately treated here, it is important to bear in mind his claim that "a definition is concerned wholly with symbols and not with what they symbolize"; they are, "strictly speaking, typographical conveniences" (Whitehead and Russell 1925–7: 11). This strongly suggests a reading of the definitions as merely providing abbreviations for complex formulae. Yet, Russell does note that although a definition in his sense is always "theoretically . . . superfluous," it does retain a certain pragmatic significance, especially in those cases where the definiendum (the expression being defined) is familiar. In such cases the "definition contains an analysis of a common idea, and may therefore express a notable advance" (Whitehead and Russell 1929: 13). Applied to definite descriptions, this is significant: it suggests that we notice a parallel between the definiendum and something already familiar to us – definite descriptions as they occur in, say, English. To choose to abbreviate '$\exists x(\forall y(Fy \equiv y = x) \wedge Gx)$' with a formula that has the argument structure of a simple predication (e.g. 'Gt') may be theoretically arbitrary – in principle, another abbreviatory convention would have served as well – but it makes a point about the logical form of English sentences exemplifying *the F is G* that the alternative would not have: namely, that they can be eliminated in a similar manner.

This point is worth emphasizing, since it is often assumed that Russell's *Principia* theory of descriptions has no bearing on the interpretation of definite descriptions in English and other natural languages, and, indeed, that this was Russell's official position. Writing in response to Moore, Russell remarked that "the whole of my theory of descriptions is contained in the beginning of 14 of *Principia Mathematica*," adding that "the reason for using an artificial language was the inevitable vagueness and ambiguity of any language used for every-day purposes" (Russell 1944: 890). While some (e.g. Mates 1973) have seen in this remark a concession that his theory is really *only* a formal definition, this is in fact not the case: "The two definitions which embody the theory of descriptions (*14.01.02), though formally they are merely nominal definitions, in fact embody new knowledge; but for this, they would not be worth writing about" (Russell 1944: 891). This suggests quite clearly that Russell intends his definitions to provide an analysis of definite descriptions. If he concedes anything, it is that his definitions do not provide a general theory of *the definite article*, since, as Moore pointed out, the definite article is used in ways that the theory cannot accommodate (notoriously, the generic use of *the*).

Descriptions as singular terms

An alternative paradigm to Russell's maintains that definite descriptions are semantical singular terms. On this approach, surface grammar does not mislead with respect to logical form: '$(x)Fx$' is both grammatical subject *and* logical subject of '$G(\iota x)Fx$.' Initially suggested by Frege, the singular term proposal is defended in Strawson (1950), which criticizes Russell for identifying the presuppositions characteristic of description sentences with the actual content of such sentences. As Strawson writes: "To use the word 'the' in [the uniquely referring] way is . . . to imply (in the relevant sense of

'imply') that the existential conditions described by Russell are fulfilled. But to use 'the' in this way is not to *state* that those conditions are fulfilled" (Strawson 1950: 147).

Consider, in this connection, (4):

(4) The present king of France is bald.

For Strawson, an utterance of (4) presupposes, and does not assert, that there exists a unique French monarch. To see this, he has us imagine the following scenario. Someone asks you, apparently seriously, whether the present king of France is bald. If Russell is correct, it is appropriate to respond, "no, that's false." Clearly, however, a more appropriate response would be to address the speaker's beliefs, and not what she said. One might respond, for example, by saying: "You seem to be laboring under a false belief – France is not a monarchy." This suggests that the propriety of the description – its having a unique denotation – is not an aspect of the content of description sentences, but of the rules dictating their correct use.

The notion of presupposition Strawson appeals to can be defined as follows: An utterance u presupposes p just in case: u is true or u is false only if p is true. It follows that sentences containing vacuous descriptions have no truth-value (since the relevant uniqueness and existence propositions are false). This does appear to be the case with post-Revolutionary utterances of (4). But if there are intuitions that favor Strawson, there are also intuitions that favor Russell. Consider (5a/b):

(5a) If Ferdinand is not drowned, Ferdinand is my only son. (Russell)
(5b) Yesterday, I dined with the King of France. (Neale)

Russell remarks that the King in *The Tempest* might have uttered (5a), and suggests that it would be true even if Ferdinand – the King's only son – had, in fact, been drowned. In addition, while Strawson would hold that an utterance of (5b) presupposes that France has a single monarch, and thus should be without a truth-value, one is hard-pressed to hear it as anything but false. So, it seems that appeals to usage provide little guidance in determining whether or not the relevant uniqueness and existence propositions are part of what is said by a description sentence. Without decisive evidence in favor of the presupposition doctrine, the Strawsonian challenge to Russell is inconclusive.

In addition, the singular-term interpretation of descriptions fails to provide a satisfying account of the propositional content of description sentences. A naïve application of this interpretation would identify the proposition expressed by an utterance u of 'the current US president is a man' with the ordered pair ⟨Clinton, is a Democrat⟩. This would be to associate the wrong proposition with the utterance. What u said should be true at any circumstance of evaluation E at which the US president (at E) is a democrat. However, a naïve application of singular-term approach will entail that what I said in uttering u is true at E just in case Clinton exists at E and is a man at E – even if the US president at E is woman! Thus, the naïve singular-term approach must be rejected.

The obvious solution is to relativize the reference relation to context. On this view, u expresses ⟨f, is a man⟩, where f is a partial function from a circumstance of evaluation E to the unique president, if there is one, at E. (Note that f corresponds to one inter-

pretation of Frege's notion of sense.) The proposition ⟨*f*, is a man⟩ is true at E just in case the US president (at E) is a man (at E). This assigns the correct proposition to *u*, at least, if we restrict our attention to those contexts at which there is a unique US president. However, this proposal, while truth is conditionally adequate, attributes a dimension to referential singular terms that is not independently motivated. As Evans writes:

> Simply in order to assimilate descriptions to referring expressions, we introduce a major change in the semantic apparatus in terms of which we describe the functioning of referring expressions in general. As a consequence of this change, we ascribe to names, pronouns, and demonstratives semantical properties of a *type* which would allow them to get up to tricks they never in fact get up to; since their reference never varies from world to world, this semantic power is never exploited. (Evans 1982: 56)

Adding an extra parameter to the reference relation is necessary if we are to provide an adequate analysis of description sentences consistent with the thesis that descriptions are referring expressions. Yet, it arbitrarily weakens the reference relation. While in no way a decisive refutation of the referential interpretation of descriptions, this consideration raises a genuine concern that the classification resulting from the analysis "may not correspond to any natural semantical kind" (Evans 1982: 57). In sum: the Strawsonian view can assimilate descriptions to the class of referential singular terms only by characterizing that class in an manner that appears dangerously *ad hoc*.

While these considerations do not settle the question against the singular term approach, the Russellian analysis seems at this point more promising.

2 Descriptions and Quantification

Restricted quantification

In Chomskian linguistic theory, 'Logical Form' (or LF) refers to a level of syntactic representation at which the scope properties of quantified noun phrases are made explicit. This level is posited to account for the distinct readings that can be assigned sentences such as 'Everyone thanked someone.' A sentence that is *n* ways structurally ambiguous at surface structure – the level that is realized phonologically – is assigned *n* distinct representations at LF. Although motivated by purely syntactic concerns, LF can function as input to a semantic theory – a theory that assigns propositions or truth conditions to sentences of English – more effectively than the language of *Principia*. This is due to the fact that, at LF, natural language quantifiers are represented as restricted quantifiers. In particular, they are represented as variable-binding devices constructed by pairing a determiner (itself represented as a variable-binding device) with a predicate and enclosing the result in brackets. Using this notation, we can represent (1) as follows:

[the *x*: F*x*] (G*x*)

This allows for a perspicuous representation of scope. For example, the readings captured by (3a/b) correspond to (3a′/b′):

(3a′) [the x: Fx] (Gx ⊃ p)
(3b′) [the x: Fx] (Gx) ⊃ p

Note that the current proposal does not entail that descriptions, any more than other quantifier phrases, can occur as logical subjects. Indeed, it shows why certain apparent grammatical subjects – that is expressions, like quantifiers, that are subjects at surface structure – are not, ultimately grammatical subjects at all. The proposal maintains that it is characteristic of a quantifier that it can be 'raised' from the position in which it occurs in *S* to a position preceding *S*, leaving behind a 'trace' which it binds from its new location. To illustrate: the quantifier in 'John admired [some vases]$_1$' can be relocated to the left of the original sentence: '[some vases]$_1$ John admired t_1.' The quantifier and trace are co-indexed, indicating that the former binds the latter. The principle allowing this movement is referred to as QR (for *Quantifier Raising*). While constraints on QR are an important aspect of the proposal under discussion, we will have to take it as given that the position to which the quantifier is raised binds the 'evacuation site.' (For details, see Heim and Kratzer 1997.) The power of this proposal can be seen in the following application:

(6a) [each curator]$_1$ admired [some vases]$_2$

(6a) represents the surface form of 'Each curator admired some vases.' Intuitively, this sentence is ambiguous between two readings. These readings can be supplied by successive applications of QR:

(6b) [each curator]$_1$ t_1 admired [some vases]$_2$
(6c) [some vases]$_2$ [each curator]$_1$ t_1 admired t_2

Alternatively, '[some vases]$_2$' in (6b) can be raised to the position immediately preceding the sentence containing its trace, yielding the second reading:

(6c) [each curator]$_1$ [some vases]$_2$ t_1 admired t_2

Note that the movement characterized by QR is possible only for quantifiers (including *wh*-phrases): names and other singular terms cannot be raised. Had we adopted the Frege–Strawson approach, QR would not be applicable to descriptions.

The notation used above translates quite straightforwardly into the restricted quantifier notation (with traces indicating argument positions in open sentences), suggesting that the syntactic level at which disambiguation occurs is closely tied to, if not identical with, the level at which semantic interpretation occurs.

Russell has been criticized for presenting apparently divergent pictures of the logical form of description sentences. On the one hand, (*14.01) suggests that description sentences are really only typographical abbreviations of more complex formulae. Yet, it becomes virtually irresistible to read (*14.01) as providing the *truth conditions* for sentences containing the *iota* operator. On the latter reading, the logical form of '$G(\iota x)Fx$' is that of a singular sentence, albeit one whose truth conditions are given quantificationally. We have seen how this latter view is mistaken – for Russell, descriptions are

not semantical singular terms. Neale suggests that an advantage of the restricted quantifier interpretation eliminates any residual uncertainty surrounding (*14.01) (Neale 1995: 779–80). His point is that the latter interpretation reveals a feature of definite descriptions that (*14.01) obscured. Indeed, if we take the notation at face value, then it appears that descriptions really are natural language quantifiers. Small wonder, then, that when treated as singular terms in the syntax of *Principia Mathematica* they don't have meaning in isolation: they can't be assigned meanings in this manner precisely because, being quantifiers and not terms, they are not the kinds of expressions that are assigned referents (and, at least for Russell, the only meaning-candidate that a description, being a term, could have if it were, *per impossible*, to have meaning in isolation, would be a referent). It is also notable that Russell's initial presentation of the theory of descriptions in "On Denoting" classifies definite descriptions with other natural language quantifiers (his term was "denoting phrase"), such as "all men," "some men," and "no men." Thus, the restricted quantifier notation has some claim to capturing the essence of Russell's theory.

This formalization will provide us with a new means of expressing (*14.01) – namely [EQ]:

[EQ]: $[\text{the } x\text{: } \Phi x](\Psi x) =_{df} [\text{some } x\text{: } [\text{all } y\text{: } \Phi y](y = x)](\Psi x)$

Indeed, Stephen Neale has claimed that '[the x: Φx] (Ψx)' is "definitionally equivalent" to its Russellian expansion (Neale 1990: 45). Yet, while it appears to place Russell's theory in a new and illuminating light, [EQ] raises certain difficulties of its own. We now turn to a discussion of these difficulties.

The problem of incompleteness

One ubiquitous feature of natural language quantification is incompleteness or underspecification. We often utter sentences such as (7), fully intending to say, and be taken as saying, something true:

(7) Everyone left the party early.

Of course, if we assume that (7) expresses a context-independent proposition, then it would be false (and, what is more, irrelevant) at any actual context, since it would make a claim about *every* existing person. Nonetheless, there is a clear intuition that it can be used to say something true and relevant. Similarly with (8): it can be uttered in many contexts to express a true proposition, even though a naïve application of Russell's theory will assign it a false (and conversationally irrelevant) proposition:

(8) The senator will not seek re-election.

These intuitions are defeasible, of course – it may well be that speakers systematically misidentify what they say, expressing false propositions when in fact they appear to be expressing truths. The latter view – defended by Kent Bach (1988) – is attractive in that it leaves our semantics untouched, requiring no supplementary apparatus to accom-

modate contextual effects. Given the slender intuitive basis for such a view, however, it is best considered only when all the available options have been found wanting; we shall not consider it further here.

A central approach to incompleteness – which, unlike the approach just considered, takes incompleteness to be a semantic phenomenon – is the explicit strategy (Neale 1990). On this approach, an utterance of (8) expresses a proposition that completes the description 'the senator' (e.g. 'The *senior senator of New York* will not seek re-election'). Perhaps its most succinct formulation is to be found in Schiffer (1995), where it is stated in the form of a 'meaning rule' for *the F is G*:

[ES] Utter *the F is G* only if you mean that [the x: Fx $\wedge$ H*x] (Gx),

where H* is a contextually-determined, implicitly referred-to property that completes *the F* (this formulation departs slightly from Schiffer). According to [ES], descriptions have a hidden constituent whether or not they are in fact incomplete. In cases where *the F* is complete, H*x can be any property such that (Fx $\wedge$ H*x) $\equiv$ Fx.

[ES] is an instance of a more general principle, which it will be convenient to state for future reference (*det* is a placeholder for a determiner):

[ES-Q] Utter *det F is G* only if you mean that [*det* x: Fx $\wedge$ H*x] (Gx),

where H* is as before.

While the explicit approach is often interpreted as claiming that, typically, an incomplete description is *elliptical* for a contextually definite completion, the notion of ellipsis involved is left at an intuitive level – it is not to be identified with the formal notion familiar from syntactic theory. According to that notion, a sentence such as 'John loves opera and Mary does too' is elliptical for 'John loves opera and Mary loves opera' in the following sense: the words 'loves opera' are "covertly present" in the former (Stanley and Szabó 2000). That is, a grammatical rule has permitted the deletion of 'loves opera' in the former sentence in such a way that the excised material can be reconstructed from the resulting context. It is a merit of [ES] that it avoids this reading of the explicit approach, since the syntactic reading assumes that the deleted descriptive material to be recovered from the context is lexical. It is, to say the least, implausible that context alone would allow the audience to recover the completion, if the latter is identified with a string of lexical items.

In fact, a problem remains even if we relax the conditions on completion to accord with [ES]. The central obstacle to acceptance of the explicit approach is that, typically, when a speaker utters an incomplete description, there are a number of completing properties that are equally compatible with her intentions, yet no *one* that is exclusively so. Since [ES] implies that a definite description should be uttered only if there is a *particular* completion that the speaker intends, it fails to state the meaning rule that underlies the competent speaker's mastery of the definite article, since the competent speaker often utters description sentences with no particular completion in mind. Consider, for example, (8). For any normal context in which (8) is truthfully uttered, there will be a number of completions of this sentence, each of which are equally obvious ways of

picking out the intended person, but none of which is significantly more obvious than the others. On the other hand, [ES] requires that there is one property H*, such that in uttering (8) in a normal context, the speaker asserts that the H* senator will not seek re-election. As we have seen, there is reason to be skeptical that the speaker can have the 'meaning intentions' in uttering (8) that [ES] requires, since it is unlikely that there will be any distinguished property that the speaker can have intended in exclusion of all others (Schiffer 1995).

One natural modification is to relax the requirement that the completion be unique. The suggestion is that, in uttering (8), the speaker did not intend to convey a particular completion (or, equivalently, a determinate completing proposition), but rather that she "*sort-of-meant*, or vaguely meant" several completions (or completing propositions) (Schiffer 1995: 371). The suggested revision of the meaning rule would run as follows:

[ES*] Utter *the F is G* only if you mean that [the x: $Fx \wedge H^*x$] (Gx),

where, for every candidate property H, H* *indeterminately refers* to H (a *candidate property* being a completion of *the F* compatible with the speaker's intentions). The problem with [ES*] is that it rather arbitrarily links quantifier incompleteness with content indeterminacy. But, in fact, there is no apparent content indeterminacy in a typical utterance of (8). A more promising approach is to retain unmodified [ES] and maintain that, despite appearances, speakers by and large have a determinate completion in mind when uttering sentences such as (8). On this interpretation a speaker need not have conscious access to the implicitly referred-to completing property in order to comply with [ES] (cf. Loar 1976).

Note, however, that even if we assume that one or another implementation of the explicit strategy is correct, another worry arises. I have been suggesting that Russell's theory requires contextual supplementation if it is to provide an adequate account of our implicit capacity to assign the correct truth conditions to utterances containing incomplete descriptions. Indeed, Neale writes that this idea was not foreign to Russell:

> [O]nce the philosophical underpinnings of the Theory of Descriptions are in focus, it is clear that Russell is concerned with the propositions expressed by particular utterances of sentences containing descriptive phrases; he is *not* primarily concerned with the more abstract notion of the linguistic meanings of sentences-*types*. (Neale 1990: 67; emphasis in text)

Yet, if we also follow Neale in accepting that '[the x: Fx](Gx)' is logically equivalent to its expansion we are in for trouble. To see this, consider a revision of [EQ] suggested by Neale's remarks:

[EQ*] An utterance u at C of *the F is G* is equivalent to an utterance u^* at C* of its Russellian expansion *exactly one thing is F and each thing that is F is G* (where C* differs from C only in respect of its containing u^* where C contains u^*).

It should be clear that [EQ*] is stronger than [ES] combined with [EQ] – indeed, for the Russellian, [EQ] should add nothing to [ES], which could be equivalently formulated as follows:

[ES-2] Utter *the F is G* only if you mean that [some x: [all y: $Fy \wedge H^*y$] $(y = x)$] (Gx).

As a number of philosophers point out – for example, Reimer (1992), Larson and Segal (1995) and Stanley and Williamson (1995) – a description sentence S as uttered at a context C cannot be assumed to be equivalent to its Russellian expansion S^* uttered at a relevantly similar context C*. Consider, for example, an utterance u at C of (9a):

(9a) The party was a success.

The Russellian must maintain that *the same proposition*, or, at least, a *proposition with the same truth conditions*, would have been expressed by an utterance of (9b):

(9b) There was exactly one party, and every party was a success.

As suggested, [ES] and [ES-Q] do not provide a guarantee that the respective utterances will be completed in the same way. The successive clauses in (9b) will not necessarily refer implicitly to the completing property referred to in (9a). Nor will they necessarily refer to one and the same property. They may refer, for example, to *party we attended last night* and *party we missed last night*, respectively. If so, (9a) and (9b) will express distinct propositions with potentially distinct truth-values.

In fact, the failure of [EQ*], while fatal to Russell's theory, does not undermine an attractive and closely related approach, one which drops [EQ*], but which accepts [ES] (and thus [EQ]). This approach entails that, for every context C, the proposition expressed by *the F is G* at C – namely, that [the x: $Fx \wedge H^*x$] (Gx) – is equivalent to the proposition that [some x: [all y: $Fy \wedge H^*y$] $(y = x)$] (Gx). Of course, this fails to give us an English equivalent for *the F is G*; it simply tells us, in a context-independent idiom, what proposition it expresses. In fact, demanding an equivalent *English* sentence might be asking for too much in any case, since failures of equivalence occur in other contexts. For example, the same sorts of considerations apply to Russell's analysis of cardinality quantifiers. There is no guarantee that (10a) and (10b) uttered at relevantly similar contexts, will be equivalent. Yet this is precisely what Russell's analysis, together with [ES-Q], would entail:

(10a) Two cars approached.
(10b) A car approached, a car distinct from the first approached, and no other cars approached.

Thus, the theory that remains, while not quite Russell's theory, cannot be assumed to be inadequate merely because it fails to meet the equivalence condition, since that condition appears implausibly strong, threatening not simply Russell's analysis of descriptions, but his analysis of cardinality quantifiers as well.

3 Descriptions and Predication

What is Russell's theory of predicative descriptions?

Russell's treatment of the indefinite description *an F* is well-known to logic students: it simply assimilates *an F* to the quantifier *some F*. (11a) and (11b) alike are thus rendered as (11c) (the use of 'human' in the analysis is discussed below):

(11a) Socrates met a man.
(11b) Socrates met some man / someone.
(11c) $\exists x(\text{Human}(x) \wedge \text{Socrates met } x)$

Russell's approach to occurrences of indefinite descriptions in predicative position – so-called predicate nominals – is modeled on his treatment of indefinite noun phrases in subject position. (12a) is rendered as (12b):

(12a) Socrates is a man.
(12b) $\exists x(\text{Human}(x) \wedge \text{Socrates} = x)$

In contrast, a simple predication such as (13a) is rendered as the atomic sentence (13b):

(13a) Socrates is human.
(13b) Human(Socrates)

(11a) and (12a) do not share the same logical form *per se*: the relation obtaining between Socrates and the individual quantified over in (11a) is the *has met* relation, whereas in (12a) it is the identity relation – a logical constant. Nonetheless, there is an important sense in which Russell does ascribe the very same form to sentences exemplifying (11a) and sentences exemplifying (12a). They both exemplify the structure: '$\exists x(\text{Human}(x) \wedge \mathbf{R}(\text{Socrates}, x))$.'

It may initially appear perplexing that predicates and predicate nominals should receive differential treatment, given that, as James Higginbotham (1993) has suggested, the indefinite article in (12a) seems the "merest syntactic grace note." Addressing this concern, Russell writes:

> The proposition 'Socrates is a man' is no doubt equivalent to 'Socrates is human,' but it is not the very same proposition. The *is* of 'Socrates is human' expresses the relation of subject and predicate; the *is* of 'Socrates is a man' expresses identity. It is a disgrace to the human race that it has chosen to employ the same word 'is' for these two entirely different ideas – a disgrace which a symbolic logical language of course remedies. The identity in 'Socrates is a man' is identity between an object named . . . and an object ambiguously described. (Russell 1919: 71)

Russell supplies no argument for his claim that the copula in 'Socrates is a man' is, in fact, the identity relation, or, more precisely, the relational property of being identical to something human. (Strictly speaking, 'Socrates is a man' is not an identity sentence

for Russell, since, as he would be the first to point out, 'a man' cannot occur as an argument of 'Socrates = *x*'.) Surely, intuition is at best silent on this question, if not decidedly opposed to the identity interpretation. Still, Russell does provide us with a uniform treatment of indefinites, and this counts in favor of his proposal. It would be strange if indefinite descriptions unaccountably played two distinct logical roles – corresponding to their respective *syntactic* roles as noun phrase and as predicate – especially given the fact that interpreting the predicative occurrences according to Russell provides the intuitively correct truth-conditions for the relevant class of sentences. So, while it isn't directly supported by intuition, we have strong methodological reasons for favoring Russell's proposal. All other things being equal, then, it is to be preferred to an account that assigns indefinites two distinct logical roles – as quantifiers and as predicates.

It will be observed that Russell's approach requires, for every general term, a corresponding predicate adjective true of exactly those things in its extension. For example, to regiment (12a) there must be a (simple) predicate adjective true of exactly those things in the extension of 'man.' While such an expression exists in the current case, a corresponding adjective will not be available for every general term – for example 'logician.' This presents a difficulty: the Russellian cannot suppose that the predicate corresponding to 'a logician' is just 'a logician,' as this presupposes an account of the expression being analyzed. But, neither can she assume that 'Logician(Russell)' makes sense – that a general term can function as an adjective – since this would effectively undermine a distinction that she is at pains to uphold. What the Russellian must maintain is that the failure for there to be a predicate for every general term is a linguistic accident – a defect of natural language. Although not fatal, this has the unwelcome consequence that, for many sentences containing predicate nominals (such as 'Russell is a logician'), a Russellian paraphrase (in the same language) is unavailable.

Russell's analysis of predicative occurrences of *definite* descriptions recapitulates his strategy in analyzing indefinites. For Russell, the sentences 'Whitehead *met* the author of *Principia*' and 'Whitehead *is* the author of *Principia*' exemplify a common structure: '[the *x*: *x* wrote *Principia*] (**R**(Whitehead, *x*)).' (I adopt the restricted-quantifier notation for readability.) The former translates as: '[the *x*: *x* wrote *Principia*] (Met(Whitehead, *x*)),' whereas the latter is: '[the *x*: *x* wrote *Principia*] (Whitehead = *x*).' Again, this account is attractive in that it treats definite descriptions in a uniform and truth-conditionally adequate manner – showing, in effect, how their diverse surface syntax belies a uniform logical role. In addition, it dovetails with Russell's account of predicate nominals, providing a uniform account of the predicative occurrence of both definite and indefinite descriptions. It would be rather *ad hoc* to suppose that predicative *an F* functions logically as a predicate but that predicative *the F* functions either as a term or a quantifier. (But see Fiengo and May (1996) for considerations that favor such a treatment.)

The picture that emerges is that sentences exemplifying α *V-s an F* possess the logical form '[some *x*: F*x*] (**V**(α, *x*))' (where α is a singular term and **V** is the relation corresponding to 'V'), while sentences exemplifying α *BE an F* possess the logical form '[some *x*: F*x*] (α = *x*)'. Similarly, sentences of the form α *V-s the F* are analyzed '[the *x*: F*x*] (**V**(α, *x*)), whereas sentences of the form α *BE the F* are analyzed as '[the *x*: F*x*] (α = *x*).' For reference, let's call this proposal Russell's Theory of Predicative Descriptions (*RTPD*).

The argument from awkwardness

An important challenge to Russell's proposal that quantifiers can be realized, at surface grammar, as predicates is that it yields odd results. For example, it sanctions as grammatical such sentences as 'John is most Democrats' or 'John is twelve apostles,' both of which seem to be uninterpretable and, indeed, ungrammatical. Prima facie, what I shall call *the argument from awkwardness* appears to be a significant worry.

The first thing to say is that there is reason to think that our resistance to these sentences is pragmatic, since other such contexts are less objectionable:

John is everyone / one person who has read Richardson's *Clarissa* in its entirety.
John is no one / someone you should meet.

In addition, similar examples involving quantifiers in subject position are equally unacceptable. That is, in general, those sentences with quantifiers in predicative position that strike us as unacceptable do not become any more acceptable if we place the predicatively-occurring quantifier in subject position. Consider 'John is most Democrats' or 'John is both candidates.' Moving the quantifier in 'John is most Democrats' to subject position ('Most democrats are [identical to] John') does not increase acceptability. Similarly, 'Both candidates are [identical to] John' is scarcely better than 'John is both candidates.' This suggests that what explains the unacceptability of quantifiers in predicative position is not that they are playing a role that quantifiers are strictly prohibited from playing. For, in general, substituting a determiner for *det* (other than *the*, *a*, and *some*) and a name for t in '[*det* x: Fx] ($x = t$)' will either produce nonsense or, at best, express an intelligible but nonetheless awkward-sounding sentence.

The Russellian response to the argument from awkwardness, then, is to demand a context of the form Q *BE* α that is acceptable but whose acceptability is compromised by converting it to α BE Q. Until such a case is presented, Russell's picture is intact.

The argument from scope

Another argument, discussed by James Higginbotham (1987) and Delia Graff (2001), concerns a diagnostic for determining whether or not a surface predicate is a quantifier at the level of logical form. The idea is quite straightforward: if a certain predicate is in fact just the surface realization of a quantifier, then it should exhibit properties characteristic of quantifiers. In particular, it should interact with negation and similar devices to produce distinct readings, depending on which expression is assigned primary scope.

To fix ideas, let's consider an example. According to the criticism, if, as we are supposing, (14b) provides the logical form of (14a), then the negation of (14a) – namely, (14c) – should be ambiguous as between (14d) and (14e):

(14a) John is a bachelor.
(14b) [an x: Bachelor (x)] (John = x)
(14c) John is not a bachelor.

(14d) ¬[an x: Bachelor (x)] (John = x)
(14e) [an x: Bachelor (x)] ¬ (John = x)

Yet, (14e) is clearly not an available reading of (14c). Intuitively, (14c) cannot be used to *say* that there is a bachelor who, as it happens, is not John. Similarly, 'John is not a Martian' does not have a reading according to which it entails that there are Martians. So, it seems we have a good reason to doubt that Russell's account of indefinite descriptions applies to predicate nominals.

The same considerations do *not* extend unproblematically to *definite* descriptions in predicative position, since (15a) below does appear to give rise to an ambiguity: both (15b) and (15c) seem to be available readings of (15a):

(15a) John is not the mayor.
(15b) ¬[the x: Mayor(x)] (John = x)
(15c) [the x: Mayor(x)] ¬ (John = x)

Nonetheless, Graff (2001) argues plausibly that the ambiguity admits of a pragmatic explanation – that the availability of (15c) is determined by the mutually held assumption that *someone* is the mayor (I have changed her example). To take another case, if I utter, 'Whitehead is not the sole author of *Principia Mathematica*,' the reading according to which 'the sole author of *Principia Mathematica*' takes scope over the negation seems unavailable. The utterance in no way says or implies that *Principia Mathematica* has a single author.

Thus, it looks as if a case can be made against RTPD, since it makes predictions about scope that seem not to be borne out by the data. Yet, there is a question as to the validity of the scope test for quantifierhood. To see why this is so, consider (16a):

(16a) Mary is someone who smokes.

(16a) contains a *quantifier* in predicative position. Russell would render it as (16b):

(16b) [some x: Smokes(x)] (Mary = x)

Yet, negating (16a) does not generate an ambiguity between the readings supplied by (16d/e), since there is no tendency to interpret (16c) as (16e):

(16c) Mary is not someone who smokes.
(16d) ¬[some x: Smokes(x)] (Mary = x)
(16e) [some x: Smokes(x)] ¬ (Mary = x)

Similarly for the modal case:

(17a) John is someone who might have proved Goldbach's conjecture.
(17b) *Poss.* [some x: Prove(x, p)] (x = John)
(17c) [some x: Prove(x, p)] *Poss.* (x = John)

There is no reading on which (17a) entails that someone actually proved Goldbach's conjecture. Thus, (17c) is not an available reading of (17a).

These examples show that a quantifier in predicative position does not invariably give rise to multiple readings when within the scope of negation or other operators. Why this should be the case is not immediately clear, but the evidence suggests that it occurs. If so, the Higginbotham–Graff diagnostic is inapplicable – it does not provide a positive test for the presence of a quantifier: there are some quantifiers that occur in predicative position that seem to take an obligatory narrow scope. It might be argued that all that my examples show is that the predicative occurrences of *someone who* are not quantificational, and precisely because they do not interact in expected ways with negation and related devices. But this would need further argument to be made plausible since the semantic contribution of predicative occurrences of *someone who* appears to be quantificational.

4 Conclusion

We have seen how Russell's characterization of the logical form of description sentences conflicts with a highly plausible proposal regarding incomplete quantification. One way to resolve the tension is to retain the view that descriptions are restricted quantifiers but at the same time to deny the most straightforward implementation of Russell's theory in an account of natural language quantification – namely, [EQ*]. This is to concede that Russell's theory cannot capture the competent speaker's ability to assign propositions to utterances of description sentences. But, as suggested, this is not really all that much of a worry, since [EQ*] appears to be unreasonably strong in any case. In addition, a close relative of Russell's theory remains viable. The status of RTPD is a bit more problematic. Although there are no serious arguments against the view, it can hardly be said to be independently motivated. There are other issues that intersect with the ones discussed which I have not been able to cover – the referential/attributive distinction, the 'implicit' approach to incompleteness, the Russellian treatment of unbound anaphora, to name but a few. Publications addressing these topics can be found under Further reading below.

References

Bach, Kent (1988) *Thought and Reference*. New York: Oxford University Press.

Evans, Gareth (1982) *The Varieties of Reference*. New York: Oxford University Press.

Fiengo, Robert and May, Robert (1996) *Indices and Identity*. Cambridge, MA: MIT Press.

Graff, Delia (2001) Descriptions as Predicates. *Philosophical Studies*, 102, 1–42.

Heim, Irene and Kratzer, Angelica (1997) *Semantics in Generative Grammar*. Malden, MA: Blackwell.

Higginbotham, James (1987) Indefiniteness and predication. In E. Reuland and A. ter Meulen (eds.), *The Representation of (In)Definiteness* (pp. 43–70). Cambridge, MA: MIT Press.

Higginbotham, James (1993) Grammatical form and logical form. In James Tomberlin (ed.), *Philosophical Perspectives*, (Vol 7): *Language and Logic*. Atascadero, CA: Ridgeview.

Larson, Richard and Segal, Gabriel (1995) *Knowledge of Meaning*. Cambridge, MA: MIT Press.

Loar, Brian (1976) The semantics of singular terms. *Philosophical Studies*, 30, 353–77.

Mates, Benson (1973) Descriptions and reference. *Foundations of Language*, 10, 409–18.

Moore, G. E. (1944) The theory of descriptions. In Paul Arthur Schilpp (ed.), *The Philosophy of Bertrand Russell* (pp. 177–225). Evanston, IL: Northwestern University Press.

Neale, Stephen (1990) *Descriptions*. Cambridge, MA: MIT Press.

Neale, Stephen (1995) The philosophical significance of Gödel's slingshot. *Mind*, 104, 761–825.

Ostertag, Gary (ed.) (1998) *Definite Descriptions: A Reader*. Cambridge, MA: MIT Press.

Reimer, Marga (1992) Incomplete descriptions. *Erkenntnis* 37, 347–63.

Russell, Bertrand (1905) On denoting. *Mind*, 14, 479–93. Reprinted in Ostertag (1998) pp. 35–49.

Russell, Bertrand (1919) Descriptions. In *Introduction to Mathematical Philosophy*. London: George Allen & Unwin. Reprinted in Ostertag (1998) pp. 67–76. Page references are to reprint.

Russell, Bertrand (1944) Reply to criticisms. In P. A. Schilpp (ed.), *The Philosophy of Bertrand Russell*. Evanston, IL: Northwestern University Press.

Schiffer, Stephen (1995) Descriptions, indexicals, and belief reports: some dilemmas (but not the ones you expect). *Mind*, 104, 107–31. Reprinted in Ostertag (1998) pp. 369–95. Page references are to reprint.

Stanley, Jason and Zoltán Gendler Szabó (2000) On Quantifier Domain Restriction. *Mind and Language*, 15, 219–61.

Stanley, Jason and Williamson, Timothy (1995) Quantifiers and context dependence. *Analysis*, 55, 291–5.

Strawson, P. F. (1950) On referring. *Mind*, 54, 320–44. Reprinted in Ostertag (1998).

Whitehead, A. N. and Russell, Bertrand (1925–7) *Principia Mathematica*, vol. I, 2nd edn. Cambridge: Cambridge University Press. Reprinted in Ostertag (1998) pp. 51–65. Page references are to reprint.

Further Reading

Bach, Kent (1994) Ramachandran vs. Russell. *Analysis*, 54.3, 183–6.

Barwise, Jon and Cooper, Robin (1981) Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4, 159–219.

Bezuidenhout, Anne (1997) Pragmatics, semantic underdetermination, and the referential/attributive distinction. *Mind*, 106, 375–409.

Burge, Tyler (1974) Truth and singular terms. *Journal of Philosophy*, 8, 309–25. Reprinted in Karel Lambert (ed.), *Philosophical Foundations of Free Logic*. New York: Oxford University Press, 1991, 189–204.

Chomsky, Noam (1975) Questions of form and interpretation. *Linguistic Analysis*, 1, 75–109.

Donnellan, Keith (1966) Reference and definite descriptions. *Philosophical Review*, 75, 281–304. Reprinted in Ostertag (1998).

Ganeri, Jonardon (1995) Contextually incomplete descriptions – a new counterexample to Russell? *Analysis*, 55.4, 287–90.

Kaplan, David (1978) Dthat. In A. Martinich (ed.), *The Philosophy of Language*, 3rd edn. New York: Oxford University Press.

Kripke, Saul (1998) Speaker's reference and semantic reference. In P. French, T. Uehling and H. Wettstein (eds.), *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press, 6–27. Reprinted in Ostertag (1998).

Lambert, Karel (1991) A theory of definite descriptions. In Karel Lambert (ed.), *Philosophical Foundations of Free Logic*. New York: Oxford University Press. Reprinted in Ostertag (1998).

Lewis, David (1979) Scorekeeping in a language game. *Journal of Philosophical Logic*, 8, 339–59.

May, Robert (1985) *Logical Form: Its Structure and Derivation*. Cambridge, MA: MIT Press.

Neale, Stephen (1998) Grammatical form, logical form, and incomplete symbols. In A. Irvine and G. Wedeking (eds.), *Russell and Analytic Philosophy*. Toronto: University of Toronto Press. Reprinted in Ostertag (1998).

Ostertag, Gary (1999) A Scorekeeping error. *Philosophical Studies*, 96, 123–46.

Ramachandran, Murali (1994) A Strawsonian objection to Russell's theory of descriptions. *Analysis*, 53.4, 209–12.

Ramachandran, Murali (1995) Bach on behalf of Russell. *Analysis*, 55.4, 283–87.

Recanati, François (1993) *Direct Reference: From Language to Thought*. Oxford: Blackwell.

Salmon, Nathan (1991) The pragmatic fallacy. *Philosophical Studies*, 63, 83–97.

Szabô, Zoltan-Gendler (2000) Descriptions and uniqueness. *Philosophical Studies*, 101, 29–57.

Westerståhl, Dag (1986) Quantifiers in formal and natural languages. In D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic: Topics in the Philosophy of Language*, vol. 4, Dordrecht: Kluwer.

Wettstein, Howard (1981) Demonstrative reference and definite descriptions. *Philosophical studies*, 40, 241–57. Reprinted in Ostertag (1998).

13

Russell's Theory of Definite Descriptions as a Paradigm for Philosophy

GREGORY LANDINI

In one of his posthumously published writings, Ramsey spoke of the theory of definite descriptions that Russell set out in his 1905 article "On Denoting" as a "paradigm for philosophy" (Ramsey 1931: 263n). Russell had begun a new scientific method in philosophy – the investigation of logical form – and its most salient example was the monumental work *Principia Mathematica*. But what precisely was the paradigm? As it is commonly articulated, Russell's theory of definite descriptions exemplified a "theory of incomplete symbols," and a "misleading form thesis." Haack puts it as follows: "If the grammatical form of a recalcitrant sentence is taken as indicative of its 'logical form,' then, indeed, assignment either of 'truth' or 'false' to it gives rise to difficulty. Once, however, it is recognized that the grammatical form of the sentence is misleading as to its logical form, the difficulty vanishes" (Haack 1996: 53). But what is 'logical form,' and what is it to render the logical form of a statement? Is the analysis of logical form a part of a theory of sense and reference, part of philosophical linguistics, part of the philosophy of mind, part of metaphysics?

1 Russell's Paradigm

In his now famous article "On Denoting," Russell (1905) lays down the proscription that in transcribing expressions of ordinary language into the canonical language of symbolic logic, ordinary proper names and definite descriptions should be treated alike. Moreover, transcriptions of ordinary statements involving proper names or definite descriptions into symbolic logic are to have the syntactic form of quantificational statements. Consider transcribing the sentence,

(1) Gödel was a mathematician.

In Russell's view, ordinary proper names are "disguised definite descriptions." Russell's technique requires that the name "Gödel" be associated with some definite description 'the entity x such that Ax,' where A contains descriptive information. It is not necessary that every transcription associates the same descriptive information with the name, so long as the descriptive attributes in question are coexemplifying. But some

description must be associated with the name. Let Ax be 'x proved the incompleteness of Arithemtic,' and let us write this as 'Px.' Then letting 'Mx' represent 'x is a mathematician,' sentence (1) is to be transcribed into symbolic logic as:

(1R) $(\exists x)(Pz \equiv_z z = x \,.\&.\, M(x))$.

This says that one and only one entity proved the incompleteness of Arithmetic and that one was a mathematician. For convenience, Russell introduces an abbreviated way of writing sentences such as (1R). Where Bv is some well-formed formula, *Principia* offers the following stipulative contextual definition:

(*14.01) $[(\iota z)(Az)][B(\iota z)(Az)/v] = df\ (\exists x)(Az \equiv_z z = x \,.\&.\, Bx/v)$.

The notation $[(\iota z)(Az)][\ldots]$ is Russell's scope marker, and is required because the context B might be syntactically complex. Consider, for instance,

(2) If Aeneas was not a Trojan then Virgil's great epic is fiction.

Associate with the name Aeneas the descriptive information 'x is founder of Rome' (represented by 'Ax'), and put 'Tx' for 'x is a Trojan,' and q for 'Virgil's great epic is fiction.' Then there are the following possible transcriptions:

$(2R_a)$ $\sim(\exists x)(Az \equiv_z z = x \,.\&.\, Tx) \vee q$
$(2R_b)$ $(\exists x)(Az \equiv_z z = x \,.\&.\, \sim Tx) \vee q$
$(2R_c)$ $(\exists x)(Az \equiv_z z = x \,.\&.\, \sim Tx \vee q)$

They are not all equivalent if in fact nothing, or more than one thing satisfies the description 'x is founder of Rome.' In such a case, $(2R_a)$ would be true, $(2R_b)$ would be false if q is false, and true otherwise, and $(2R_c)$ would be false no matter what q is. It would not be possible, therefore, to write '$\sim T(\iota z)(Az) \vee q$', since this would be ambiguous as to the different scopes of the definite description. Accordingly, Russell introduces his scope marker and writes,

$\sim\ (\iota z)(Az)[T(\iota z)(Az)] \vee q$
$(\iota z)(Az)[\ \sim T(\iota z)(Az)] \vee q$
$(\iota z)(Az)[\ \sim T(\iota z)(Az) \vee q]$

respectively. When the scope marker is outside the entire formula, the description is said to have 'primary occurrence,' and there will be one and only one such occurrence. All other occurrences are called 'secondary occurrences.' Primary and secondary occurrences are always equivalent in truth-functional contexts when the description is satisfied. Russell derives the theorem:

$(\exists x)(Az \equiv_z z = x) \,.\supset.\, \mathcal{T}\{(\iota z)(Az)[B(\iota z)(Az)]\} \equiv (\iota z)(Az)[\mathcal{T}\{B(\iota z)(Az)\}]$,

where $\mathcal{T}$ is any truth-functional context.

Russell adopts the convention of omitting scope markers when the smallest possible scope is intended. The convention of omitting scope markers does not, as it has some-

times been argued, cause difficulties concerning the order that definitions are to be applied (Geach 1950). In *Principia*, the identity sign is defined as follows:

(*13.01) $x = y = \text{df}\ (\varphi)(\varphi x \equiv \varphi y)$.

It is not possible, however, to apply (*13.01) to

$(\iota z)(Az) = (\iota z)(Az)$

to yield,

$(\varphi)(\varphi(\iota z)(Az) \equiv \varphi(\iota z)(Az))$.

Definitions such as (*13.01) apply only to genuine singular terms of the formal language of *Principia*, and expressions such as '$(\iota z)(Az)$' are not among its genuine singular terms. Thus, the smallest possible scope yields

$(\iota z)(Az)[(\iota z)(Az) = (\iota z)(Az)]$.

Applying (*14.01) we get:

$(\exists x)(Az \equiv_z z = x \,.\&.\, x = x)$.

It is only now that (*13.01) can be applied:

$(\exists x)(Az \equiv_z z = x \,.\&.\, (\varphi)(\varphi x \equiv \varphi x))$.

The conventions on the omission of scope markers in *Principia*, together with the fact that definite descriptions are not singular terms, fully determine the proper order of the elimination of defined signs.

Scope markers act as though they bind occurrences of expressions of the form '$(\iota z)(Az)$.' If there are multiple occurrences of the same descriptive expression, it is Russell's intent that they each be tied to their scope marker. But it is possible to represent different scopes. For this, Russell adopts the convention that the left-most description be taken first, and then in order as one proceeds to the right. Thus, for instance, restoration of scope markers in,

$(\iota z)(Az)\ R\ (\iota y)(Ay)$

yields the following:

$(\iota z)(Az)[\ (\iota y)(Ay)[(\iota z)(Az)\ R\ (\iota y)(Ay)]\ ]$.

Applying the contextual definition (*14.01), this is:

$(\exists x)(Az \equiv_z z = x \,.\&.\, (\exists y)(Az \equiv_z z = y \,.\&.\, xRy))$.

Ordinary grammar presents statements involving definite descriptions as if they are subject predicate, but on Russell's theory, ordinary grammar is misleading with respect to the role such expressions play when transcribed into symbolic logic. The proper logical grammar has a quantificational structure. The quantificational structures and scope distinctions that accrue to the transcription of ordinary proper names and definite descriptions on Russell's theory may be exploited to explain the role of such names in existential, identity, doxastic, modal, fictional, and counterfactual contexts. To get a glimpse of this, let us examine a few examples. Consider the statement,

Aeneas exists,

whose natural language syntax predicates 'existence' of the subject expression 'Aeneas.' This is transcribed as

$$(\exists x)(Az \equiv_z z = x).$$

As we see, 'existence' is not to be adopted as a logical predicate. For convenience, Russell introduces the definition,

$$(*14.02) \quad E!(\iota z)(Az) = df\ (\exists x)(Az \equiv_z z = x).$$

Surprisingly, the superficial similarity of $E!(\iota z)(Az)$ and $\varphi(\iota z)(Az)$, where φ is a predicate letter of the formal language, has mislead some into arguing that Russell has not given a uniform treatment of the expressions of his formal language (Lambert 1990). But quite clearly, (*14.02) is not intended to introduce a new predicate expression E! into the formal language. In the definition, E! is not separable from $E!(\iota z)(Az)$, and it is quite ungrammatical to write E!x. To be sure, one can write, $(\exists x)(x = v)$, and so it has appeared to some that one can predicate existence in spite of Russell's best efforts to the contrary. But the objection is misguided. The above formula does not express the statement that v exists. It expresses the statement that some individual is identical with v. It is the presence of the free individual variable v that commits one to existence here and not identification with some individual. Indeed, formulas $(\varphi)(\varphi v \supset \varphi v)$ and φv, with a free variable v, have equal claim with $(\exists x)(x = v)$ to be called 'existence predications.' Russell anticipates Quine in maintaining that ontological commitment is given with the variables of quantification, and not by predication of a special existence (or being) predicate.

The presence of quantificational structure also enables Russell's theory to resolve Frege's famous puzzle concerning the informativity of identity statements made with proper names. A *pure* predicate calculus contains no individual constants, or function constants. But in applying the calculus, modern mathematical logic allows that one may form a theory by adding proper axioms and any countable number of individual constants and function constants to the base language of the predicate calculus. In rendering the semantics of such an extended language, an interpretation of the language is given which fixes an assignment of referents to the constants which does not vary in the way that assignments to the individual variables does. The syntax does not encode

any semantic information concerning interpretation of the constants. To understand the semantic contribution that an individual constant makes to the meaning of an expression containing it, one must understand the referent of the constant assigned by the interpretation. For instance, if an applied symbolic logic employed the individual constants 'Hesperus' and 'Phosphorus,' then in grasping what the interpretation assigns to

'Hesperus = Phosphorus'

we must grasp what the interpretation assigns to 'Hesperus' and to 'Phosphorus.' It either assigns each the same entity (in the domain of the interpretation) or it does not. In the first case, Russell rightly points out, the statement is uninformative, in the second it is simply false (Whitehead and Russell 1962: 67). Russell, like Quine (1979) after him, holds that in applying a symbolic logic to form a theory, we are not to add individual constants (or function constants) to the language of symbolic logic. The only singular terms of a theory are to be the individual variables. In Russell's view, syntactic structures should encode as much semantic information as possible. If 'Hesperus' is associated with a definite description such as 'the morning star' and 'Phosphorus' with 'the evening star,' an ordinary language identity statement such as 'Hesperus = Phosphorus' will be transcribed as:

$$(\iota z)(Mz) = (\iota z)(Ez).$$

Applying (*14.01), this is:

$$(\exists x)(Mz \equiv_z z = x \,.\&.\, (\exists y)(Ez \equiv_z z = y \,.\&.\, x = y)).$$

(This says that there is exactly one M and exactly one E and that they are identical.) Russell's transcriptional technique of associating an ordinary proper name with a definite description enables his formal syntax to encode semantic information into identity statements. In this way, the semantic informativity of an identity statement is part of the formal syntax.

Russell holds that since the ordinary language syntax of definite descriptions does encode semantic information, he can generate a 'proof' that definite descriptions should not be transcribed into symbolic logic as singular terms and *must* be treated as 'incomplete symbols' to be contextually defined. His proof is simple. Expressions of the form,

$$c = (\iota z)(\varphi z)$$

Russell explains, are never 'trivial,' for unlike expressions such as 'c = d,' they encode semantic information in their syntax. Thus '$(\iota z)(\varphi z)$' cannot be a genuine singular term, else the fact that its syntax encodes semantic information is lost. Russell's 'proof' does show that definite descriptions are not 'singular terms' in the sense of being individual constants whose syntax encodes nothing of the semantics. But this falls short of demonstrating that the *only* way to syntactically encode this semantic information is

by means of a theory of incomplete symbols. It is possible to both introduce proper axioms which keep definite descriptions as singular referring expressions, and at the same time capture the fact that they syntactically encode semantic information. Add to the language one individual constant t, and the following axiom schema:

$$(\iota x)(Ax) = y .\equiv:. Ay .\&. Az \equiv_z z = y :v: \sim(\exists x)(Az \equiv_z z = x) \,\&\, y = t$$

This approach has come to be called the 'chosen object view,' and a version of it was adopted by Frege. The approach conveniently avoids the many complications of scope imposed by Russell's approach of contextual definition. It is, however, highly artificial. If more than one entity satisfies the description, or if nothing does, the referent of the definite description is simply identified as whatever the interpretation assigns to t.

The scope distinctions that accrue to proper names and definite descriptions in Russell's approach of contextual definition to transcription are indeed inconvenient, but they are also precisely what is most attractive about the theory. They are the very feature that is called upon to solve the puzzles that infest the use of names and definite descriptions in ordinary inferences. For example, they explain how it is that the unassailable law of identity appears, nonetheless, to fail in contexts which are not truth-functional. Let $\mathcal{T}$ represent a sentential context that is not truth-functional. Then primary and secondary scopes of a definite description $(\iota z)(Az)$ will not be equivalent, even when $E!(\iota z)(Az)$. We may have,

$\mathcal{T}\{B(\iota z)(Az)\}$
$(\iota z)(Az) = (\iota z)(Bz)$
$\sim\mathcal{T}\{B(\iota z)(Az)\}$

For instance, let $\mathcal{T}$ be the context of an ascription of belief to Galileo Galilei. If the name 'Hesperus' refers non-descriptively in its occurrence in

(3) Galileo believed that Hesperus orbits the sun,

then the law of identity would apply, and the substitution of 'Phosphorus' for the name should preserve truth. But the identity of Hesperus and Phosphorus does not entail that Galileo believed that Phosphorus orbits the sun simply because he believed that Hesperus does. Russell's theory provides a solution. The names 'Hesperus' and 'Phosphorus' carry descriptive information relevant to the nature of Galileo's belief. By associating 'Hesperus' with a definite description such as 'the morning star,' and 'Phosphorus' with 'the evening star,' we see that (3) is ambiguous between different scopes. It may mean,

(3a) Galileo believes $(\exists x)(Mz \equiv_z z = x .\&. x$ orbits the sun)

which is a *de dicto* ascription to Galileo of particular descriptive information that he employs in using the name Hesperus. Or it may mean,

(3b) $(\exists x)(Mz \equiv_z z = x .\&.$ Galileo believes x orbits the sun).

In this case, one does not intend to give any information about the way in which Galileo himself would express his belief. The ascription is said to be *de re*. Galileo is said to have a '*belief of*' an object, the morning star, that it orbits the sun. We can see this even more saliently if we quantify over attributes, replacing (3b) with

(3bb) $(\exists\varphi)(\varphi z \equiv_z Mz$.&. Galileo believes $(\exists x)(\varphi z \equiv_z z = x$.&. x orbits the sun)).

In this way, we see clearly that in a *de re* ascription of belief the descriptive content Galileo himself employs to single out Hesperus is left unspecified. Accordingly, since Hesperus (the morning star) is identical with Phosphorus (the evening star), we have,

$Mz \equiv_z Ez$,

and so (3b) is equivalent to

(3c) $(\exists x)(Ez \equiv_z z = x$.&. Galileo believes x orbits the sun).

Similarly, (3bb) is equivalent to

(3cc) $(\exists\varphi)(\varphi z \equiv_z Ez$.&. Galileo believes $(\exists x)(\varphi z \equiv_z z = x$.&. x orbits the sun)).

Thus if 'Galileo believes Hesperus orbits the sun' is to be understood by means of (3bb), it follows that 'Galileo believes Phosphorus orbits the sun' as understood by means of (3cc). No similar move is possible for (3a), and the contextual elimination of the ordinary names leaves nothing to which the law of identity could apply.

As we see, the benefits of adopting a Russellian approach to the transcription of ordinary names and definite descriptions are many. They are due to the possibility of finding complex quantificational structures, logical forms, where ordinary language employs simple grammatical forms. Interestingly, we shall see that it is precisely this feature that has been the focus of criticism from those who object to the theory.

2 The Description Theory and Logical Form

The *Description Theory* of what an ordinary proper name denotes holds that associated with each name as used by a group of speakers who believe and intend that they are using the name with the same denotation, is a description or set of descriptions cullable from their beliefs which an item has to satisfy to be the bearer of the name. The theory owes its origins to Russell's thesis that ordinary proper names are disguised descriptions, and to Frege's famous position that a proper name expresses a *Sinn* (sense) (for a given person at a time). Frege did not take the logical form of expressions involving proper names and definite descriptions to be quantificational, but in his 1892 article "On Sense and Reference" he held that a name, just as a definite description, presents descriptive qualities which the purported object referred to by the name must satisfy (Frege 1980). Frege offers as an example the name 'Aristotle,' writing that its sense

might, for one person, be taken to be 'the pupil of Plato and teacher of Alexander the Great' and for another person, 'the teacher of Alexander the Great.' This suggests that the sense of a name (for a speaker at a time) will be the same as the sense of some definite description. For this reason, the Description Theory is often attributed to both Frege and Russell.

As Evans rightly points out, the description may be arrived at by averaging out the beliefs of different speakers; the theory is by no means committed to the thesis that every user of the name must be in possession of a description, or figure in the cluster of descriptive information every user of the name associates with the name (Evans 1973). Thus the Description Theory must be distinguished from what Evans calls the "description theory of speaker's denotation," which holds that an ordinary proper name denotes an entity upon a particular occasion of its use by a speaker just in case that entity uniquely satisfies all (or most) of the descriptive information the speaker associates with the name. In any event, the Description Theory construes Russell's theory of descriptions as a meaning analysis of the use of names and descriptions by speakers of a language.

Strawson (1950) famously objected that Russell's theory of definite descriptions misunderstands the function of singular noun phrases in communication and fails to do justice to the use of referring terms in natural language. Donnellan (1966) continued this line of criticism of Russell's theory, maintaining that one must distinguish 'attributive' from 'referential' uses of definite descriptions in natural language. With the right stage setting, a person may succeed in referring to a person drinking water by uttering 'Who is the man drinking a martini?' This is a 'referential' use of a definite description, not the 'attributive' use which picks out an entity only in so far as it satisfies the description.

Russell held that a statement involving a definite description in a primary occurrence *entails* an existential statement that some entity satisfies the description. Strawson challenged this, and Donnellan agreed that the relationship is properly one of the presuppositions behind speech acts of communication. Donnellan modifies Strawson's account, arguing that different presuppositions explain the fact that both referential and attributive uses presuppose existential statements. A referential use of a given definite description presupposes an existential statement that something satisfies the description simply because, under normal circumstances of communication, a person tries to describe correctly what he/she want to refer to because this is the best way to get his/her audience to recognize what is being referred to. Nonetheless, it may well be possible for the audience to locate the referent independently of its satisfying the descriptive information. So there is no entailment. On the other hand, an attributive use of a definite description presupposes that something satisfies the description because if nothing fits the description the linguistic purpose of the speech act (of asserting, questioning, or ordering) will be thwarted. Donnellan's distinctions are designed to amend Strawson's theory that definite descriptions are referential – a theory that appeals to the presuppositions of acts of communication to explain whether statements involving definite descriptions are truth-valued, or simply such that the speech act in question misfires.

Following Russell's own lead in his reply to Strawson, defenders of Russellian description theories reject such objections because they seem based upon a misunder-

standing of the intents and purposes of the theory. As Russell put it, the theory of descriptions did not intend to account for the 'egocentric' use of words – words whose reference is dependent on the pragmatic circumstances, times, and places of utterance (Russell 1959: 239f). Strawson's appeal to *presupposition* rests upon intuitions about the kinds of assertions made by an *utterance* on an occasion of *use*. As Bach puts it: "There is no legitimate notion of a semantic presupposition (as a property of sentences). And it turns out that there are several different kinds of pragmatic presupposition, each of which is a property of utterances" (Bach 1987: 98). Russell's theory of definite descriptions does not concern utterances or assertions. Donnellan's distinction between 'referential' and 'attributive' occurrences of definite descriptions properly applies to the *use* of definite descriptions. Whether a definite description is used attributively or referentially is a function of the sort of speech act a speaker makes on an occasion of utterance. This is a pragmatic consideration, not a semantic one. And the same may be said of the many other objections to Russell's theory of definite descriptions that follow the lead of Strawson. They rely upon the improper infusion of pragmatic elements into semantics.

Of course, the distinction between pragmatics and semantics can be slippery. Russell's early ontology posited the existence of a true or false 'proposition' as the 'meaning' of an 'asserted' sentence, and this is easily conflated with the postulation of utterances construed as the 'meanings' of a given type of speech act. Linguists tend to assume that language must be semantically analyzed in terms of mental constructs. Philosophers favor ontological approaches that render semantic analyses of natural language in terms of intensional entities such as properties, propositions, nonexistent objects, and the like. Both approaches to semantics leave themselves open to a blurring of the semantics/pragmatics distinction. Propositions, for instance, may seem like utterances, assertions, or speech acts of a sort, made on occasion of use. But 'reference' and 'truth,' which are normally semantic notions, are not properly semantic when taken to be properties of utterances. One must be on the lookout for conflations of this sort. Utterances involve the production of tokens of certain types of speech act, and belong to the pragmatic study of how context of utterance and speaker's intent contribute to the communication of meaning. In Austin's theory of speech acts, for instance, utterances of complete sentences are classified as 'locutionary', 'illocutionary,' or 'perlocutionary.' Acts of referring and communication of one's intended reference, are components of illocutionary speech acts. When an illocutionary act is a statement or a predication or other 'conative' act, it may be said to be true or false. These notions of reference and truth are a part of pragmatics and not semantics. As Bach (1987: 4) points out, not all notions of 'truth' and 'reference' are semantic. Perhaps we can see the source of Russell's sarcasm in writing that adherents of the ordinary language philosophy of Austin and the later Wittgensteinians "are fond of pointing out, as if it were a discovery, that sentences may be interrogative, imperative, or optative, as well as indicative" (Russell 1959: 217).

Bach's account of the pragmatics/semantics distinction is particularly illuminating. Pragmatics is the theory of communication and speech acts. The semantics of an expression, on the other hand, gives the information that a competent speaker can glean from it independently of any context of utterance. Whenever he hears a particular utterance of it in a given context, he uses this information, in tandem with specific

information available in that context, to understand the speaker's communicative intent. Semantic knowledge is not, on this view, supposed to be a compilation of general pragmatic information governing different possible circumstances of utterance for an expression of a given type. Semantic information can be gleaned independently of context of utterance only in so far as it is encoded in the syntactic structures of the language. The notion of semantics here is compositional – that is, there are complex expressions whose content is determined by the content of their parts. The independence of the compositional semantics from pragmatics (where context of utterance is involved) is a consequence of the adoption of theory of grammar according to which the syntactic types of the language in question encode the whole of its combinatorial semantics. In short, the province of semantics is linguistic grammatical types. An example of the compositional approach is Tarski-style model theoretic formal semantics, and its extensions to possible-worlds semantics for modal theories. The intent of such accounts is to give a systematic combinatorial account of logical consequence for syntactically formalized languages (whose formation and deductive transformation rules are explicit) in terms of truth (reference and satisfaction) in the domain of an interpretation.

Indeed, in contemporary discussions in the philosophy of mind and language, the combinatorial semantic theories of modern philosophical linguistics is often offered as an explanation of what Russell meant when he proclaimed that his theory of definite descriptions reveals that ordinary grammatical form can by misleading with respect to logical form. The so-called 'logical form' of a proposition or 'assertion' (in the semantic sense) specifies the truth conditions of propositions in terms of the recursive operations of a logical syntax. Cocchiarella (1989) characterizes an even stronger sense of logical form according to which logical forms specify not only the truth conditions of an assertion, but they also specify the cognitive structure of the assertion itself by providing an appropriate representation of the referential and predicable mental concepts that underlie the assertion.

This is an important enterprise in philosophical linguistics, and it is naturally allied with Chomsky's research program in linguistics. The leading idea here is that at least some grammatical structures are transformations of other structures, where words and phrases are displaced from syntactic positions typically associated with their semantic roles. The idea of a *transformational grammar* places a premium upon reconciling the quantificational structures produced by Russellian analyses of ordinary proper names and definite descriptions, with certain features of the ordinary grammar of categorical phrases. Phrases such as 'all *a*,' 'some *a*,' 'any *a*,' 'every *a*,' 'the *a*' (as well as 'most *a*,' and 'few *a*') where *a* is a common noun or noun phrase, do act as if plural subjects. Consider the phrase

Some moment does not follow any moment.

In their efforts to transcend the subject-predicate forms of categorical logic, Russellian and Fregean analyses abandoned the transformational nature of categorical phrases, writing

$(\exists x)(Mx \,\&\, (\forall y)(My \supset x \text{ does not follow } y))$.

This does not respect the fact that in the original phrase, the expressions 'some moment,' and 'any moment' appear in grammatical positions of singular terms. They may be removed to form,

$[]_i$ does not follow $[]_j$

If the integrity of the restricted quantifiers 'some moment' and 'any moment' as syntactic/semantic units can be preserved, one can regard

[Some moment]$_i$ [any moment]$_j$ {$[]_i$ does not follow $[]_j$},

as a transformation of the original, enacted by a displacement of the phrases (governed by the hypothesized transformation linguists call 'quantifier raising'). Instead of 'some moment,' and 'any moment' one can write '(∃xM),' '(AyM)' respectively and represent the logical form with:

(∃xM)(AyM)(x does not follow y).

By construing categorical phrases as restricted quantifiers in this way, rules such as the Subjacency Principle might then be called upon to explain transformational restrictions of scope governing the use of determining phrases in natural language. For example, 'some *a*' normally has a wider scope that 'any *a*,' but in

A moment precedes any moment,

we see that 'any *a*' has wider scope than 'a(n) *a*' for this means that for every moment there is some moment that precedes it. Allies of transformational grammar endeavor to preserve the Frege/Russell view that the proper logical form of categorical phrases is quantificational, while at the same time preserving the integrity of categorical phrases as syntactic/semantic units that may be moved in syntactic transformations.

There is a large body of empirical work in linguistics suggesting that many logical properties of quantifiers, names, definite and indefinite descriptions, and pronouns are best understood as involving such restricted quantificational logical forms. A number of philosophers, notably Montague (1974), in the context of a type-stratified set theory, Cocchieralla (1981), in the context of a type-free intensional logic of attributes, Evans (1985), and Neale (1990), have developed philosophical theories of this sort. Cocchieralla (1977) construes ordinary proper names as sortal common names, whose identity criteria single out at most one entity. Just as the referential concept which underlies the use of sortal common nouns or noun phrases are associated with certain identity criteria for identifying and reidentifying entities of a kind, so also do ordinary proper names come with certain identification criteria – namely, those provided (in a given context) by the most specific sortal concept associated with the name's introduction into discourse. Thus the proper name 'Ponce de Leon,' just as a categorical phrase 'some *S*,' is construed as involving the quantificational determiner 'some' and a common noun sortal *S*. In the case of a proper name, however, the sortal provides identity criteria for singling out at most one entity.

Cocchiarella (1989) employs his type-free intensional logic of attributes to represent referential concepts, be they for definite descriptions, indefinite descriptions or proper names, as properties of properties. In his logic, attributes (properties and relations) have both a predicable and an individual nature. They may have predicative occurrences or themselves be subjects of predication. The referential (predicable) occurrence of a referential concept for a definite description 'the S' is represented as '$(\exists^1 xS)$.' Using Church's lambda notation for properties, the referential concept '$(\exists^1 xS)$' is identified as the property $[\lambda\varphi(\exists x)(Sz \equiv_z z = x \,.\&.\,\varphi x)]$. The cognitive structure underlying an assertion such as 'the S is G' is perspicuous in the following:

$$[\lambda\varphi\ (\exists x)(Sz \equiv_z z = x \,.\&.\,\varphi x](G).$$

Here we see the referential concept occurs predicatively in the assertion. By lambda conversion, this is equivalent to the more usual

$$(\exists x)(Sz \equiv_z z = x \,.\&.\, Gx).$$

The occurrence of referential concepts (construed as properties of properties) as subjects of predication explains how, in the presence of intentional verbs, their ordinary referential use may be disengaged. Consider the following:

Ponce de Leon seeks the fountain of youth.

Russell's analysis of definite descriptions cannot account for the difference in any staightforward way. It would clearly not do to put:

$$(\exists x)(Yz \equiv_z x = z \,.\&.\, \text{Ponce de Leon seeks } x).$$

It does not follow from that fact that Ponce seeks the fountain of youth that there is such a fountain of youth that he seeks. On the other hand, if Ponce finds the fountain of youth, there is a fountain that he finds. The difference must, it seems, be grounded in a difference of logical form, surface grammatical form notwithstanding. But to find a secondary occurrence of the definite description, and so a difference in the logical form, a Russell's analysis would require a complicated reconstruction of the nature of the intentionality hidden in the verb 'to seek.' Cocchiarella's approach is to regard logical form as reflecting the cognitive structure of the assertion by providing an appropriate representation of the referential and predicable mental concepts that underlie the assertion. The definite description, 'the fountain of youth' and the proper name 'Ponce de Leon' correspond to referential concepts, which have restricted quantificational forms. The relation of seeking is intensional for its range but extensional for its domain. The structure of the assertion that 'Ponce de Leon seeks the fountain of youth' is:

$$\text{Seeks}\{(\exists x\text{Ponce}),(\exists^1 xY)\}.$$

Since 'seeks' is extensional in its domain, transformation of the referential concept '$(\exists x\text{Ponce})$' to a predicational (and thus referential) position is possible. Thus we get,

$(\exists x\text{Ponce})\text{Seeks}\{x, (\exists^1 y Y)\}$.

No such transformation is possible for expressions occurring in the second argument place of the relation sign. On the other hand, in the case of

$\text{Finds}\{(\exists x\text{Ponce}),(\exists^1 y Y)\}$.

Transformations are possible for both the domain and the range because 'finds' is extensional in both occurrences. Thus the above is equivalent to:

$(\exists x\text{Ponce})(\exists^1 y Y)(\text{Finds}\{x,y\})$.

The difference between the domain and the range of the intentional relation 'seeks' is made more manifest if we take the following, which differs from Cocchiarella's account because it appeals to a partial analysis of the relation:

$(\exists x\text{Ponce})(\exists m)(\text{Mental-state-of-seeking}(m) \,\&\, \text{Has}(x,m) \,\&\, (\exists \psi)(\text{In}\{(\exists^1 y Y)\psi y, m\}))$

Ponce obviously is not seeking a property. On Cocchiarella's analysis it is a referential concept of a fountain of youth (represented as a certain property of properties) that Ponce uses in the mental acts he employs in seeking.

By assimilating ordinary proper names to sortal quantifiers, a similar construction may be employed for examples such as 'Caesar worshipped Jupiter.' Cocchiarella has:

$(\exists x\text{Caesar})\text{Worships}\{x, (\exists^1 y\text{Jupiter})\}$.

Using our technique of a partial phrarphrase, we have:

$(\exists x\text{Caesar})(\exists m)(\text{Mental-state-of-worshipping}(m) \,\&\, \text{Has}(x,m) \,\&\, (\exists \psi)(\text{In}\{(\exists^1 y\text{Jupiter})\psi y, m\}))$.

As before, because of the intentional nature of the relation at its range, it does not follow from the fact that Caesar worshipped Jupiter that there exists some entity Jupiter that Caesar worshipped. Cocchiarella's techniques have particularly useful applications to the phenomena of anaphora. Consider the following difficult case:

Hob thinks some witch is afoot, and Nob wonders whether she (that witch) is evil.

The problem is to explain how the pronoun 'she' in the second clause is bound to the quantifier 'some witch' in the first. Obviously, the following will not do

$(\exists^1 y W)\{(\exists x\text{Hob})\text{Thinks}(x, \text{Afoot}(y)) \,\&\, (\exists z\text{Nob})\text{Wonders}(z, \text{Evil}(y))\}$.

This introduces an ontological commitment to witches. By appealing to the referential concept employed by both Hob and Nob, we can begin to see how to go about solving the puzzle. I shall not spell out Cocchiarella's complete solution here, but only suggest

the direction. Where '($\exists^1$y Wy & (∃xHob)Thinks-About(x,y))' represents the referential concept 'the witch that Hob thinks about,' we have:

(∃xHob)Thinks{x, (∃yW)Afoot(y)} & (∃xNob)Wonders{x, ($\exists^1$y Wy & (∃xHob)Thinks-About(x,y)) Evil(y))}.

Referential concepts seem to play an important role in recovering the conceptual structure of the assertion.

As we are beginning to see, questions about the nature of reference and logical form are entangled with the many issues in philosophical linguistics and cognate fields such as the philosophy of mind and cognition. Indeed, in many cases the philosophy of language is being altogether subsumed by philosophy of mind. Classical cognitivism posits syntactically structured symbolic representations and defines its computational, rule-based, operations so as to apply to such representations in virture of their syntactic structures. Cognition is computational, and computations are defined over symbols (representations) encoded into data structures that can be stored, moved, retrieved, and manipulated according to a recursive set of rules. The representations of cognitive structures offered by logical languages (and their formal semantics) have a lot to offer here. By appeal to such represention, many standard problems (e.g. the incompleteness and non-monotonicity of reasoning, the frame problem, etc.) are tamable. The representation of logical form has important applications as an analytic tool; it offers a formalization of knowledge-representation, and a model of reasoning. Indeed, it can also be used as part of a programming language (e.g. Prolog). A computational model offers a formal analysis of the sentences of natural language – a theory of logical form that renders a perspicuous logical representation of the truth-conditions determined by the content of those sentences. In this way, cognitive models based on logical form serve to guide and test general arguments concerning the nature of cognitive processes.

Formal logic has been a very attractive tool for traditional models in cognitive science, but we must not neglect the fact that there are new models of cognition that employ connectionist (parallel *distributed* processing), and many of these are at the forefront of recent research. Connectionist ('non-representational') architectures have been found to enjoy success where classical cognitivism is weakest – *viz.* in modeling perceptual tasks such as face recognition, speech processing, and visual discrimination. Connectionist models forgo decompositional recursive architectures; the contribution of individual component units are minimized and the behavior of the system results from the strength and kinds of interactions between the components rather than from a recursive rule-governed process of manipulation of units. There are no fixed representations upon which operations are performed. Instead, there are activated units which function to increase or decrease the activation patterns of other units until a stable configuration is reached. On such models, the notions of 'reference,' 'representation,' 'proposition,' 'belief,' and even 'truth' take on a new naturalized meaning. A connectionist system does not rely upon internal representations as its processing units, and it does not need to represent all relevant aspects of its environment. It 'tunes itself' to its environment without operating on syntactically encoded representations retrieved from memory. On a connectionist model, mental states do not have a combinatorial or structural semantics – the content of complex units is not determined, in

some recursive way, by the content of their more simple parts. The complex behavior of the system emerges in a way that is not built up piecemeal from operations at the next lower level. Connectionist models are flexible and can respond to deformed inputs or new inputs without supplement of new rules and new stored data. The performance of the system degrades smoothly when parts are destroyed or overloaded, settling in spite of the adversity into a state of equilibrium.

Now Fodor and Pylyshyn (1988), among others, have pointed out that features of cognition that are involved in problem solving and reasoning are precisely the sort of features that connectionist architectures find most difficult to model. Verbal behavior is paradigmatic of structured combinatorial semantics, for it seems to require that complex verbal forms be syntactically composed of recurring units. When one understands an utterance of a sentence, one constructs a mental representation – a *parsing tree* which displays the semantic content (truth-conditions) of the whole complex as a function of the semantic content of its syntactically more simple parts. Psycholinguistic theories differ in the nature of such trees and in how they are composed, but in all such theories quantificational logical forms play a central role. Speakers of a language can effectively determine the meaning or meanings of an arbitrary expression, and it is the central task of a linguistic theory to show how this is possible. On mastering a finite vocabulary and sets of rules, we are able to produce and understand a potentially infinite number of sentence types. This seems impossible to explain without the postulation of semantically structured representations. If one adopts a computational theory of cognition, Fodor (1987) argues, then one must be prepared to accept that the transformational grammar of current cognitive science is empirically well-corroborated, and that this speaks in favor of an ontology of structured mental representations – a *language of thought*.

The connectionist admits that perceptual (e.g. auditory and visual) experiences associated with parsing the contents of utterances must be explained, but denies that the actual cognitive architecture (the neural networks) which make linguistic understanding and communication possible has anything to do with their realizing structured mental representations. Given Church's Thesis that the imprecise notion of 'computation' be identified with the rigorous notion of 'recursiveness,' traditional and connectionist architectures will be able to emulate the behavior of one another. The debate between traditional cognitive science (as a computational account of mind) and connectionism is properly a debate about which research program is more likely to render a naturalistic (causal/evolutionary) explanation of how neurons actually give rise to human consciousness and animal cognition. There is a danger, therefore, in aligning the Russellian notion of logical form, and the Description Theory of Reference, too closely with philosophical linguistics and cognitive science. Science may, in the end, find that best account of language apprehension and cognition rejects structured mental representations and transformational grammar.

3 Rigid Designators

As Russell saw matters, a good many metaphysical theories are generated from a failure to properly analyze logical form. Unfortunately, the Russellian emphasis in analytic phi-

losophy on logical form fell out of fashion with the collapse of the Frege/Russell logicist program and the logical empiricism it spawned. This is writ large in modern modal logic, with its semantics of possible worlds, entities *existing* at one world and not another, and its ascriptions *de re* of essential properties. The new essentialist modal logic has been a rich resource for those who challenge the Description Theory, and the Russellian notion of logical form itself.

In the context of modal ascriptions, the law of identity, together with innocuous looking assumptions, can yield startling results. Let us represent necessity by $\Box$ and possibility by $\Diamond$. Assuming that

$$(x)\ \Box(x = x),$$

that is that every entity is necessarily self-identical, one can derive:

$$(x)(y)(x = y \supset \Box(x = y)).$$

quite straightforwardly from the law of identity. Now if proper names are genuine singular terms, then by *universal instantiation*, we arrive at:

$$\text{Hesperus} = \text{Phosphorus} \supset \Box(\text{Hesperus} = \text{Phosphorus}).$$

This seems astonishing. By astronomical investigation *a posteriori* we come to discover that Hesperus is identical with Phosphorus, and yet from the above this yields knowledge of a necessity! Worse, with definite descriptions construed as singular terms, *universal instantiation* would seem to yield:

$$\text{The morning star} = \text{the evening star} \supset \Box(\text{the morning star} = \text{the evening star}).$$

Yet surely the morning star is contingently identical with the evening star. It may have turned out that they were not both the planet Venus.

Russell's theory of definite descriptions offers an explanation. Proper names are to be transcribed in symbolic logic as definite descriptions, and definite descriptions are 'incomplete symbols' to be contextually defined. *Universal instantiation* does not apply to definite descriptions, for they are not genuine terms of the formal language. On Russell's theory, one has:

$$E!(\iota z)(Az) :\supset: (x)Bx \supset (\iota z)(Az)[B(\iota z)(Az)].$$

Accordingly, since we have $E!(\iota z)(Mz)$ and $E!(\iota z)(Ez)$, *Universal instantiation* yields

$$(\iota z)(Mz)\ (\iota z)(Ez)[(\iota z)(Mz) = (\iota z)(Ez)\ .\supset.\ \Box\{(\iota z)(Mz) = (\iota z)(Ez)\}\ ].$$

Eliminating the descriptions, this is:

$$(\exists x)(Mz \equiv_z z = x\ .\&.\ (\exists y)(Ez \equiv_z z = y\ :\&:\ x = y \supset \Box(x = y))).$$

The result now appears innocuous.

In a now famous argument, however, Quine attempted to show that singular expressions embedded in the context of necessity are non-referential, and quantified modal logic is illicit. The context of necessity is, as Quine puts it, 'referentially opaque.' Quine observed that if '9' in the true statement

(4) $\Box(9 > 7)$,

refers to the number 9, then by the law of identity it may be replaced by the singular expression 'the number of planets' without loss of truth value. But of course such a replacement does alter the truth value, for

"$\Box$(the number of planets > 7)"

is false. In Quine's view, the failure of the substitutivity of the co-referrential expressions '9' and 'the number of planets' in the context of necessity shows that the name '9' in the expression '$\Box(9 > 7)$' is an orthographical accident like the 'nine' as it occurs in 'Quinine water is therapeutic' (Quine 1976). Quite obviously, 'nine' does not refer to the number 9 in such an occurrence. The context is a referentially opaque with respect to the occurrence of the expression 'nine.' It would be improper to form the context

'Qui(x) is therapeutic'

and permit the variable x to then be bound by a quantifier. Similarly, Quine maintains that the expression "$\Box(x > 7)$" is ill-formed.

With the help of the scope distinctions afforded by Russell's theory of definite descriptions, Smullyan (1948) shows how to maintain, in spite of Quine's argument, that the occurrence of '9' in '$\Box(9 > 7)$' is referential and refers to the number 9. Let 'Px' represent 'x numbers the planets.' Then the statement '$\Box$(the number of planets > 7),' that is,

(5) $\Box((\iota x)(Px) > 7)$

is ambiguous between the following:

(5a) $\Box(\exists x)(Pz \equiv_z z = x \,.\&.\, x > 7)$
(5b) $(\exists x)(Pz \equiv_z z = x \,.\&.\, \Box(x > 7))$.

Sentence (5a) is false, but sentence (5b) is true and provable from (4) by the law of identity. There is a number that contingently numbers the planets and it is necessarily greater than the number 7.

Quine, of course, was no stranger to the apparatus of Russell's theory of definite descriptions, and he certainly would have anticipated Smullyan's use of the theory against his argument. So it might at first appear perplexing why Quine had not realized that his argument for the referential opacity of the context of necessity could be undermined by Russell's theory. But it must be understood that the source of Quine's objection to quantifying into the context of necessity lies in his empiricist conviction

that the only necessity is logical necessity and that logical necessity is 'truth in virtue of logical form.' The legacy of Frege and Russell is to have replaced the early empiricist notion of 'truth in virtue of meaning' with the more refined notion of 'truth in virtue of logical form (generated by the logical particles alone).' This is fundamentally a *de dicto* notion whose semantics is rendered most straightforwardly in a Tarski-style formal semantic account of logical truth. To embrace a Russellian approach to the failure of the substitutivity of co-referentials in the context of necessity, as Smullyan does, one has to allow expressions of *de re* necessity. Orthodox empiricism is deeply troubled by *de re* ascriptions which ground necessity in the metaphysical essential natures of entities and not in the form of propositions. The intelligibility of *de re* ascriptions of necessity required by a Russellian analysis, would, as Quine puts it, require the metaphysical tangles of an Aristotelian Essentialism.

If necessity is to be understood as fundamentally an anti-essentialist notion of form, then Smullyan's employment of Russell's theory of definite descriptions will be of little help as a response to Quine's argument for the referential opacity of contexts of necessity and the illegitimacy of quantified modal logic. Nonetheless, Quine is mistaken in thinking that quantified modal logic is committed to any form of essentialist notion of necessity. In the Kripke-style semantics for quantified modal logic, there is no assurance that for every admissible extension of the predicate letters of a formula in a domain (of a Tarski-style semantics for logical truth), there is a possible world in which just those entities of the domain satisfy the predicate (Kripke 1963). For instance, where F is a predicate letter of the language, there will be Kripke models in which an essential sentence such as,

$(\exists x)\Box Fx$

is true. This can only be so if no possible world in the model is a world where nothing has F. The interpretation which assigns the empty-class to F has been left out. Parsons (1969) points out, however, that even in a Kripke-style semantics for quantified modal logic, with its different entities in different worlds, some among the models will be 'maximal models' in which for each admissible extension of the predicate letters there is a possible world where just those entities in the extension satisfy the predicate. Accordingly, since Kripke's notion of *universal validity* is understood as invariant truth in every possible world of every model, no essentialist sentence will be *universally valid*. Though some essentialist sentences will be true in a given model, no essential sentence will be a *thesis* of a sound axiomatization of quantified modal logic. In a *universally valid* formula, the only properties that will be necessarily possessed by entities are purely logical properties such as '$[\lambda x\, Fx \supset Fx]$.'

Cocchiarella (1975) goes even further. Kripke's notion of *universal validity* arbitrarily omits some logically possible worlds. At first blush this is easy to miss, for Kripke's notion of *universal validity* is defined in terms of *every* possible world of *every* model. But as we saw, the Kripke semantics does not measure what counts as a 'possible' world in terms of the Tarski semantic conception of an admissible interpretation over a domain. If necessity in quantified modal logic is to be interpreted as logical necessity (in such a way that it coincides with the Tarski-style semantics of logical truth), then one must adopt a 'primary semantics' for quantified modal logic in which *every* model is *maximal*.

The Kripke semantics is a 'secondary semantics' for necessity because it omits some logically possible worlds. The differences are striking. For instance, Cocchiarella shows that monadic modal logic is decidable in its primary semantics. It is undecidable, as Kripke has demonstrated, in the secondary semantics. Moreover, Cocchiarella shows that in its primary semantics modal logic is semantically incomplete. (Its logical truths coincide with those of second-order logic, which is known to be semantically incomplete.) In Kripke's secondary semantics, modal logic is semantically complete. Quine's empiricist objections to *de re* ascriptions of necessity are assuaged in the 'primary semantics.' In such a semantics, each *de re* ascription is semantically equivalent to some *de dicto* ascription (McKay 1975). In the primary semantics for quantified modal logic, logical necessity is a formal notion – truth in virtue of form and not truth in terms of the metaphysical essences of entities. Smullyan's employment of the Russellian approach to proper names and definite descriptions in quantified modal logic does not, therefore, require any essentialist statement to be true.

Metaphysicians who agree that logical necessity should coincide with the semantic conception of logical truth may, nonetheless, wish to reject the empiricist cannon that the only necessity is logical necessity. They may wish to embrace a causal/physical form of necessity. Ordinary language is rich with *de re* essentialist statements of this sort. Indeed, such framework seems embedded in ordinary biological taxonomies based upon genus and species. If there are natural kinds, then there is a form of causal/physcial essentialism. As Cocchiarella (1984) points out, it is not the primary semantics for logical necessity that would be appropriate in such contexts, but rather the Kripke-style secondary semantics of 'metaphysical' necessity (to use Kripke's expression). Kripke's metaphysical necessity would then be interpreted as causal necessity.

To semantically underwrite *de re* ascriptions of metaphysical necessity, Kripke and Putnam have argued that mass terms for substances like 'water,' and 'gold,' and terms for biological kinds like 'horse,' 'cat,' and 'lemon,' are rigid designators, properly understood in terms of a causal theory of reference. Putnam developed a causal theory of kind terms extensively (Putnam 1975). Natural kind terms are associated with sortal concepts. But how precisely does the sortal concept direct the classification of entities as being of the same kind? An account that seeks to specify the concept of, say, 'gold' by a description in terms of the manifest properties, relations, and appearances, will be quite unsatisfactory. For example, early users of the natural kind word 'gold' could not distinguish it from 'fool's gold' (chalcopyrite). Not only do such accounts often fail to provide necessary and sufficient conditions for being of the kind in question, they leave wholly unexplained how it is that scientific categorizations have evolved. Newton's conception of mass, for example, was quite different from that of Einstein, for central among the manifest attributes he associated with the concept was that mass cannot be altered by acceleration. A descriptivist approach threatens to leave the history of science as irrealist and non-convergent, with new scientific theories changing the very meanings of the fundamental terms of the old. Putnam is concerned to protect convergent scientific realism.

Putting aside for the moment Kripke's conception of the philosophical underpinnings of his secondary semantics for metaphysical necessity, Putnam's employment of a causal theory of reference to underwrite convergent scientific realism can be interpreted as Cocchiarella suggests – *viz.* as an interpretation which takes Kripke's meta-

physical necessity to be causal/physical necessity. On Putnam's causal account of reference, an entity x is (an) f (horse, birch tree, orange, gold, etc.) if and only if, given *good* exemplars of f, the most explanatory and comprehensive true theoretical account of the causal structure of the exemplars would group *x* alongside these exemplars. Whether something is an f turns on the causal structure of the word; it is a matter of whether it bears the relation 'same f as,' construed as a cross-*causally*-possible world relation, to the *good* exemplars. Putnam's semantics for natural kind world will accommodate the fact that one and the same concept of what it is to be an f would be unfolded gradually in a succession of different and improving scientific conceptions of the 'same f as' relation. The theory can explain how it is that what have appeared astonishingly like fs (and may even have been thought to be among the exemplars of fs) turn out not to be fs, and that because hidden structures dominate appearances, it explains how the most improbable seeming specimens may in fact turn out to be fs.

We know *a posteriori* that water is 'necessarily' H_2O. It cannot *causally* have been otherwise. The necessity here applies not to *identity* as a logical relation, but to the relation of 'sameness of causal structure.' A substance *x* is water if and only if it, in fact, bears the trans-world relation of *sameness of causal structure* to the particular exemplars of the substance we call 'water' in the actual world. Given that water is, in fact, H_2O, nothing counts as a causally possible world in which water does not have that structure. Use of a natural kind term f is not understood in terms of some set of ideas or concepts (intensions) associated with the term which supposedly determine its extension, but rather by fixing on certain actual exemplars of substances that are thought to have a common nature in being f's. Psychological states of linguistic speakers, concepts, ideas, images, and the like, which are associated with use of the term 'water' do not determine the extension of the term. What determines the extension is the actual chemical structure of water itself. In this use, natural kind terms such as 'water' behave like the indexicals 'I', 'now,' 'this,' 'that.' Their use is explained by *pragmatics*, not semantics. That is, the reference of such terms is fixed by causal relations that are external to the concepts employed by speakers of the language. In Putnam's view, natural kind predicates no not have sense.

Natural kind terms function as 'rigid designators,' indexically picking out the same substances in all possible worlds in which they exist. The indexicality of such terms, as they are often used in natural language, manifests itself in modal and counterfactual contexts. The following is a true sentence:

> It might (causally) have been the case that the substance that has all the manifest properties and relations and appearances of water is not water but a substance XYZ.

The term 'water' in this sentence, is not synonymous with any evolving cluster of descriptive information of manifest properties, relations, and appearances of water which would purport to single out the extension of the term. This is the possible world Putnam famously has called 'twin Earth' – a metaphysically (causally) possible world in which what satisfies all that would be part the best descriptivist account of the meaning of the term "water" (say before the chemical composition of water was known), is nonetheless not water. Indeed, even the cluster of description information

that included 'substance whose chemical composition is H_20' is not determinative of the extension of the kind term 'water.' For it is causally possible that modern physical chemistry is incomplete, and that water turns out not to be a substance whose chemical composition is simply H_20, but some more complicated molecule.

The extension of a natural kind predicate is not given by the descriptive information associated with the predicate as it is used within a community. Something is water if and only if it, in fact, has the same causal nature as the good exemplars of water. This is so regardless of whether members of a linguistic community who use the term 'water' know what that causal structure is, and regardless of their conception of what water is. The extension of a natural kind term (if it is known at all) may be known only to a small community of scientific experts. According to Putnam's thesis of the 'division of linguistic labor,' the criteria of application of a natural kind term may be known only to experts and every one else who acquires the kind term, and uses it indexically, implicitly defers to the experts regarding its application. The pay-off that Putnam hopes to obtain is the revitalization of convergent scientific realism. To return to our earlier example, Newton was talking about mass because the *good* exemplars of the phenomena of body's having mass have the same causal nature as those studied by Einstein. But, we shall have to willingly accept that Newton's concept of mass (including the laws he thought constitutive of the notion) did not determine its extension. Meaning, in the sense of what it is that determines extension, is not matter of concepts in the minds of speakers of a language.

The fact that we do use scientific kind terms such as 'water,' 'H_20,' 'Hydrogen,' 'electron,' and the like indexically, leaving it to the world's causal structure to fix extension, seems no great surprise. The sting comes only if the causal theory is parlayed into a *general* theory of reference – an externalist theory of the content of cognitive states. It need not be so. In his discussion of 'egocentric particulars,' Russell himself admitted that indexicals (he reduced them all to expressions involving the word 'this') fix reference via a causal chain: "the shortest possible chain from a stimulus outside the brain to a verbal response" (Russell 1966: 112). But he emphatically asserted that no egocentric particulars are needed in a scientific account of the world. In his effort to defend convergent scientific realism, Putnam may disagree, but the jury is out. Indeed, Laudan (1981) has argued convincingly that Putnam's causal theory of reference does not best serve scientific realism. In any event, accepting the pragmatic fact that a natural kind term may be used indexically (so that its extension is fixed externally and contextually), does not by itself undermine the Description Theory's role in computational accounts human thought and cognition, or in accounts of the compositional semantic structures (such as those of a Chomsky-style transformational grammar).

There are, however, far more unruly notions of metaphysical necessity that Kripke's secondary-semantics allows, and providing a viable semantics for these seems to call for a thorough rejection of the Description Theory. With the collapse of logicism, the nature of mathematical truth has remained a mystery, and its statements seem to be prime examples of a new form of metaphysical essentialism about numbers. The number 9 is necessarily odd. Goldbach's conjecture that every whole even number greater than 2 is the sum of exactly two primes, if true, is a necessary truth. But even more radically, Kripke's secondary semantics opens the door to a unique form of *de re* essentialism that is closer to a logical notion than a causal one, and yet it is a concep-

tion of necessity that is based on neither the notion of truth in virtue of ontological structure nor the Tarski-style semantics for logical truth. An Aristotelian essentialism with respect to a conception of causal (physical) necessity is problematic enough for empiricism. But this sort of *logico-metaphysical* necessity seems beyond the pale.

In embracing *de re* metaphysical necessity of this extreme form, as opposed to orthodox empiricism's conception of necessity as 'truth-in-virtue of form', Kripke is driven to his anti-Russellian position that ordinary proper names are rigid designators (Kripke 1971). An individual constant 'a' is a rigid designator if and only if $(\exists x)\,\Box(x = a)$. That is, it designates the very same entity in every possible world. The use of ordinary proper names in modal and counterfactual contexts reveals that they are rigid designators, not definite descriptions. Consider the sentence,

> The most famous among philosophers of antiquity might not have been most famous among philosophers of antiquity.

On one reading, this sentence is true. Aristotle was indeed most famous among philosophers of antiquity, but he might not have been. On another reading, it is logically contradictory. Whoever was most famous among philosophers of antiquity was certainly most famous among those philosophers. By syntactically signaling the presence of descriptive semantic information, definite descriptions induce ambiguities of scope in modal and counterfactual contexts. The use of ordinary proper names on the contrary, relies upon the context of utterance to secure reference. In virtue of this, ordinary proper names in modal and counterfactual contexts do not produce scope ambiguities. They are not, therefore, synonymous with any definite description.

The Description Theory of reference in natural language offers a theory according to which the sense of an ordinary proper name is the sense of some definite description, and accordingly the name refers to whatever satisfies the description. Ordinary proper names, Kripke admits, are introduced into a language by means of a reference fixing definite description, but he denies that this fixes the sense (meaning) of the proper name. The descriptive apparatus initially employed to fix the reference of a name does not, in general, continue to fix its reference in all further uses of the name. Taking an example from Evans (1973), observe that

> "It was Elhannan (of 2 *Samuel* 21:19) and not Goliath, who was the Philistine giant slayed by David,"

is possibly true. Indeed, there is now significant historical evidence that it is in fact true. But this certainly does not lead us to say that the name 'Goliath' refers, after all, to Elhannan. The name 'Goliath' refers to the same person, irrespectively of whether or not David slayed him. It cannot, therefore, be synonymous with a description such as 'the Philistine giant slayed by David.'

Proper names are rigid, descriptions (except when the descriptive properties are essential properties) are not rigid. Kripke explains the rigidity of ordinary proper names by appeal to a causal theory of reference. The reference of a name is made rigid by the existence of a certain reference-preserving causal/historical chain leading back to an entity, and not by the fact that the referent satisfies a set of descriptive information asso-

ciated with the sense of the name. Evans gives a succinct characterization: A speaker, using a name N on a particular occasion, will denote some item x if there is an appropriate causal chain of reference-preserving links leading back from his use on that occasion ultimately to the item x itself being involved in a name-acquiring transaction such as an explicit dubbing (Evans 1973). Kripke denies that even an evolving and amendable cluster of descriptive identification criteria, some but not all of which must be satisfied by the name, can serve in a semantic theory of proper names. This parallels an interesting result in formal semantics. No axiomatic first-order theory can fix its interpretation. Indeed, according to the Löwenheim-Skolem theorem, any first-order axiomatic theory with identity that has an infinite model, has a denumerable normal model (where the identity sign is interpreted as identity) in the natural numbers. No matter what new axioms are added to try to delimit the referents of its terms, it remains that there are unintended interpretations that satisfy all the axioms. Similarly, no matter what cluster of descriptive information is chosen to supplant a proper name, one can always find an epistemically plausible situation in which the referent of the proper name does not satisfy the descriptive information. It simply will not work for a Description Theory to attempt to subsume the causal theory of reference by forming a definite description which characterizes the relevant reference-preserving causal chain associated with the use of the proper name. Such a description, like adding more axioms to a first-order theory, cannot fix an intended interpretation. It is the world – the causal chain itself – that fixes reference, not satisfaction of any description. Kripke concludes that long ago Mill had matters right: ordinary proper names do not have sense.

Working out precisely what is required of a causal chain that it be 'appropriate' proves difficult. There can be troublesome cases of branching and deviant chains, and of course the familiar problem explaining the use of fictional names. The theory also faces very serious problems as to how to explain the failure of substitution of co-referential proper names in the contexts of propositional attitudes. In fact, Kripke himself has generated a new 'puzzle about belief' involving translation and disquotation that arises in such contexts (Kripke 1976). But we shall not be concerned with these details. As we have seen, the distinction between semantics and pragmatics may be exploited to come to the rescue of a Description Theory. The Description Theory was originally intended as a purely semantic theory, not a pragmatic (*cum* semantic) theory of speaker's reference or communication. Indeed, while the causal theory of names often secures the right references for ordinary proper names in modal and counterfactual contexts, Evans (1973) has pointed out that it too fails to fully appreciate the extent that determination of reference is contextual and pragmatic. He illustrates the point by discussing an example from E. K. Chambers's *Arthur of Britain*. Arthur, it seems, had a son Anir whom legend has perhaps confused with his burial place. Evans writes: "If Kripke's notion of reference fixing is such that those who said Anir was a burial place of Arthur might be denoting a person, it seems that it has little to commend it." The causal theory must accept that there can be cases of reference shifting. Accordingly, Evans offers a hybrid theory of what is required for and expression to be a genuine proper name. In general, he says, a speaker intends to refer to the item that is the dominant causal source of his associated body of descriptive information.

The sources of Kripke's rigid designators and his objections to the Description Theory are, however, quite different than the pragmatic considerations of Strawson,

Donnellan, and the like, who rightly find the Description Theory to be an inadequate account of how the names are actually used in communication. A Russellian might simply acknowledge that a rigid pragmatic use of an ordinary proper name demands that when transcribing modal sentences, the definite description chosen to replace the ordinary proper name must always be rendered with primary scope. The source of Kripke's objections to the Description Theory lie in his advocacy of a secondary semantics for a *logico-metaphysical* necessity, where worlds may well have greater or fewer entities than there are in the actual world. Indeed, if one were to take such worlds realistically, a singular term may refer rigidly to an entity that is not actual. A primary occurrence of a definite description will always refer (if it refers at all) to an actual entity.

Russell's view that ordinary proper names are 'disguised definite descriptions' was the result of his quest to find logical structure where surface grammatical structure had none. By doing such a conceptual analysis he thought philosophy could free itself from what he regarded as muddles of metaphysics. On Kripke's view, *de re* metaphysically necessary truths are not to be construed as conceptual truths of form or meaning, and philosophy is not to be regarded as a discipline engaged in *conceptual* analysis. Philosophy is engaged in discovering (at times *a posteriori*) *de re logico-metaphysical* essences, just as the science of natural kinds is involved in the empirical discovery of causal structures underlying substances. There is, therefore, on this conception of philosophy, no need to follow Russell in searching for logical forms (logical structures) obscured by surface grammatical forms in an effort to explain away metaphysical necessity. If we take Kripke's *de re* metaphysical necessity seriously, then we should be prepared to reject a quantificational account of the logical form of statements involving proper names. We should be prepared to reject the Russellian quest for logical form altogether, and be content to say that proper names act as if indexicals, rigidly picking out their referents because of the metaphysical nature of the world and independently of any speaker's descriptive information. To return to the example that began this section, from,

(6) $(x)(y)(x = y \supset \Box(x = y))$,

and the astronomical discovery that the morning star = the evening star, one may *not* conclude (by *universal instantiation*, and *modus ponens*) that,

$\Box$(the morning star = the evening star).

In the contexts of Kripke's secondary semantics for metaphysical necessity, the axiom of *universal instantiation* undergoes modification. The system does not allow universal instantiation to definite descriptions. Thus the acceptance of (6) does not rule out *de dicto* identity statements that are contingently true *in virtue of their form*. From (6) we shall only be able to arrive at:

$(\exists x)\,\Box(x = a) \,\&\, (\exists x)\,\Box(x = b) .\supset. a = b \supset \Box(a = b)$,

None the less, from (6) and the astronomical discovery that Hesperus = Phosphorus, we saw that Kripke arrives *a posteriori* at the following:

$\Box$(Hesperus = Phosphorus).

'Hesperus' and 'Phosphorus' are to be read transparently in virtue of their being rigid designators. The logical form of this statement is not quantificational. We have *de re* metaphysical (*logical*) necessity, and not a necessity grounded in propositional form.

4 Russell on Logical Form

We have come full circle. We argued that Russell's theory of definite descriptions can defend itself against the sort of objections voiced by Strawson, Donnellan, and the like, by carefully distinguishing issues that pertain to pragmatics from those that are relevant to combinatorial semantics. This, however, lends itself to too narrow a construal of Russell's notion of logical form – aligning it with the transformational grammars of contemporary philosophical linguistics. Moreover, we saw that the deep source of Kripke's objections to the Description Theory are not to be found in appeal to pragmatic features of reference. They lie in his advocacy of a secondary semantics for a *de re* and metaphysical necessity. In this regard, it is interesting to return to Russell's own conception of the paradigm for a new scientific philosophy that is exemplified by his 1905 theory of definite descriptions.

Russell, as Frege before him, embraced a conception of logic that is quite different from the contemporary. Logic is not the mathematical study of formal systems, their semantic completeness, consistency, and the like. Logic does not have as its main goal the investigation of the combinatorial semantic notion of logical consequence (the conditions of truth-preservation in inference so elegantly captured in a Tarski-style formal semantics). For Russell, logic is a general science of ontological structure. The psycholinguistic semantic structures postulated by philosophical linguistics and cognitive science in their efforts to underwrite truth-preservation in inference will be included, but the science of logic is not dependent upon any particular theory of language learning, meaning, or cognition. On the conception of logic that Russell held while advancing the 'misleading form thesis' of his theory of definite descriptions, logical analysis is ontological analysis.

If we look at examples of analytic work on logical form that Russell endorsed, we will be immediately struck by what is included. Russell took his program for a scientific philosophy based on the analysis of logical form to be exemplified by the achievements of mathematicians, such as Frege on the notion of cardinal number, Cantor on infinity and continuity, Dedekind on the notion of irrationals, and Weierstrass on the notion of the 'limit' of a function (Russell 1901). Their studies eventuated in new logical analyses of these notions. In Russell's view, Cantor's work on the transfinite put to rest centuries of speculative metaphysics surrounding the 'infinite' and the notion of 'continuity.' Russell writes: "Continuity had been, until he [Cantor] defined it, a vague word, convenient for philosophers like Hegel, who wished to introduce metaphysical muddles into mathematics. . . . By this means a great deal of mysticism, such as that of Bergson, was rendered inadequate" (Russell 1946: 829). With Cantor, the former notion of continuity which seemed impossible to render by any notion of magnitude, depends only on the notion of *order*. The new constructions arithmetizing Analysis revealed that it is order, not magnitude, that is basic to continuity. The *derivative* and the *integral* became, through the new definitions of 'number' and 'limit,' not *quantitative* but

ordinal concepts. Continuity lies in the fact that some sets of discrete units form a dense compact set. "Quantity," wrote Russell, ". . . has lost the mathematical importance which it used to possess, owing to the fact that most theorems concerning it can be generalized so as to become theorems concerning order" (Russell 1946: 829). Weierstrass had banished the use of infinitesimals in the calculus. He showed that the notion of the 'limit' of a function which used to be understood in terms of quantity, as a number to which other numbers in a series generated by the function approximate as nearly as one pleases, should be replaced by a quite different *ordinal* notion.

Naturally, Frege's analysis of the notion of cardinal number is an important example of logical form, and Russell heralds it as "the first complete example" of "the logical-analytic method in philosophy" (Russell 1969: 7). But we do well to observe that Russell also included Einstein on space–time, as an example of work that revealed logical form. "Physics," Russell tells us, "as well as mathematics, has supplied material for the philosophy of philosophical analysis. . . . What is important to the philosopher in the theory of relativity is the substitution of space–time for space and time." With respect to quantum theory, Russell continues, "I suspect that it will demand even more radical departures from the traditional doctrine of space and time than those demanded by the theory of relativity" (Russell 1946: 832).

Looking at the examples that Russell took to be paradigmatic of work towards a theory of *logical form* a new perspective emerges. The interpretative tradition is misguided when it maintains that for Russell a theory of logical form renders an account of compositional semantic structures – a 'meaning analysis' which reveals that the structures that underlie cognition may be hidden in the misleading surface grammar of statements. Quite clearly the analyses offered in the work of Weierstrass, Cantor, Frege, and Einstein are not accounts of the psycho-linguistic structures grounding assertions involving notions such as 'limit,' 'continuity,' 'natural numbers,' or 'space,' and 'time.' They offer analyses that are quite different from the ordinary language meanings of such notions. The fundamental idea underlying Russell's science of logical form is not properly characterized as one of a meaning analysis (or semantics) of a statement; it is rather that of an eliminativistic ontological analysis.

For example, eighteenth- and nineteenth-century physics and chemistry offered a number of subtle fluid and aether theories that were highly successful at explaining a wide variety of phenomena. In the process of theory change, the research programs that gave rise to such theories were supplanted by atomistic physical theories couched within a new research program. Empirical and conceptual problems pertaining to the aether (such as its elasticity) were dropped, and an entirely new research program, with a new language and a new set of empirical and conceptual techniques, was inaugurated. Many successes of the earlier aether theories were retained by the theories of the new research program. Retention, however, is only partial; the confirmed predictions of an earlier theory in a rival research tradition do not always survive into the supplanting research tradition. Indeed, theoretical processes and mechanisms of earlier theories are at times treated as flotsam (Laudan 1977). The supplanting tradition may come to regard the terms of the earlier theories as non-referential, or regard earlier ontologies as idle wheels that serve no explanatory purpose. This is precisely how Russell viewed philosophy as a quest for *logical form*.

Russell's own work on logical form illustrates the method. His substitutional theory of propositions, which plies his 1905 theory of definite descriptions toward a solution of the paradoxes plaguing logicism, showed that a type-stratified theory of attributes powerful enough to generate arithmetic can be proxied within a type-free 'no-classes' and 'no-propositional functions' theory of propositions. Russell knew that a type-stratified theory of attributes in intension ('propositional functions') would block his paradox of predication – the paradox of the property P that a property exemplifies if and only if it does not exemplify itself. But Russell held that any calculus for the science of logic must adopt only one style of variables – individual/entity variables. The Russell Paradox of the attribute which an attribute exemplifies just when it does not exemplify itself, and the analogous paradox of the class of all classes not members of themselves, were solved by Russell's substitutional theory. The theory succeeds in finding a logical construction which builds the type distinctions that dismantle the paradoxes into the formal grammar of a 'no-classes' and 'no-propositional-functions' theory of propositional structure. The type-stratified language of attributes can be proxied in the type-free grammar of the calculus for the logic of propositions. In this way, Russell hoped to recover Logicism.

Russell's substitutional theory shows that a type-stratified theory of attributes powerful enough to generate arithmetic, can be proxied within a type-free theory of propositions. In the substitutional theory, the type-stratified language and ontology of attributes in intension (and the contextual definition of class expressions set within) is to be supplanted by the type-free substitutional theory, which would explain in an entirely new way what the naïve theory of classes was (albeit confusedly) getting at, and preserve, wherever possible, its mathematical uses.

At times Russell spoke of his theory of classes as if it offered a conceptual analysis of the statements of the naïve theory of classes, showing that class expressions of ordinary language, like definite descriptions, are not referential expressions. But properly speaking Russell is offering a retentative eliminativistic analysis. This explains why it is that Russell vacillated between describing his approach as the positive denial that there are classes (so that class expressions are non-referential expressions), and describing it as a form of agnosticism – recognizing from the perspective of the supplanting research program that classes, if they exist, are idle wheels that play no role in mathematical constructions. The approach is eliminivistic, but structurally retentive. "The only legitimate attitude about the physical world," Russell writes, "seems to be one of complete agnosticism as regards all but its mathematical properties" (Russell 1927: 271). This view might best be called 'structural realism.' Einstein's theory of relativity, for example, preserves Maxwell's equations concerning the propagation of electromagnetic energy in the aether. But it wholly abandons the ontology of the aether. Similarly, the major successes obtained by appeal to the existence of classes, the positive constructions of Cantor, Dedekind, Weierstrass, and Frege are to be retained within Russell's substitutional theory. Russell explained that "the principles of mathematics may be stated in conformity with the theory," and the theory "avoids all known contradictions, while at the same time preserves nearly the whole of Cantor's work on the infinite" (Russell 1906: 213). The substitutional theory involves, as Russell put it, "an elaborate restatement of logical principles." The results obtained by appeal to the existence of classes are conceptualized in an entirely new way within the research program

of the substitutional theory. There will be some loss – some flotsam – such as Cantor's transfinite ordinal number ω_ω, the usual generative process for the series of ordinals, and the class of all ordinals. But this loss is to be measured against the successes of the new program. Indeed, had the program yielded the conceptual successes that Russell had anticipated, one might venture to say that present mathematics would regard the notion of a class as present physics regards phlogiston, caloric fluid, the aether, and other relics of the past.

Russell's work to build types (and *Principia's* order/types) into formal grammar, reveals how he understood the analysis of logical form. One language (and the ontological entailments its predicates and grammar embody) is to be supplanted by another, technical language, for the purposes of science. In the new language, the old philosophical problems are solved. There are, for example, no entities of modern physics to identify with phlogiston, or caloric (of the caloric theory of heat), or the aether (of the wave theories of light). Transcription of the primitive's ontological problem of the elasticity of the aether, for example, will be impossible. It is rather that the new theory renders an explanation of what (if anything) was correct in the primitive's world view, and shows why the primitive's mistaken ontology was (to a limiting extent) on track. So also, in the new language of logical form that Russell envisioned – the 'logically perfect language' if you will – there are no predicate expressions '. . . exists,' or '. . . is a class, ' or '. . . is true,' or '. . . is a propositional function.' These are pseudo-predicates. But the logical grammar of the proper language for the calculus of the science of logic, shows the extent to which the naïve ontologies of earlier metaphysical systems were on the right track while capturing their important successes. Russell's eliminativistic conception of logical form offers a middle way between the Tarski-semantic conception of logical form employed by the Description Theory and the abandonment of logical form found in Kripke's defense of metaphyscial necessity. Russell's account of natural number, for example, is neither a meaning analysis of the concept 'natural number' nor is it properly understood an account of the metaphysical essence of natural numbers. Russell's program is one of analysis and reconstruction, where the "supreme maxim of all scientific philosophizing" is to be this: "Wherever possible, logical constructions are to be substituted for inferred entities" (Russell 1914: 115). Inspired by advances in mathematics, he contended that logic is the essence of philosophy: "every philosophical problem, when it is subjected to the necessary analysis and purification, is found either to be not really philosophical at all, or else to be, in the sense in which we are using the word logical" (Russell 1969: 42).

References

Bach, Kent (1987) *Thought and Reference.* Oxford: Clarendon Press.

Chomsky, Noam (1981) *Lectures on Government and Binding.* Dordrecht: Foris.

Cocchiarella, Nino (1975) On the Primary and Secondary Semantics of Logical Necessity. *Journal of Philosophical Logic*, 4, 13–27.

Cocchiarella, Nino (1977) Sortals, Natural Kinds and Reidentification. *Logique et Analysé*, 80, 439–74.

Cocchiarella, Nino (1981) Richard Montague and the logical analysis of language. In G. Fløstad (ed.), *Contemporary Philosophy: A New Survey*, vol. 2 (pp. 113–55). The Hague: M. Nijhoff.

Cocchiarella, Nino (1984) Philosophical persepectives on quantification in tense and modal logic. In D. Gabbay and F. Guenthner (eds.), *The Handbook of Philosophical Logic* (pp. 309–53). Dordrecht: Kluwer Academic.

Cocchiarella, Nino (1989) Conceptualism, realism and intensional logic. *Topoi*, 7, pp. 15–34.

Donnellan, Keith (1966) Reference and definite descriptions. *Philosophical Review*, 77, 281–304.

Dowty, D., Wall, R. and Peters, S. (1985) *Introduction to Montague Semantics*. Dordrecht: Reidel.

Evans, Gareth (1973) The causal theory of names. *Proceedings of the Aristotelian Society*, 47, 87–208. Reprinted in Peter Ludlow (ed.), *Readings in the Philosophy of Language*. Cambridge, MA: Bradford Books, 1997, 635–55.

Evans, Gareth (1985) *Collected Papers*. Oxford: Clarendon Press.

Fodor, J. and Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: a critical analysis. In S. Pinker and J. Mehler (eds.), *Connections and Symbols*. Cambridge, MA: MIT Press.

Fodor, J. and Pylyshyn, Z. W. (1987) Why there still has to be a Language of Thought. In *Psychosemantics* (pp. 135–67). Cambridge, MA: Bradford Books, MIT.

Frege, Gottlob (1980) On sense and reference. In P. Geach and M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege* (pp. 56–78). Oxford: Basil Blackwell.

Geach, P. T. (1950) Russell's theory of descriptions. *Analysis*, 10, 84–8.

Haack, Susan (1996) *Deviant Logic, Fuzzy Logic*. Chicago: Chicago University Press.

Kripke, Saul (1963) Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16, 83–94.

Kripke, Saul (1971) Identity and necessity. In M. K. Munitz (ed.), *Identity and Individuation* (pp. 135–64). New York University Press.

Kripke, Saul (1976) A puzzle about belief. In Avishi Margalit (ed.), *Meaning and Use: Papers Presented at the Second Jerusalem Philosophical Encounter*. Dordrecht: Reidel.

Kripke, Saul (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lambert, Karel (1990) Russell's theory of definite descriptions. *Dialectica*, 44, 137–52.

Laudan, Larry (1977) *Progress and its Problems*. Berkeley: University of California Press.

Laudan, Larry (1981) A confutation of convergent realism. *Philosophy of Science*, 48, 19–49.

McKay, Thomas (1975) Essentialism in quantified modal logic. *Journal of Philosophical Logic*, 4, 423–38.

Montague, Richard (1974) The proper treatment of quantification in ordinary English. In R. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague* (pp. 247–70). New Haven: Yale University Press.

Neale, Stephen (1990) *Descriptions*. Cambridge, MA: MIT Press.

Parsons, Terence (1969) Essentialism and quantified modal logic. *The Philosophical Review*, 78, 35–52.

Putnam, H. (1975) The meaning of "meaning". In K. Gunderson (ed.), *Language, Mind and Knowledge* (pp. 131–93). University of Minnesota Press. Reprinted in H. Putnam, *Mind, Language and Reality: Philosophical Papers*, vol. 2, Cambridge: Cambridge University Press, 1975.

Quine, W. V. O. (1976) Three grades of modal involvement. In *The Ways of Paradox and Other Essays* (pp. 158–76). Harvard University Press.

Quine, W. V. O. (1979) *Word and Object*. Cambridge, MA: MIT Press.

Ramsey, Frank (1931) Philosophy. In R. B. Braithwaite (ed.), *The Foundations of Mathematics and Other Logical Essays by Frank Plumpton Ramsey*. London: Harcourt, Brace.

Russell, Bertrand (1901) Mathematics and the metaphysicians. Printed with the title Recent Work in the Philosophy of Mathematics, *The International Monthly*. Reprinted in *Mysticism and Logic and Other Essays* (pp. 59–74). Barnes & Noble, 1976.

Russell, Bertrand (1905) On denoting. *Mind*, 14, 479–93.

Russell, Bertrand (1906) Les paradoxes de la logique. *Revue de Métaphysique et de Morale*, 14, 627–50. The English title is "On 'Insolubilia' and Their Solution By Symbolic Logic," and is reprinted in Lackey (ed.), *Essays in Analysis By Bertrand Russell*. London: Allen & Unwin, 1973.

Russell, Bertrand (1914) On the relation of sense-data to physics. *Scientia*, vol. 4. Reprinted in *Mysticism and Logic and Other Essays*. Barnes & Noble, 1976.

Russell, Bertrand (1927) *The Analysis of Matter*. London: Harcourt, Brace.

Russell, Bertrand (1946) *History of Western Philosophy*. New York: Simon & Schuster.

Russell, Bertrand (1959) *My Philosophical Development*. New York: Simon & Schuster.

Russell, Bertrand (1966) *An Inquiry into Meaning and Truth*. London: Allen & Unwin.

Russell, Bertrand (1969) *Our Knowledge of the External World*. London: Allen & Unwin.

Smullyan, Arthur (1948) Modality and description. *Journal of Symbolic Logic*, 13, 31–7.

Strawson, Peter (1950) On referring. *Mind*, 59, 320–44.

Whitehead, A. N. and Russell, Bertrand (1962) *Principia Mathematica*. Cambridge: Cambridge University Press.

This page intentionally left blank

Part V

CONCEPTS OF LOGICAL CONSEQUENCE

This page intentionally left blank

14

Necessity, Meaning, and Rationality: The Notion of Logical Consequence

STEWART SHAPIRO

There are many ways of saying that a given proposition, or sentence, Φ is a *logical consequence* of a set Γ of propositions, or sentences: Γ *entails* Φ, Γ *implies* Φ, Φ *follows from* Γ, Φ is a *consequence of* Γ, and the pair $\langle \Gamma, \Phi \rangle$ is *valid*. If a given Φ is a logical consequence of the empty set, we say that Φ is *logically true*, Φ is a *tautology*, or Φ is *valid*.

The notion of logical consequence has always been an important item on the agenda of philosophy. What is it? How do we determine that a given Φ is a consequence of a given Γ? Can we know this infallibly? A priori? What role does consequence play in our efforts to obtain knowledge? If Φ is a logical consequence of Γ, then what is the epistemic status of Φ vis-à-vis the members of Γ?

Logic is the study of correct reasoning, and has something to do with justification. Logical consequence is an important ingredient in *proof*. Thus, we broach issues concerning what *reasoning* is, and questions about what it is to reason *correctly*. The very notion of rationality is tied in here. Other slogans are that logic is topic neutral, and completely general, and that logical consequence is a matter of *form*. What do these slogans mean?

There are actually several different notions that go by the name of 'logical consequence,' 'implication,' etc. Some of them are controversial and some are, or may be, related to others in interesting and important ways. The purpose of this article is to sort some of this out.

1 Modality

As far as we know, the first systematic treatment of logic is found in Aristotle's *Prior Analytics*. In chapter 2 of book 1, we find:

> A deduction is a discourse in which, certain things having been supposed, something different from the things supposed results of necessity because these things are so. By "because these things are so," I mean "resulting through them" and by "resulting through them," I mean "needing no further term from outside in order for the necessity to come about".

I will not attempt to recapitulate the efforts of scholars to probe the subtleties in this text. To attempt a paraphrase, Aristotle's thesis is that a given proposition Φ is a consequence of a set Γ of propositions if (1) Φ is different from any of the propositions in Γ, (2) Φ *necessarily* follows from the propositions in Γ ('because these things are so'), and (3) propositions not in Γ are not needed for this necessity 'to come about.'

Contemporary practice is to drop clause (1) and allow that Φ follows from Γ when Φ is a member of Γ, as a trivial instance of logical consequence. Aristotle's phrase "because these things are so" seems to imply that in order to have a consequence, or 'deduction,' the premises in Γ must all be *true*. With one notable exception (Gottlob Frege), most modern conceptions of consequence do not follow this, and allow instances of logical consequence in which the premises are false. For example, 'Socrates is a puppy' follows from 'All men are puppies' and 'Socrates is a man.'

What of Aristotle's gloss of "because these things are so" as "resulting through them," and that as "needing no further term from outside in order for the necessity to come about." These clauses might indicate that the premises *alone* guarantee the conclusion, or that the premises are sufficient for the conclusion. Our first conception of logical consequence is modeled on this reading of Aristotle's definition:

(M) Φ is a logical consequence of Γ if it is not possible for the members of Γ to be true and Φ false.

It is common nowadays to think of modal notions in terms of possible worlds. For what that is worth, our thesis (M) becomes:

(PW) Φ is a logical consequence of Γ if Φ is true in every possible world in which every member of Γ is true.

According to (M) and (PW), 'Al is taller than Bill' seems to follow from 'Bill is shorter than Al,' since it is impossible both for Bill to be shorter than Al and for Al to fail to be taller than Bill. Surely, 'Al is taller than Bill' holds in every possible world in which 'Bill is shorter than Al.' Or so one would think. For another example, according to (M) and (PW), 'Hilary is wealthier than Barbara' and 'Barbara is wealthier than Nancy' seems to entail that 'Hilary is wealthier than Nancy.' Again, it is simply not possible for Hilary to be wealthier than Barbara, Barbara to be wealthier than Nancy, and for Hilary to fail to be wealthier than Nancy. To adapt an example from Bernard Bolzano, a religious person who accepts (M) or (PW) might say that 'Caius has an immortal soul' follows from 'Caius is a human' since (according to the person's theology), the premise *cannot* be true and the conclusion false.

On most contemporary accounts of logic, none of these conclusions is a logical consequence of the corresponding premise(s). It is a routine exercise to formalize these arguments and show that the conclusions do not follow (see below). Perhaps we can bring (M) and (PW) closer to the contemporary notions by articulating the involved modality, invoking a special notion of *logical* possibility and necessity. One tactic would be to invoke Aristotle's final clause, that a true logical consequence needs "no further term from outside in order for the necessity to come about." In the example about Bill and Al, we need to invoke some 'outside' fact about the relationship between shortness

and tallness in order for the 'necessity to come about.' In the second example, we need the fact that relative wealth is transitive, and in the third example, our theologian needs to invoke some theology for the conclusion to follow. The idea is that 'Caius has an immortal soul' follows from 'Caius is a human' *together with* the relevant theology, but 'Caius has an immortal soul' does not follow from 'Caius is human' alone.

On the other hand, I would think that there just is no possible world in which Bill is shorter than Al without Al being taller than Bill. Al being taller than Bill is part of what it is for Bill to be shorter than Al. And I presume that most theologians would insist that there are no possible worlds in which the relevant theology is false. Having an immortal soul is part of what it is to *be a human*.

2 Semantics

According to Alberto Coffa (1991), a major concern of philosophers throughout the nineteenth century was to account for the necessity of mathematics and logic without invoking Kantian intuition. Coffa proposed that the most successful line came from the 'semantic tradition,' running through the work of Bolzano, Frege, and Ludwig Wittgenstein, culminating in the Vienna Circle. The idea is that the relevant necessity lies in the use of language, or *meaning*. This suggests the following proposal:

(S) Φ is a logical consequence of Γ if the truth of the members of Γ guarantees the truth of Φ in virtue of the meanings of the terms in those sentences.

Thesis (S) rules out our theological example. I presume that even the most religious linguist or philosopher of language does not take it to be part of the *meaning* of the word 'human' that humans have immortal souls. One can perfectly grasp the relevant meaning and not know the relevant theology. So according to (S), 'Caius has an immortal soul' does not follow from 'Caius is human.' We are, however, still left with our other two examples. According to (S), 'Al is taller than Bill' does indeed follow from 'Bill is shorter than Al,' since the meanings of 'taller' and 'shorter' indicate that these relations are converses to each other. Similarly, meaning alone determines that 'Hilary is wealthier than Barbara' and 'Barbara is wealthier than Nancy' together guarantee that 'Hilary is wealthier than Nancy.' The meaning of 'wealthier' indicates that it is a transitive relation.

The thesis (S) captures what is sometimes called 'analytic consequence,' which is often *distinguished* from logical consequence, due to examples like those considered here. We now turn to a refinement of the semantic idea.

3 Form

As noted above, there is a longstanding view that logical consequence is a matter of *form*. As far as I know, Aristotle does not explicitly endorse this, but his work in logic is surely consonant with it. He sometimes presents 'deductions' by just giving the forms of the propositions in them. Moreover, to show that a given conclusion does *not* follow

from a given pair of premises, Aristotle typically gives an argument in the same form with true premises and false conclusion. It is most straightforward to interpret these passages as presupposing that if an argument is valid, then every argument in the same form is valid.

Consider a paradigm case of a valid argument:

All men are mortal; Socrates is a man; therefore, Socrates is mortal.

The validity of this argument does not turn on anything special about mortality and Socrates. Any argument in the form

All *A* are *B*; *s* is an *A*; therefore *s* is a *B*

is valid. That is, if one fills in the schematic letters *A*, *B* with any predicates or common nouns and *s* with any name or definite description, the result is a valid argument.

We might say similar things about the examples we used to illustrate the theses (M), (PW), and (S). Consider the following 'forms':

s is human; therefore *s* has an immortal soul
s is shorter than *t*; therefore *t* is taller than *s*
s is wealthier then *t*; *t* is wealthier than *u*; therefore *s* is wealthier than *u*

An argument in one of these forms has the same status, vis-à-vis (M), (PW), or (S), as the argument it was taken from. So one might think of these arguments as valid in virtue of form.

Nevertheless, the prevailing view is that the examples illustrating the above theses do not have a valid form. Although the one about shortness and tallness, for example, does not turn on anything special about the denotations of 'Bill' and 'Al,' it does turn on specific facts about (the meaning of) 'shorter' and 'taller.' On the prevailing view, the requisite logical forms of the above arguments are

s is *A*; therefore *s* is *B*
s is *S* than *t*; therefore *t* is *T* than *s*
s is *S* than *t*; *t* is *S* than *u*; therefore *s* is *S* than *u*

It is, of course, straightforward to find arguments in each of these forms that have true premises and false conclusion. So the original arguments are not valid in virtue of *these* forms.

At this point, a stubborn opponent might complain that even though the paradigm argument does not turn on anything special about Socrates, humanity, or mortality, it does turn on the specific meaning of 'all,' 'are,' and 'is.' We can give the following 'form' to our paradigm valid argument:

Π *A* is *B*; *s* is *A*; therefore, *S* is *B*,

and then give a counter-argument in that same 'form':

Some men are British; Clinton is a man; therefore Clinton is British

This has true premises and a false conclusion. So even the paradigm argument is not valid in virtue of the last-displayed 'form.'

The standard response would be to claim that the last-displayed 'form' is not a *logical* form of the paradigm argument. How, then, are we to characterize logical form? One might say that a form is logical if the only terms it contains (besides the schematic letters) are *logical terms*. Typically, these consist of truth-functional connectives ('not,' 'and,' 'or,' 'if . . . then'), quantifiers ('some,' 'all'), variables, and perhaps the sign for identity.

We now face the task of characterizing the logical terms. How do we go about designating a term as logical? The logician, or philosopher of logic, has three options. One is to attempt a principled definition of 'logical term,' perhaps by focusing on some of the traditional goals and purposes of logic (see, for example, Peacocke 1976; Hacking 1979; McCarthy 1981; Tarski 1986; Sher 1991). The proposals and theories cover a wide range of criteria and desiderata, such as a priori knowledge, analyticity, formality, justification, and topic-neutrality. It would take us too far afield to examine the proposals here. A second tactic, implicitly followed in most logic textbooks, is to merely provide a list of the logical terms, and to leave our task with this act of fiat. This is perhaps a safe route, since it avoids some sticky philosophical questions, but it might leave the readers wondering what is going on, and of course it provides no insight into the choice of logical terms. A third option (following Bolzano) is to make the notions of logical form and logical consequence relative. That is, one defines an argument to have a certain logical form *relative to* a given choice of logical terms. The same argument might be valid relative to one set of logical terms and invalid relative to another.

Some medieval logicians combined the notion of form with the *modal* conception of consequence. They defined a conclusion to be a *formal consequence* of some premises if (1) the argument is a consequence in the sense much like our (M) above, and (2) the result of any uniform substitution of the terms results in an argument that is also a consequence in that sense. In other words, suppose that Φ is a formal consequence of Γ. Then if Γ', Φ' are the result of a uniform substitution of the terms, then it is not possible for the members of Γ' to be true and Φ' false.

Our next notion of consequence combines the notion of logical form with a *semantic* conception like (S) above:

(FS) Φ is a logical consequence of Γ if the truth of the members of Γ guarantees the truth of Φ in virtue of the meanings of the logical terminology.

Another popular option is to avoid explicit mention of semantic notions like meaning altogether:

(Sub) Φ is a logical consequence of Γ if there is no uniform substitution of the nonlogical terminology that renders every member of Γ true and Φ false.

According to (Sub), logical consequence is defined solely in terms of logical form, substitution, and ordinary truth and falsehood. *Prima facie*, no metaphysically troublesome modal or semantic notions are involved.

4 Epistemic Matters

We still have not directly addressed the role of logical consequence in organizing and extending knowledge. As noted above, a common slogan is that logic is the study of correct *reasoning*. In particular, we reason from premises to conclusion via valid arguments. If we believe the premises, we must believe the conclusion, on pain of contradiction – whatever that means.

Let us propose another definition of consequence:

(R) Φ is a logical consequence of Γ if it is irrational to maintain that every member of Γ is true and that Φ is false. The premises Γ alone *justify* the conclusion Φ.

A theologian might admit that it is not irrational to hold that Caius is a human being without an immortal soul. The theologian should concede that someone can know that Caius is a human without knowing that he has an immortal soul. Some poor folks are ignorant of the relevant theology. So prima facie, the conception of consequence underlying (R) differs from the one underlying (M) above. On the other hand, there does seem to be something irrational in maintaining that Bill is shorter than Al while denying that Al is taller than Bill – unless of course one does not know the meaning of 'shorter' or 'taller.' But perhaps one can also rationally deny that Socrates is mortal while affirming that all men are mortal and Socrates is a man – if one pleads ignorance of the meaning of 'all.'

What is the penalty for being irrational? What exactly is the 'pain' of contradiction? The idea is that one who affirms the premises and denies the conclusion of a valid argument has thereby said things which *cannot* all be true. This broaches modal notions, as in (M) and (PW) from Section 1 above, but the pain of contradiction goes further than this. The charge is that our subject *could have known better*, and indeed should have known better. In this sense, logical consequence is a *normative* notion.

The most common way to articulate the modality and normativity here is in terms of *deduction*. If Φ is a consequence of Γ in this sense, then there should be a process of inference taking one from members of Γ to Φ. One purpose of such a deduction is to provide a convincing, final case that someone who accepts the members of Γ is thereby committed to Φ. So we have:

(Ded) Φ is a logical consequence of Γ if there is a deduction of Φ from Γ by a chain of legitimate, gap-free (self-evident) rules of inference.

Arguably, this notion also has its pedigree with Aristotle. He presents a class of syllogisms as 'perfectly' valid, and shows how to reduce other syllogisms to the perfectly valid ones by inference (see Corcoran 1974).

5 Recapitulation

I do not claim that the foregoing survey includes every notion of logical consequence that has been seriously proposed and maintained. For example, there is a tradition,

going back to antiquity and very much alive today, that maintains that Φ is not a logical consequence of Γ unless Γ is *relevant* to Φ. But to keep the treatment from getting any more out of hand, we will stick with the above notions. Here they are:

(M) Φ is a logical consequence of Γ if it is not possible for the members of Γ to be true and Φ false.

(PW) Φ is a logical consequence of Γ if Φ is true in every possible world in which every member of Γ is true.

(S) Φ is a logical consequence of Γ if the truth of the members of Γ guarantees the truth of Φ in virtue of the meanings of the terms in those sentences.

(FS) Φ is a logical consequence of Γ if the truth of the members of Γ guarantees the truth of Φ in virtue of the meanings of the logical terminology.

(Sub) Φ is a logical consequence of Γ if there is no uniform substitution of the non-logical terminology that renders every member of Γ true and Φ false.

(R) Φ is a logical consequence of Γ if it is irrational to maintain that every member of Γ is true and that Φ is false. The premises Γ alone justify the conclusion Φ.

(Ded) Φ is a logical consequence of Γ if there is a deduction of Φ from Γ by a chain of legitimate, gap-free (self-evident) rules of inference.

Our next question concerns what to make of all these notions. They do not *seem* to be pointing in the same direction. Nevertheless, one might hold that there is but a single underlying notion of logical consequence. On this view, if there is a divergence between two of the above notions, then we must conclude that (at least) one of them is incorrect. It fails to capture the true notion of logical consequence. On the other hand, the logician might be more eclectic, proposing that there are different notions of consequence, some of which are captured by the above notions. In this case, of course, the various notions are not necessarily rivals, even if they differ from each other.

In any case, there are connections between the above notions. Trivially, if an argument is valid in the sense (FS) then it is valid in the sense (S). If the premises guarantee the conclusion in virtue of the meaning of the logical terminology, then the premises guarantee the conclusion in virtue of meaning. As we have seen, the converse of this fails in cases where premises guarantee a conclusion in virtue of the meanings of the *non-logical* terminology. If an argument is valid in the semantic sense (S) then presumably it is valid in the modal sense (M) (and perhaps (PW)). That is, if the meaning of the terms guarantees that the premises cannot all be true and the conclusion false, then surely it is not possible for the premises to be true and the conclusion to be false. The converse, from (M) to (S), depends on whether there are necessary truths that do not turn on the meanings of terms. Our foregoing theologian thinks that there are such truths (e.g. about immortal souls).

The relationship between (Sub) and (S) turns on the boundary between logical and non-logical terms and the expressive resources available in the base language. For example, suppose that we are dealing with a 'language' in which the only predicates are 'was US President sometime before January 1, 2000' and 'is male,' and the only singular terms are 'Bill Clinton' and 'Hilary Clinton.' Then the argument:

Bill Clinton is (or was) US President; therefore Bill Clinton is male

comes out valid according to (Sub). Any uniform substitution of the (available) non-logical terminology that makes the premise true also makes the conclusion true. But, of course, this argument is not valid on any of the other conceptions. Its coming out as a (Sub)-consequence turns on the fact that the 'language' in question is amazingly impoverished.

Suppose that we follow standard practice and assume (or stipulate) that the logical terminology consists of truth-functional connectives, quantifiers, variables, and the sign for identity. Consider the following argument:

(weird) for every x there is a z such that $x \neq z$; therefore for every x and every y there is a z such that $x \neq z$ and $y \neq z$.

The premise 'says' that there are at least two things and the conclusion 'says' that there are at least three things. Both are true. Notice that neither of these propositions contains any non-logical terminology. So there are no substitutions to make, and so the argument is valid according to (Sub). This is not a comfortable result. Clearly, it is not part of the *meaning* of the logical terminology that if there are at least two things then there are at least three things. So (weird) is not valid according to (S) or (FS). Whether (weird) is valid according to (M) depends on whether it is necessary that if there are two things then there are three things. Whether (weird) is valid according to (R) depends on whether one can rationally maintain that there are two things while denying that there are three things in the universe. I do not venture an opinion on these matters of metaphysics and epistemology.

Suppose that an argument is valid according to the modal conception (M), so that it is not possible for its premises to be true and its conclusion false. Does it follow that it is *irrational* to believe the premises and deny the conclusion? Can an argument be valid in the sense of (M) even if no one knows, or can know, that the argument is valid? Conversely, suppose that it is irrational to believe some premises and still deny a conclusion. Does it follow that it is *impossible* for the premises to be true and the conclusion false? The philosophical literature reveals no consensus on these matters, and I propose to stay out of the debates here.

Suppose that an argument $\langle \Gamma, \Phi \rangle$ is valid in the sense of (S) (or (FS)). Then if someone knows that each member of Γ is true, then she can determine that Φ is true just by reflecting on the meanings of the words. In other words, anyone who knows the language and also knows every member of Γ thereby has the wherewithal to know that Φ is true. To adapt Aristotle's phrase, nothing 'from outside' the premises is needed to determine the truth of the conclusion. Presumably, the meaning of the premises is not outside of them. Thus, it is *prima facie* irrational to believe that the premises are true and the conclusion false. So the argument is valid according to (R). Turning to the converse, suppose that it is irrational to believe some premises while denying a conclusion. Does it follow that the premises guarantee the conclusion in virtue of meaning? Once again, it depends on the nature of the underlying notions. Are there any beliefs whose irrationality does not turn on meaning?

In addition to the nature of rationality, any connections between (Sub) and (R) turn on issues concerning the logical/non-logical boundary and the expressive resources of the language. I leave this as an exercise.

Turning to the deductive notion (Ded), we encounter the notion of a legitimate, gap-free, self-evident rule. Another slogan of logic is that rules of inference are truth-preserving. This seems to entail that if a legitimate, gap-free (self-evident) rule of inference takes one from some premises to a conclusion then it not possible for the premises to be true and the conclusion false. Thus, if an argument is valid in the sense of (Ded) then it is valid in the sense (M).

In articulating (Ded), we can maintain the theses that consequence turns on meaning (S) and that consequence is a matter of form (FS) by insisting that the *only* legitimate, gap-free rules of inference are those that flow from the meaning of the logical terminology (see Hacking 1979; Tennant 1987).

So it is plausible that if an argument is valid in the sense (Ded) then it is valid in the senses (M), (PW), (S), and (FS). W. V. O. Quine argues that if the logical/non-logical boundary is chosen judiciously and the language has sufficient expressive resources (as above), then an argument is valid in the sense (Ded) only if it is valid in the sense (Sub).

The converses of these implications are more problematic. Are there necessary truths that are not knowable via a derivation using only legitimate, gap-free self-evident rules? If so, then there are arguments that are valid in the sense (M) (and (PW)) but invalid in the sense (Ded).

What if the necessity in question turns on meaning alone (as in (S)), or what if the necessity turns on the meaning of the logical terminology? In that case, can we conclude that there is a chain of legitimate, gap-free, self-evident rules that go from premises to conclusion? This depends on whether all truths concerning meaning can be negotiated via the requisite type of derivation. I reiterate the emerging policy of not taking sides on debates like this.

The notions (Ded) and (FS) are equivalent (at least in extension) if the meaning of every logical term is exhausted by legitimate, gap-free (self-evident) rules of inference involving the term. This is also a matter of controversy. Some philosophers claim that a term is logical only if its meaning is determined completely by matching introduction and elimination rules (Hacking 1979; Tennant 1987). This view rules out the sort of non-effective consequence relation advocated by other philosophers and logicians (see Shapiro 1991: chapter 2).

6 Mathematical Notions

The foregoing notions, from (M) to (Ded), are intuitive conceptions of logical consequence, dealing with either sentences in natural languages or propositions expressed by such sentences. Most textbooks in logic give scant treatment to these intuitive notions, and quickly move to developing a formal language, which is a rigorously defined set of strings on a fixed alphabet. The books then focus exclusively on this 'language,' and at least seem to leave the intuitive notions behind. The resulting mathematics is, of course, interesting and important, but we can query its philosophical ramifications.

Typically, parts of a formal language correspond, roughly, to certain parts of a natural language. Characters like '&,' '$\vee$,' '$\rightarrow$,' '$\neg$,' '$\forall$,' and '$\exists$' approximately correspond to the English expressions 'and,' 'or', 'if . . . then,' 'it is not the case that,' 'for every,' and "there is," respectively. As above, these are logical terms. Some formal languages include specific non-logical terms, such as the sign for the less-than relation over the natural numbers, but it is more common to include a stock of schematic letters which stand for arbitrary, but unnamed, non-logical names, predicates, and functions. So one can think of a formula of a formal language as corresponding to a logical *form* in a natural language (or in the realm of propositions). The correspondence thus engages the slogan that logic is a matter of form (as in (FS)).

Let γ be a set of formulas and ϕ a single formula of a formal language. A typical logic text formulates two rigorous notions of consequence, two senses in which ϕ follows from γ. For one of the notions of consequence, the author presents a *deductive system* S, which might consist of a list of axioms and rules of inference. An argument $\langle\gamma,\phi\rangle$ in the formal language is *deductively valid* (via S) if there is a sequence of formulas in the formal language ending with ϕ, such that each member of the sequence is either a member of γ, an axiom of S, or follows from previous formulas in the sequence by one of the rules of inference of S. If $\langle\gamma,\phi\rangle$ is deductively valid via S, we write $\gamma \vdash_S \phi$, or simply $\gamma \vdash \phi$ if it is safe to suppress mention of the deductive system.

The other rigorous notion of consequence invokes a realm of *models* or *interpretations* of the formal language. Typically, a model is a structure $M = \langle d, I\rangle$, where d is a set, the *domain* of M, and I is a function that assigns extensions to the non-logical terminology. For example, if c is a constant, then Ic is a member of the domain d, and if R is a binary predicate, then IR is a set of ordered pairs on d. Then one defines a relation of *satisfaction* between interpretations M and formulas ϕ. To say that M satisfies ϕ, written $M \models \phi$, is to say that ϕ is true under the interpretation M.

Finally, one defines ϕ to be a *model-theoretic* consequence of γ if every interpretation that satisfies every member of γ also satisfies ϕ. In other words, ϕ is a model-theoretic consequence of γ if there is no interpretation that satisfies every member of γ and fails to satisfy ϕ. In this case, we write that the argument $\langle\gamma,\phi\rangle$ is model-theoretically valid, or $\gamma \models \phi$.

Model-theoretic consequence and deductive validity (via S) are both sharply defined notions on the formal language. So relations between them are mathematical matters. The system is *sound* if every deductively valid argument is also model-theoretically valid, and the system is *complete* if every model-theoretically valid argument is also deductively valid.

Typically, soundness is easily established, by checking each axiom and rule of inference. Completeness is a usually a deep and interesting mathematical result. Virtually every system presented in a logic text is sound. Gödel's (1930) completeness theorem entails that first-order logic (with or without identity) is complete. Second-order logic is not complete (see Shapiro 1991: chapter 4).

To begin an assessment of the philosophical import of the technical work, one must explore the relation between the rigorous notions (of deductive and model-theoretic consequence) and the intuitive notions broached above ((M) to (Ded)). Probably the closest conceptual connection is that between the deductive notion of consequence (Ded) and deductive validity via a standard deductive system. In so-called "natural

deduction" systems each rule of inference corresponds to a legitimate, gap-free (self-evident) inference in ordinary reasoning. So if an argument $\langle\gamma,\phi\rangle$ in the formal language is valid via such a system, and if a propositional or natural language argument $\langle\Gamma,\Phi\rangle$ corresponds to $\langle\gamma,\phi\rangle$, then Φ is a consequence of Γ in the sense (Ded). Although it is not quite as straightforward, something similar holds for other deductive systems. One typically indicates how each rule of inference corresponds to a chain of legitimate, gap-free inferences concerning ordinary reasoning, and that each axiom can be established by such rules.

Let S be a fixed, standard deductive system. One would like to establish a converse to the above conditional linking (Ded) to deductive validity via S. Call the following biconditional Hilbert's thesis:

> There is a deduction of a proposition (or natural language sentence) Φ from a set Γ of propositions (or natural language sentences) by a chain of legitimate, gap-free, self-evident rules of inference if and only if there is a corresponding argument $\langle\gamma,\phi\rangle$ in the formal language such that $\langle\gamma,\phi\rangle$ is deductively valid via S.

Perhaps one might restrict Hilbert's thesis to cases where the 'chain of gap-free, self-evident rules of inference' flow from the meaning of terminology that corresponds to the logical terminology of the formal language. This would focus attention on arguments that are valid in virtue of their logical form.

The philosophical interest of formal deductive systems depends on something like Hilbert's thesis. If there is no interesting connection between (Ded) (or (R)) and formal deductive validity, then the technical work is a mere academic exercise. Hilbert's thesis is the same kind of thing as Church's thesis, in that it identifies an intuitive, pre-theoretic notion with a precisely defined mathematical one. The exact nature of the identification depends on the relationship between formulas in the formal language and propositions or natural language sentences (or whatever it is that (Ded) applies to). At the very least, deductive validity via S is meant as a good mathematical model of (Ded).

Let us turn to model-theoretic consequence. The technical notion of satisfaction is a relation of truth-under-an-interpretation. Roughly, the relation $M \models \phi$ says that if the domain of M were the whole universe and if the non-logical terms are understood according to M, then ϕ is true. So model-theoretic consequence recapitulates the slogan that logical consequence is truth-preserving.

One might think of an interpretation as a possible world, which would link model-theoretic consequence to the modal notion (PW) and thus to (M). However, the complete freedom one has to 'interpret' the non-logical terminology (in the realm of model-theoretic interpretations) does not sit well with the modal notions. Consider one of our standby arguments: 'Hilary is wealthier than Barbara; Barbara is wealthier than Nancy; therefore Hilary is wealthier than Nancy.' A straightforward formalization would be *Whb*; *Wbn*; therefore *Whn*. To see that this formal argument is not model-theoretically valid, consider an interpretation whose domain is the natural numbers, and where W is 'within 3' (so that IW is $\{\langle x,y\rangle : |x-y| \geq 3\}$); Ih is 0; Ib is 2; and In is 4. This interpretation satisfies (i.e. makes true) the premises but not the conclusion. However, this interpretation has nothing to do with the *modal* status of the original argument about the relative wealth. It does not represent a genuine possibility con-

cerning the relative wealth of those women. In terms of (PW), the given interpretation does not correspond to a genuine possible world.

For much the same reason, model-theoretic consequence systematically diverges from the semantic notion (S), according to which Φ is a logical consequence of Γ if the truth of the members of Γ guarantees the truth of Φ in virtue of the meanings of the terms in those sentences. Again, it is part of the meaning of 'wealthier' that the relation is transitive. This feature of the meaning is lost in the given interpretation of the formal argument over the natural numbers.

Model-theoretic consequence does better with (FS): Φ is a logical consequence of Γ if the truth of the members of Γ guarantees the truth of Φ in virtue of the meanings of the *logical* terminology. Within the framework, the extension of the nonlogical terminology varies from interpretation to interpretation. So one can claim that the model-theoretic validity of a given formal argument $\langle\gamma,\phi\rangle$ is independent of the meaning of the nonlogical terminology. So, to the extent that model-theoretic validity depends on meaning, it depends only on the 'meaning' of the logical terminology in the formulas.

But does model-theoretic consequence depend *only* on meaning (as required for (FS))? Recall that the different interpretations have different domains. This feature of model-theoretic semantics does not seem to correspond to anything in (FS). Why should we vary the domain, in order to determine what follows from what *in virtue of the meaning* of the logical terminology? What do the varying domains have to do with meaning at all? One might show that the variation in the domains keeps arbitrary and nonlogical features of the universe (such as its size) from affecting logical consequence, ruling out arguments like the above (weird). But we need an argument to establish a direct link between the semantic notion of meaning and the variation of domains from interpretation to interpretation.

The fact that each interpretation has a domain, and that different interpretations can have different domains, does fit in nicely with the *modal* notion (M). The domain corresponds to what the totality of the universe might be. If we think of an interpretation as a possible world (perhaps invoking the notion (PW)), then the domain would be the universe of that world.

So perhaps model-theoretic consequence corresponds to a blending of a modal notion like (M) or (PW) with the notion (FS) that revolves around logical form. We say that Φ is a logical consequence of Γ in this blended sense if it is not possible for every member of Γ to be true and Φ false, and this impossibility holds in virtue of the meaning of the logical terms. In the terminology of possible worlds, Φ is a logical consequence of Γ in this blended sense if Φ is true in every possible world under every reinterpretation of the non-logical terminology in which every member of Γ is true.

In sum, perhaps we have several intuitive notions of consequence corresponding, at least roughly, to the formal notions of deducibility in a deductive system and model-theoretic validity. From the soundness and completeness of first-order logic (with or without identity), we have the two rigorous notions corresponding to each other exactly. To stretch things a bit, the situation is analogous to that of Church's thesis, where it was shown that a number of different mathematical notions (recursiveness, λ-definability, Turing computability, Markov computability, etc.) each corresponding to a different pre-theoretic idea of computability, are all coextensive with each other. In the case of Church's thesis, this is sometimes taken to be evidence that all of the notions are correct

– that they do accurately capture the underlying notion of computability. Given the wide range of notions of consequence noted above (not to mention those not noted above), and the tenuous connections between them, we should not make too strong a conclusion here in light of completeness. Perhaps we can tentatively suggest that validity in a standard deductive system and model-theoretic validity correspond to something like a natural kind. The exact nature of this natural kind, and its relationship to notions like necessity, possibility, meaning, form, deduction, and rationality requires further study.

References

Coffa, A. (1991) *The Semantic Tradition from Kant to Carnap*. Cambridge: Cambridge University Press.

Corcoran, J. (1974) Aristotle's natural deduction system. In J. Corcoran (ed.), *Ancient Logic and its Modern Interpretations* (pp. 85–131). Dordrecht: Reidel, 85–131.

Gödel, K. (1930) Die Vollständigkeit der Axiome des logischen Funktionenkalkuls. *Montatshefte für Mathematik und Physik*, 37, 349–60; translated as "The completeness of the axioms of the functional calculus of logic," in Jean van Heijenoort (ed.), *From Frege to Gödel*, Cambridge, MA: Harvard University Press, 1967, 582–91.

Hacking, I. (1979) What is logic? *Journal of Philosophy*, 76, 285–319.

McCarthy, T. (1981) The idea of a logical constant. *Journal of Philosophy*, 78, 499–523.

Peacocke, C. (1976) What is a logical constant? *Journal of Philosophy*, 78, 221–40.

Shapiro, S. (1991) *Foundations without Foundationalism: A Case for Second-order Logic*. Oxford: Oxford University Press.

Sher, G. (1991) *The Bounds of Logic*. Cambridge, MA: MIT Press.

Tarski, A. (1986) What are logical notions. In John Corcoran (ed.), *History and Philosophy of Logic*, 7, 143–54.

Tennant, N. (1987) *Anti-Realism and Logic*. Oxford: Oxford University Press.

Further Reading

The following is only a sampling of the many articles and books on our subject. I apologize to neglected authors.

Anderson, A. and Belnap, N. (1975) *Entailment: The Logic of Relevance and Necessity I*. Princeton, NJ: Princeton University Press. (An extensive study of logical consequence, arguing that the premises of a logical consequence must be relevant to its conclusion.)

Anderson, A., Belnap, N. and Dunn, M. (1992) *Entailment: The Logic of Relevance and Necessity II*. Princeton, NJ: Princeton University Press. (A sequel to the above.)

Bolzano, B. (1837) *Theory of Science*, trans. R. George. Berkeley, University of California Press, 1972. (A presentation of a substitutional account of consequence, along the lines of (Sub).)

Corcoran, J. (1973) Meanings of implication. *Dialogos*, 25, 59–76. (Presents a lucid account of several different notions of consequence, focusing on the role of formal deductive systems and a relation like (Ded).)

Etchemendy, J. (1990) *The Concept of Logical Consequence*. Cambridge, MA: Harvard University Press. (A critical study of the influential account in Tarski (1935) and of contemporary model-theoretic consequence. This book generated many responses, including Sánchez-Miguel (1993), Sher (1996), and Shapiro (1998).

Quine, W. V. O. (1986) *Philosophy of Logic*, 2nd edn., Englewood Cliffs, NJ: Prentice-Hall. (An influential substitutional account, along the lines of (Sub), relating that conception to the formal notions of model-theoretic consequence and deducibility in a standard deductive system.)

Sánchez-Miguel, M. (1993) The grounds of the model-theoretic account of the logical properties. *Notre Dame Journal of Formal Logic*, 34, 107–31. (A nice study of the role of model-theoretic consequence, responding to Etchemendy (1990).)

Shapiro, S. (1998) Logical consequence: models and modality. In M. Schirn (ed.), *The Philosophy of Mathematics Today*. Oxford: Oxford University Press, 131–56. (A defense of model-theoretic consequence, relating it to the various intuitive notions.)

Sher, G. (1996) Did Tarski commit "Tarski's fallacy"? *Journal of Symbolic Logic*, 61, 653–86. (A reaction to Etchemendy (1990), relating Tarski's account to the model-theoretic one.)

Tarski, A. (1935) On the concept of logical consequence. In A. Tarski, *Logic, Semantics and Metamathematics*. Oxford: Clarendon Press, 1956; 2nd edn., edited and introduced by John Corcoran (Indianapolis: Hackett, 1983), 417–29. (A very influential article, developing a notion of consequence evolving from (Sub) but avoiding the tie to the expressive resources of the base language. The notion of satisfaction is introduced. The connection between this Tarskian notion and the contemporary model-theoretic one is a matter of controversy.)

15

Varieties of Consequence

B. G. SUNDHOLM

I

Contemporary – metamathematical – logic operates with two kinds of consequence. In both cases the consequence in question is a relation among (sets) of well-formed formulae (wffs) in a certain formal language $\mathcal{L}$. In order to keep my exposition maximally simple I shall first consider a language for the propositional calculus, using only the connectives $\supset$ ('implication') and $\bot$ ('absurdity') as primitive, and with

$$p_0, p_1, p_2, \ldots, p_k, \ldots,$$

as propositional letters.

The (well-formed formulae of the) formal language $\mathcal{L}$ are given by a standard inductive definition:

(0) $\bot$ is an (atomic) wff in $\mathcal{L}$.
(1) p_k is an (atomic) wff in $\mathcal{L}$, for every $k \in \mathbb{N}$.
(2) When A and B are wffs $\mathcal{L}$, then so is $(A \supset B)$.
(3) There are no other wffs in $\mathcal{L}$ than those one obtains through finitely repeated applications of (0)–(2).

The clauses (0) and (1) are the *basic* clauses for the inductive definition, whereas the clause (2) constitutes the *inductive* clause. Jointly they tell us what to put into the inductively defined class. The clause (3), finally, is the *extremal* clause, that tells us what to exclude from the class in question. (In languages with such a sparse collection of primitive notions, the other standard connectives are defined in the usual way from $\supset$ and $\neg$ ('negation'), where the stipulatory definition

$$\neg A =_{\text{def}} (A \supset \bot)$$

takes care of the negation.)

Both kinds of consequence are inductively defined with respect to the build-up of the well-formed formulae of the language in question. The first notion, which is the later

one from a chronological point of view, is *semantical* in that it makes use of interpretations, or models, for the formal language $\mathcal{L}$.

We consider the two Boolean truth-values **T**(rue) and **F**(alse). A *valuation* v is a function from N to {**T**, **F**}. This valuation v is then extended to a valuation v^* for all of $\mathcal{L}$ via the following inductive definition:

(0) $v^*(\bot) = \mathbf{F}$, that is, from a contentual point of view, absurdity is false (under any valuation);
(1) $v^*(p_k) = v(p_k)$ (which value $\in \{\mathbf{T}, \mathbf{F}\}$);
(2) $v^*(A \supset B) = \mathbf{T}$ when $v^*(A) = \mathbf{T}$ implies that $v^*(B) = \mathbf{T}$, and $= \mathbf{F}$ otherwise.

A valuation v such $v^*(\varphi) = T$ is a *model* of the wff φ. When Σ is a set of wffs in $\mathcal{L}$, we extend v^* also to the set Σ

$v^*(\Sigma) = \mathbf{T}$ when $V^*(\psi) = \mathbf{T}$, for all wffs $\psi \in \Sigma$.

Finally we are ready to define the notion of *(logical) consequence*

the *consequent* φ is a (logical) consequence of the set of *antecedents* Σ (in symbols $\Sigma = \varphi$),
iff $v^*(\varphi) = \mathbf{T}$ for any valuation v^* such that $v^*(\Sigma) = \mathbf{T}$.

We write

$\psi = \varphi$

for '$\{\psi\} = \varphi$.' (The sign '=' is known as a *turnstile.*)

Accordingly, φ is a (logical) consequence of ψ when every model of a model of ψ is also a model of φ.

On this construal, then, (logical) consequence is a universal notion, defined by means of universal quantification over functions (or sets), since one considers *all* models satisfying a certain condition. (Thus, consequence is refuted by a countermodel, that is, a valuation that makes the antecedent true and the consequent false.) This universality of consequence is a typical feature which is retained also for more complex languages: for instance the above pattern is kept also for the predicate calculus, albeit that the notion of valuation is considerably more intricate in that case.

II

The other notion of consequence for the language $\mathcal{L}$ is *syntactical*, rather than semantical, in character. It is defined, not in terms of truth under all valuations, but in terms of the existence of a 'derivation' from certain 'axioms.'

Any wff of $\mathcal{L}$ that is an instance of one of the following schemata is an *axiom*:

(0) $(A \supset (B \supset A))$;
(1) $((A \supset (B \supset C)) \supset ((A \supset B) \supset (A \supset C)))$;
(2) $(((A \supset \bot) \supset \bot) \supset A)$;

The *theorems* ('derivable' formulae) are then defined via (yet again!) an inductive definition:

(0) Any axiom is derivable (is a theorem).
(1) If $(\varphi \supset \psi)$ and φ are derivable (theorems), then so is ψ.
(2) There are no other theorems than those obtained from repeated applications of (0) and (1).

When the wff φ is derivable we use a single turnstile, rather than the double semantical turnstile '$\models$', and write '$\vdash\varphi$.'

By the above definition every theorem is a theorem in virtue of a *derivation*. Such derivations are in tree form and have axioms at their topmost leaves: there is no other way to commence a derivation save by an axiom. Deeper down the tree is regulated by the rule of *modus ponens*:

$$\frac{\vdash A \supset B \quad \vdash A}{\vdash B.}$$

Properties of *all* theorems can then be established by 'induction over the (length of the) derivation.'

In order to obtain the syntactic notion of consequence we must extend the notion of derivability to 'derivability from assumptions in the set Σ.' We proceed (yet again) via an inductive definition:

(0) φ is derivable from assumptions Σ whenever φ is an axiom;
(1) φ is derivable from assumptions in Σ whenever $\varphi \in \Sigma$;
(2) If $(\varphi \supset \psi)$ and φ are derivable from assumptions in Σ, then so is ψ.
(3) No wff is derivable from assumptions in Σ save by a finite number of applications of (0)–(1).

The syntactic turnstile is then extended to cover also derivability from assumptions: we write '$\Sigma \vdash \varphi$' when φ is derivable from assumptions in Σ. Also theorems from assumptions in Σ have derivations (from assumptions in Σ); such derivations from assumptions in Σ allow as top-formulae, not only axioms, but also wffs from the set Σ. We then see that derivability from assumptions, that is syntactic consequence, does not share the universal form of semantic consequence. On the contrary, syntactic consequence holds in virtue of the *existence* of a suitable derivation. This generation of the syntactic notion of consequence via axioms, rules of inference, and added assumptions is not the only way of proceeding. In the early 1930s, Gentzen and Jaskowski took derivability from assumptions as the basic notion in their systems of natural deduction, using no axioms, but inference rules only, where outright derivability can be defined as derivability from no assumptions.

III

The *soundness* and *completeness* theorems for a formal system relate the semantic and syntactic notions of consequence. Soundness states that every syntactic consequence

is also a semantic consequence, while the opposite direction is taken care of by completeness.

The above pattern of semantic and syntactic consequence relations is omnipresent in current metalogic. Predicate logic, second- and higher-order systems, extensions to infinitary languages, modal logics; all and sundry confirm to the basic pattern. In the early days of mathematical logic the syntactic consequence-relation was the primary one. A formal system was given showing how its theorems were generated from axioms via rules of inference. However, as more experience was gained of matters semantical, through the work of Alfred Tarski and his pupils, notably Dana Scott, the semantical perspective gained prominence. Today it is fair to say that the semantical way of proceeding is the more fundamental one, partly also because some logics (systems), such as full second-order logic or the logic of the so-called Henkin-quantifier, do not allow for complete axiomatization.

The extension of the above (excessively simple) notion of valuation to the language of first order logic proved non-trivial. In the case of a first-order language $\mathcal{L}$, containing only the two-place predicate symbol R, the individual constant c, and for simplicity, no further function symbols, we interpret with respect to a relational structure

$$A = \langle A, R^A, c^A \rangle,$$

where the set $A \neq \emptyset$. The problem here is that, in general, the domain of discourse, that is, the set A, contains more elements than can be named by constants of $\mathcal{L}$. This problem – technical, rather than conceptual – was solved by Tarski using *assignments*. An assignment is a function $s \in N \rightarrow A$, and the terms of the language $\mathcal{L}$ are evaluated relative to this assignment:

(0) $s^*(c) = c^A$;
(1) $s^*(x_k) = s(k)$.

The formulae are then evaluated in the obvious way mimicking the inductive steps for the propositional calculus in the definition of the three-place metamathematical relation $A =_s \varphi$ – 'the assignment s satisfies the wff φ in the structure A':

(0) $A =_s R(t_1, t_2)$ iff $< s^*(t_1), s^*(t_2) > \in R^A$;
(1) not: $A =_s \bot$:
(2) $A =_s (\varphi \supset \psi)$ iff $\mathbf{A} =_s \varphi$ implies $A =_s \psi$
(3) $A =_s (\forall x_k \varphi)$ iff for all $a \in A$, $A =_{s[a/k]} \varphi$,

where the function $s[a/k] \in N \rightarrow A$ is defined by

$$s[a/k](m) =_{def} s(m) \text{ if } m \neq k;$$
$$=_{def} a \text{ if } m = k.$$

One should here note that traditionally, and unfortunately, the double turnstile is used for *two different notions*, namely

satisfaction – a three-place relation between a structure A, a wff φ and an assignment s,

and

(logical) consequence – a two-place relations between (sets of) wffs.

The above definitions, with the relativization to varying domains of discourse, are essentially due to Tarski, and were, perhaps, first published in final form only as late as 1957. (Tarski's earlier (1936) work on the definition of logical consequence had left this relativization out of account.) Once this definition of satisfaction is given, the definition of logical consequence also for this extended language of first-order predicate logic is readily forthcoming, namely as the preservation of satisfaction by the same assignment from antecedents to consequent.

IV

The above orgy of inductive definitions, which commenced in his famous work on the definition of truth, was not Tarski's only contribution to the theory of consequence (-relations). Already in 1930 he considered an abstract theory of consequence that was obtained by generalization from the syntactic consequence relation above. We consider a set S of 'sentences' and a consequence operator Cn defined on sets of sentences. Tarski then uses axioms such as:

(0) $S \neq \emptyset$ and $\text{card}(S) \leq \mathcal{N}_0$;
(1) If $X \subseteq S$, $X \subseteq Cn(X) \subseteq S$;
(2) If $X \subseteq S$, $Cn(Cn(X)) = Cn(X)$;
(3) If $X \subseteq S$, $Cn(X) = \cup \{Cn(Y): Y \subseteq X \text{ and } \text{card}(Y) < \mathcal{N}_0\}$;
(4) For some $x \in S$, $Cn(\{x\}) = S$.

These axioms are clearly satisfied by the above notion of syntactic consequence: axiom (2) says that using derivable consequences as extra assumptions does not add anything, and axiom (3) expresses that a derivation makes use only of finitely many assumptions, while the absurdity $\perp$ serves as the omniconsequential sentence demanded by axiom (4).

Around the same time, Gerhard Gentzen, building on earlier work by Paul Hertz, gave a formulation of elementary logic in term of *sequents*. A sequent is an array of wffs

$$\varphi_1, \ldots, \varphi_k \Rightarrow \psi.$$

(In some systems Gentzen allows more then one 'succedent-formula' after the arrow.)

Then, the derivable objects of his sequent calculi are sequents, rather than wffs. Derivations begin with *axioms* of the schematic form

$$A \Rightarrow A,$$

that is, the wff A is a consequence of, is derivable from, the assumption A. Depending on which kind of calculus one chooses, the derivation then proceeds by adding complex formulae using either (left and right) *introduction*-rules only, in which case we have a *sequent calculus*, or introduction- and elimination-rules, which operate solely to the right of the arrow, in which case we have a *sequential formulation of natural deduction*. For instance, in the sequent calculus as well as in the sequential natural deduction calculus, the (right) introduction rule for conjunction & (where the language has been extended in the usual fashion) takes the form

$$\frac{\Gamma \Rightarrow A \quad \Sigma \Rightarrow B}{\Gamma, \Sigma \Rightarrow A \,\&\, B}$$

where Γ and Σ are lists (or sets, or 'multisets') of wffs, depending on what representation has been chosen for sequents. The left introduction rule has the form

$$\frac{A, B, \Gamma \Rightarrow C}{A \,\&\, B, \Gamma \Rightarrow C}$$

and is justified by (and describes) the natural deduction elimination-rules

$$\frac{\Gamma \Rightarrow A \,\&\, B}{\Gamma \Rightarrow A} \quad \text{and} \quad \frac{\Gamma \Rightarrow A \,\&\, B}{\Gamma \Rightarrow B}$$

If C can be obtained from assumptions A, B, then C can be obtained from an assumption A&B, since, by the elimination rules, from A&B one gets both A and B.

V

The above pattern with two metamathematical consequence relations, one syntactic and one semantic, is present throughout the whole gamut of (metamathematical) logic; it has been carried out for classical logic (and its intuitionistic rival). Among so called 'philosophical logics' not only familiar modal logic(s) and the logic of counterfactual conditionals have been so treated, but also more exotic members of the wide logical family such as doxastic and erotetic logic, relevance logic, paraconsistent logic, and so on, have been brought within the fold. You name your favorite logical system and the chance is very high, indeed, that it has a syntax and semantics, with ensuing soundness and completeness theorems. When the entire pattern cannot be upheld, the semantic definition is generally given pride of place. Soundness of the syntactic consequence relative to the semantic one is a *sine qua non*, whereas completeness of the syntactic rule-system with respect to the semantic consequence is a strong desideratum, naturally enough, but cannot always be guaranteed. As already noted, full second order logic, with quantification over really *all* subsets of the universe cannot be effectively axiomatized with decidable axioms and rules of inference. As is well-known (from the work of Richard Dedekind), using full second-order quantification, it is possible to characterize the natural numbers up to isomorphism. Thus, in view of Tarski's theorem

concerning the arithmetical undefinability of arithmetical truth, theoremhood in second-order logic cannot be arithmetical, much less recursively enumerable. So, therefore, there are no appropriate syntactic characterizations of this prior semantic notion of second-order logical truth and consequence. This failure – unexpected, unavoidable, and unwanted – of completeness in full second-order logic holds with respect to a prior more or less 'natural' semantics. Sometimes though, especially in the case of various (artificial) modal and tense logics, the opposite direction poses the challenging task of actually designing syntactic rule-systems that (provably) have no complete semantics of a given kind. Such constructions, though, are of limited philosophical interest in themselves. To my mind, they can be compared to the construction of pathological counter-examples in real analysis, for example of a non-differentiable contiunuous function: *that* there is such a function is interesting, but the function itself is not very interesting.

VI

The wffs are considered solely as metamathematical objects and also their 'interpretation' was metamathematical rather than semantic, that is, no proper meaning has been assigned to the formulae. When considering natural-language interpretations of the formal calculi, I shall use the following terminology. An assertion is commonly made through the utterance of a declarative that expresses a statement. (This is not to say that every utterance of a declarative is an assertion; it is, however, a convention concerning the use of language that an utterance of a declarative, in the absence of appropriate counter-indications, is held to be an assertion.) The content of the statement expressed by a declarative is a proposition. Propositions can be indicated by means of nominalized declaratives, that is, by *that clauses*. Thus, for instance,

that snow is white,

is a proposition. However, one cannot make an assertion by means of a proposition only; for this we need to add

. . . is true,

to the that-clause, in order to get a statement in declarative form, by means of which an assertion can be effected. Thus

that snow is white is true

is the explicit form of the statement expressed by the declarative

snow is white.

The content of the statement in question is the proposition that snow is white. Thus the declarative *snow is white* expresses the statement

that snow is white is true

which has the proposition *that snow is white* as its content.

When the wffs are interpreted as propositions, they may be thought of as that-clauses, that is, nominalizations of declarative natural language sentences, such as

that snow is white and *that grass is green*,

an *implication* wff $(\varphi \supset \psi)$ is interpreted as, for instance, the proposition

that *that snow is white* implies *that grass is green*,

which is the same proposition as

the implication of *that snow is white* and *that grass is green*.

The sequent

$\varphi \Rightarrow \psi$,

on the other hand, is then interpreted as the consequence-statement

that grass is green is true under the assumption (on condition, provided) that *that snow is white* is true.

However, the statement

that snow is white is true

is the same as the statement

snow is white,

that is, the same assertion would be effected by uttering either.

Accordingly, the above consequence-statement is the same statement as

grass is green under the assumption (on condition, provided) that snow is white,

or, indeed, in conditional form,

if snow is white, then grass is green.

(Thus, I take it, these example show that 'implication' is different from the conditional 'if . . . , then ____'; an implication takes propositions (that is, what that-clauses stand for) and yields a statement, whereas the conditional takes statements and yields a statement. Finally, in order to saturate the expression 'the implication of . . . and ____,'

two that-clauses are needed, and we then get a term that stands for an implicational proposition.)

VII

Both approaches to consequence – semantic and syntactic – have counterparts in a long-standing logical tradition. In order to understand this it is necessary to memorize one of the decisive steps in the development of logic that was taken by Bolzano in his monumental *Wissenschaftslehre*, Theory of Science, from 1837. There he discarded the traditional, two-term form of judgement [S is P] and replaced it with the unary form

the proposition A is true.

Bolzano's propositions are *Sätze an sich* and serve as contents of judgments. Frege, indeed, used "judgable content" for the very same notion. They are independent of any *Setzung* whether by mind or language and do not belong in the physical or mental realm, but belong to a platonic third realm, having no spatial, temporal, or causal features. Thus they exhibit the same pattern as Frege's *Gedanken* ("Thoughts"), that is, the judgable contents in a later guise. Bertrand Russell, in what is surely the worst mistranslation in the history of logic, rendered Frege's "Gedanke" as "proposition" in *The Principles of Mathematics*, and he and G. E. Moore, who had inspired Russell's use, bear the responsibility for the resulting confusion. Throughout the earlier logical tradition the term *proposition* was invariably used for speaking about judgments and not about their contents. This (unacknowledged and maybe even unwitting) change in the use of the term has had dire consequences for our understanding of the theory of inference.

In the late middle ages (±1300) a novel genre was added to the logical repertoire. Around that time tracts "on consequences" (*De Consequentiis*) begin to appear, in which the theory of inference was treated differently from what had been common up till then. The old theories had been squarely syllogistic in nature, studying (what amounts essentially to) Aristotelian term-logic, whereas now one begins to find treatments of a more propositional kind. Today, introductory courses in logic commonly teach students to look for 'inference indicators' when analyzing informal arguments. Typical such indicator-words are

therefore, thus, whence, hence, because, and, sometimes even, *if . . . , then.*

The medieval *consequentia* were at least of four kinds and knew the indicator words:

(i) *si* (if): If snow is white, then grass is green;
(ii) *sequitur* (follows from): That grass is green follows from that snow is white;
(iii) *igitur* (therefore): Snow is white. Therefore: grass is green;
(iv) *quia* (because): Grass is green.

Note that these words did not serve to indicate different notions: on the contrary, all four point to one and the same notion. Thus the laws for *consequentia* should hold under

all four readings. Today, it must be stressed, it would seem more apposite to distinguish four different notions, rather than to have the four versions of the medieval notion:

(i′) *Conditional*, which forms a statement out of statements;
(ii′) *Consequence*, which forms a statement out of (modern) propositions;
(iii′) *Inference*, which is a passage from known statement(s) to a statement;
(iv′) *Causal grounding*, which is a relation between state of affairs or events.

The medievals applied a single correctness-notion *tenere* (holds) to *consequentia*. Each of the four current notions, however, matches its own correctness-notion. The appropriate notions of correctness are, respectively:

When correct
(i″) conditionals are *true*;
(ii″) consequences *hold*;
(iii″) inferences are *valid*;
(iv″) causal groundings *obtain*.

Unless we wish to follow the medieval pattern, we shall have to inquire into the conceptual order of priority, if any, among the various kinds of *consequentia* and their matching correctness notions. It will then also prove convenient to add a fifth notion, namely

(v) Implication, which takes two propositions and yields a proposition, namely: the implication of that snow is white and that grass is green
[= the proposition that *that snow is white* implies *that grass is green*].

Here the appropriate correctness notion is *truth* (for propositions), naturally enough.

VIII

Aristotle, in the *Posterior Analytics*, imposed three conditions on the principles that govern demonstrative science: ultimately, a proof, or demonstration, has to begin with principles that are (1) general, (2) *per se*, and (3) universal. The generality in question means that first principles should be in a completely general form: they speak about all things of a certain kind. Particular knowledge of particulars does not constitute the right basis for logic. To some extent the demand for universality is related to this: it comprises a demand for topic-neutrality. The general principles must be applicable across the board; not only within geometry or arithmetic or biology, but in any discipline. These demands for generality and universality on the basic principles of demonstrative science have a counterpart in one of the ways in which the medieval treated of the validity of inference, namely the *Incompatibility* theory. It goes back to Aristotle's *Prior Analytics* and was perhaps first clearly enunciated in the Stoic propositional approach to logic. It was firmly upheld by Parisian logicians in the early fifteenth century. The general inference I:

$$\frac{J_1 \ldots J_k}{J}$$

is held to be valid if the *truth of the premises* $J_1, \ldots, J_k$ is *incompatible* with *the falsity of the conclusion* J. Thus, by trivial computation in Boolean and modal logic, we get

[A is true. *Therefore*: B is true] is valid
iff
[A true **and** B false] are incompatible
iff
$\neg\Diamond$[A true **and** B false]
iff
$\Box\neg$[A true **and** B false]
iff
$\Box$ [**if** A true, **then not**-(B false)]
iff
$\Box$ [**if** A true, **then** B true].

The question is now how the modal box '$\Box$', that is, the necessity in question, should be interpreted. One natural way of proceeding here is to take necessity in the sense of 'holds in all alternatives.' This was done by an influential school of medieval logicians, who read the universality and topic neutrality as 'holds *in omnis terminis*' and so the logically valid is that which holds in all terms. The above chain of equivalences the continues:

for any variation ′ with respect to sub-propositional parts
if A′ true, **then** B′ true.

This is how Bernard Bolzano defined his notion of *Ableitbarkeit* (consequence) in 1837; note that this is a *three*-place relation between antecedent(s), consequent(s) and a collection of ideas-in-themselves (that is, the relevant sub-propositional parts, with respect to which the variation takes place). *Logische Ableitbarkeit – logical* consequence – then involves variation of with respect to all non-logical sub-propositional parts. Similarly, Bolzano held that a *Satz an sich*, that is, a proposition, was 'analytic in the logical sense' if the proposition remained true with respect to arbitrary variation at all nonlogical sub-propositional parts. One century later, essentially the same characterization of logical truth was offered by Ajdukiewicz and Quine (for *sentences* though, rather than Bolzano's propositions).

To some extent this notion of truth under variation is captured by the modern model-theoretic notion. The parallel is not exact, though. In the Bolzano–(Ajdukiewicz–Quine) conception variation takes place with respect to the proposition (or sentence), whereas in the semantic, model-theoretic notion what is varied is *not* the metamathematical counterpart to the proposition (sentence), that is, the well-formed formula. On the contrary, the variation takes place with respect to the relational structure A. Thus, if anything, it is the *world*, rather than the *description* thereof, that is varied. Thus, the notion of a *tautology*, that is, a proposition of logic, from Wittgenstein's *Tractatus* is a better contentual counterpart to the model theoretic notion of logically true wff. A tautology is a proposition which is true, come what may, independently of what is the case in the world (irrespective of how the world is or of what states of affairs obtain in the

world), and similarly for the notion of consequence. This onto-logical conception of logical truth and validity seems to me to capture best the intuitions that are formalized in the model-theoretic notion of semantic consequence.

IX

Given Bolzano's form of judgment, the general form of inference I is transformed into I′:

$$\frac{A_1 \text{ is true} \ldots A_k \text{ is true}}{C \text{ is true.}}$$

Bolzano reduces the validity of this inference I′ to the *logical* holding of the sequent

$$A_1, \ldots, A_k \Rightarrow C.$$

This, in turn, is equivalent to that

$A_1 \& \ldots \& A_k \supset C$ is logically true.

This reduction is exactly parallel to his reduction of the correctness ('truth') of the statement to the *propositional* truth of the content A. Bolzano here says that the judgment [A is true] is correct (*richtig*) when the proposition A really is true. Stronger still, the judgment

[A is true] is (a piece of) knowledge (*ist eine Erkenntnis*)

when the proposition A is true. This, however, admits of the unpalatable consequence that blind judgments are knowledge, irrespective of any epistemic grounding. (The apt term *blind judgment* was coined by Franz Brentano.) The statement

The Palace of Westminster has 1,203,496 windowpanes

is *knowledge* if by fluke, but not by telling, I have happened to choose the right number when constructing the example, that is, if the proposition

that The Palace of Westminster has 1,203,496 windowpanes

is a propositional truth (*an sich*, as Bolzano would say).

Entirely parallel considerations yield that also *blind* inference, without epistemic warrant, is valid under the Bolzano reduction. This, to me, is sufficient to vitiate the Incompatibility theory with its Bolzano reductions and thus I prefer to search for other accounts of validity that do not allow for the validity of blind inference. One such is readily forthcoming in the *Containment* theory. This also has Aristotelian roots, was perhaps first adumbrated by Peter Abailard, and was squarely defended by 'English

logicians' at Padua in the fifteenth century. Here an inference is valid if the truth of the conclusion is somehow analytically contained in the truth of premises. The Bolzano-reduction reduced the correctness of a judgment to the propositional (bivalent) truth of its content. This, while pleasingly simple, leaves the vital epistemic justification completely out of the picture and it was left to Franz Brentano to suggest an evidence theory of correctness ('truth') for statements:

a statement is correct if it can be made evident.

Correctness, or truth, at the level of statements (judgments), is accordingly a modal notion.

Indeed, at this level, the equation

true = evidenceable, knowable, justifiable, warrantable,

holds. It must be stressed here that it is at the level of what is known that correctness coincides with knowability. A true, or correct, statement is knowable, but one must not export this to the propositional content of the statement in question. The object of the act of knowledge is a judgment concerning the truth of a propositional content and it is the statement which is knowable if correct. The notion of propositional truth, whether bivalent or not, is not couched in terms of knowability; propositions are not the objects of (acts of) knowledge.

X

Turning now to the validity of inference, we recall that the premise(s) and conclusion of the completely general inference-figure, inference I, are statements (judgments). Accordingly the appropriate notion of truth to be used here is that of knowability, and the inference I has to preserve knowability from premise(s) to conclusion. Thus, one has to know the conclusion under the assumption that one knows the premise(s). In other words, the conclusion must be made evident, given that the premise(s) have been made evident. This is now where the insights of the containment theory come to aid. All (true, correct, that is) knowable judgments can be made evident, and for some judgments their evidence(ness) rests ultimately upon that of other evident judgments. Certain correct judgments, though, are such that their evidenceability rests upon no other judgments than themselves: these are judgments which are *per se nota*, or analytic in the sense of Kant. The can be known *ex vi terminorum*, in virtue of the concepts out of which they have been formed. Axioms in the original (Euclidean, but not Hilbertian, hypothetico-deductive) sense are examples of this: they can be known but they neither need nor are capable of further demonstration by means of other judgments. In the same way certain inferences are 'immediately' evident upon knowledge of the constituent judgments. Note though that the immediacy is not temporal but conceptual; the inference in question neither can nor needs to be justified in terms of further inferences. The introduction and elimination rules in the natural-deduction systems of Gerhard Gentzen are examples of such immediate inferences.

The validity of an inference is secured by means of the (constructive) existence (= possibility to find) of a chain of immediately evident inferences linking premise(s) and conclusion: the transmission of evidence finds place by means of immediate evident steps that are such that when one knows the premise(s) and understands the conclusion nothing further is needed in order to know the conclusion. The possession of such a chain guarantees that the conclusion can be made evident under the assumption that the premises have been made evident, that is, are known. Mere possession, though, does not suffice for drawing the inference; in order to know the conclusion I must actually have performed the immediate inferences in the chain. Thus, the modern notion of syntactic consequence, under the containment theory of inferential validity, has a counterpart in the chain of immediate inferences that constitutes the ground for the validity of an inference.

Thus, we have found a difference between inferences and consequence: a correct consequence, be it logical or not, preserves truth from antecedent propositions to consequent proposition (possibly under all suitable variations) whereas a valid inference-figure preserves knowability from premise-judgment(s) to conclusion judgment. In particular, the inference Σ:

$$\frac{A \Rightarrow B \text{ holds} \quad A \text{ is true}}{B \text{ is true}}$$

is valid; indeed, the holding (but not the *logical* holding, under all variations) of the sequent $A \Rightarrow B$ is explained in such a way that the inference from the truth of proposition A to the truth of proposition B is then immediate. Thus the holding of sequents is reduced to, or explained in terms of, the validity of inference.

An attempt, on the other hand, to reduce the validity of inference to the (possibly logical) holding of consequences to validity will engage us in an infinite regress of the kind that Lewis Carroll ran for Achilles and the Tortoise in *Mind* 1895. Then the inference

A is true. *Therefore*: B is true

is valid if the sequent $A \Rightarrow B$ holds. But the inference

$A \Rightarrow B$ holds, A is true. *Therefore*: B is true

is certainly valid by the explanation of $\Rightarrow$: $A \Rightarrow B$ holds when B is true if A is true. Thus, by the reduction of validity, the (higher-level!) sequent

$[A \Rightarrow B, A] \Rightarrow B$

must hold. But then the inference

$[A \Rightarrow B, A] \Rightarrow B$ holds, $A \Rightarrow B$ holds, A is true. *Therefore*: B is true

is valid. Thus, by the reduction of inferential validity to that of holding for consequence, the (even) higher-level sequent

$[[A \Rightarrow B, A] \Rightarrow B, A \Rightarrow B, A] \Rightarrow B$

must hold. But then, yet again, a certain inference is valid and so we get a tower of ever higher-level consequences that have to account for the validity of the first inference in question.

The criticisms that have been voiced against Frege's account of inference, on the present analysis, are nugatory. Frege was absolutely right in that inference proceeds from known premises and obtains new knowledge. This is also accounted for by the explanation of validity. The conclusion-judgment must be made evident, given that – under the assumption that – the premises are known. In the (logical) holding of consequence, on the other hand, there is no reference to knowledge: a sequent holds if propositional truth is transmitted from antecedent(s) to consequent. Thus, the criticisms of Frege seem to stem from a conflation of (the validity of) inference with (the holding of) consequence.

Further Reading

Bolzano, Bernard (1837) *Wissenschaftslehre*, I–IV. Sulzbach: J. Seidel. English translation edn. by Jan Berg: *Theory of Science*, Dordrecht: Reidel, 1973 (especially §§ 34, 36, 148 and 155).

Enderton, Herbert (1972) *A Mathematical Introduction to Logic*. New York: Academic Press. (Rigorous, yet accessible, treatment of the model-theoretic notion of consequence from a metamathematical perspective.)

Gentzen, Gerhard (1969) *Collected Papers*, ed. Manfred Szabo. Amsterdam: North-Holland. (Contains the original papers on the sequent calculus and natural deduction systems.)

Monk, Donald (1976) *Mathematical Logic*. Berlin: Springer. (Contains pellucid expositions of model-theoretic consequence. Possibly hard for philosophers.)

Sundholm, Göran (1998) Inference, consequence, implication. *Philosophia Mathematica*, 6, 178–94. (Spells out the present framework in greater detail.)

Tarski, Alfred (1956) *Logic, Semantics, Metamathematics*. Oxford: Clarendon. (Contains his classical papers on truth and consequence.)

16

Modality of Deductively Valid Inference

DALE JACQUETTE

1 Validity and Necessity

An inference is deductively valid if and only if it is logically necessary that if its assumptions are true, then its conclusions are also true; or, alternatively, if and only if it is logically impossible for its assumptions to be true and its conclusions false.

Some type of modality evidently governs the truth conditions of assumptions and conclusions in deductive inference. There are many different systems of alethic modal logic, however, and the question of which modal system is appropriate for understanding the modality of deductive validity has not been rigorously investigated. In what exact sense is it logically necessary for the conclusions of a deductively valid argument to be true if its assumptions are true? In what exact sense it is logically possible for the conclusions of a deductively invalid argument to be false when its assumptions are true? Does deductive inference presuppose the modality of, say, modal system $S1$, or T, $S2$, $S3$, $S4$, $S5$, the Brouwersche system, or yet another modal logic?

I argue in what follows that the failure of the validity or Pseudo-Scotus paradox in normal modal logics weaker than $S5$, and its provability in $S5$ and conservative extensions of $S5$, suggests that the modality of deductively valid inference must be weaker than $S5$. The sense in which it is logically necessary for the conclusions of a deductively valid inference to be true if its assumptions are true, or logically impossible for its assumptions to be true and its conclusions false, in that case must be defined in terms of a modal logic weaker than $S5$.

2 The Validity Paradox

The validity paradox, also known as the Pseudo-Scotus, is most easily understood in an impredicative formulation. Consider the following inference:

(V) 1. Argument (V) is deductively valid.

---

2. Argument (V) is deductively invalid.

The paradox proceeds by projecting argument (V) into a dilemma. We assume that argument (V) is either deductively valid or deductively invalid. If (V) is deductively valid, then it is also sound, since the assumption in (1) declares that the argument is deductively valid. Sound arguments by definition have true conclusions. So, if (V) is deductively valid, then, as its conclusion states, it is deductively invalid. The second horn of the dilemma is more difficult. If (V) is deductively invalid, then, according to the definition of deductive validity, it is logically possible for the assumption of (V) to be true and the conclusion false. The assumption of the second dilemma horn thus implies only that it is logically possible, not categorically true, that argument (V) is deductively valid. It does not follow simply that if argument (V) is deductively invalid, then it is deductively valid, but at most only that if (V) is deductively invalid, then it is logically possible that (V) is deductively valid. To go beyond this, trying to deduce that (V) is deductively valid if and only if it is deductively invalid, is to commit the inelegant modal fallacy of inferring that a proposition is true from the mere logical possibility that it is true (see Jacquette 1996).

3 Gödel Arithmetizing the Validity Paradox

It might be thought that the validity paradox is improper because of its impredicative form, violating the vicious circle principle. The impredicative expression of the validity paradox as presented is nevertheless inessential. Impredication can be avoided by Gödelizing the syntax of the inference.

The validity paradox (V) is Gödelized as (GV) for $g^{\ulcorner}V[sub_g(n)] \vdash \overline{V}[sub_g(n)]^{\urcorner} = n \wedge sub_g(n) = {\ulcorner}V[sub_g(n)] \vdash \overline{V}[sub_g(n)]^{\urcorner}$, in order to prove that $V[sub_g(n)] \leftrightarrow \overline{V}[sub_g(n)]$. The Gödel number of the argument is determined by assigning natural numbers to syntax items in the expression to be arithmetized. Each such number is made the exponent of a corresponding prime number base taken in sequence in the same order of increasing magnitude as the syntax (standardly left-to-right) in the expression to be coded. The Gödel number of the expression is the product of these primes raised to the powers of the corresponding syntax item code numbers. A Gödel substitution function, sub_g, substitutes for any whole number to which it is applied the unique syntax string, if any, which the Gödel number encodes.

V	[	sub_g	(	_	)	]	$\vdash$	$\overline{V}$	[	sub_g	(	_	)	]
|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
1	2	3	4	5	6	7	8	9	2	3	4	5	6	7

The Gödel number of the validity paradox on this assignment of Gödel numbers to syntax items in the formula is: $2^1 \times 3^2 \times 5^3 \times 7^4 \times 11^5 \times 13^6 \times 17^7 \times 19^8 \times 23^9 \times 29^2 \times 31^3 \times 37^4 \times 41^5 \times 43^6 \times 47^7 = n$. This number is substituted for blank spaces (alternatively, free variables) to which the number 5 is here assigned in the open sentence above to complete the Gödel arithmetization in $g^{\ulcorner}V[sub_g(n)] \vdash \overline{V}[sub_g(n)]^{\urcorner} = n$, where by stipulation, $sub_g(n) = {\ulcorner}V[sub_g(n)] \vdash \overline{V}[sub_g(n)]^{\urcorner}$.

Angle quotes, $^{\ulcorner},^{\urcorner}$, are used conventionally to indicate that the Gödel-numbering context is intensional, since a Gödel numbering context does not support intersubsti-

tution of logically equivalent expressions that differ syntactically in any way. A distinct Gödel number obtains for every distinct syntax combination, including logical equivalents, like $\varphi \vee \psi$ and $\neg\varphi \rightarrow \psi$, where $g^{\ulcorner}\varphi \vee \psi^{\urcorner} \neq g^{\ulcorner}\neg\varphi \rightarrow \psi^{\urcorner}$, even though $[\varphi \vee \psi] \leftrightarrow [\neg\varphi \rightarrow \psi]$.

The Fundamental Theorem of Arithmetic guarantees that every number can be decomposed into a unique factorization of prime number bases raised to certain powers. When number n is factored in this way and the factors arranged in ascending order (again, from left to right) according to the increasing magnitude of prime number bases, the expression mapped into Gödel-numbered space can be read directly from the exponents of each prime, and translated back into the original logical syntax by the glossary of natural number assignments.

The Gödelized validity paradox is not impredicative, because the Gödelized paradox argument is not defined in terms of propositions that explicitly mention the argument's label or name, (V). Self-reference is instead achieved only indirectly by the stipulation that the Gödel number of the inference $V[sub_g(n)] \vdash \overline{V}[sub_g(n)]$ is n, and the definition of the Gödel substitution function sub_g, by which the Gödel coded inference is recovered in its exact syntax-item-by-syntax-item formulation. Gödelization avoids impredication in the same way that it circumvents Russell's simple type theory restriction on syntactical self-predications. The Gödel sentence predicates a semantic property only of an object, a substituend identical to the sentence obtained by applying the Gödel substitution function to a Gödel number, and not to another property represented by a predicate of the same type. Gödelization thereby also avoids the need for explicit mention of the name or label of a sentence or argument, achieving self-reference indirectly in the inference by predicating a semantic property, validity or invalidity, of the substituend represented by a Gödel code number defined as the Gödel code number of the inference itself.

4 The Validity Paradox in S5

A proof that the second dilemma horn fails in modal systems weaker than $S5$, but succeeds in modal $S5$ and its conservative extensions, can be formalized in this way for Gödelized validity paradox (GV).

The role of the iterated modalities, and their implications for the second validity paradox dilemma horn, are seen in the following derivation. Here it is obvious that the inference from the assumption that (GV) is invalid to the conclusion that (GV) is valid holds only in some but not all systems of modal logic, according to the world- or model-accessibility relations by which each distinct modal logic is defined.

To symbolize the paradox requires a metalinguistic vocabulary to formally represent specific logical and semantic properties of propositions and inferences. We stipulate as primitive metalogical predicates that A is the property of being an assumption, C the property of being a conclusion, effectively, of an argument. We assume *Truth*, T, as a primitive bivalent relation of positive correspondence between a proposition and an existent state of affairs that the proposition describes or otherwise linguistically represents. If the state of affairs the proposition represents does not exist, then the proposition is false. A state of affairs is the possession of a property by or involvement in a relation of the objects in a well-defined semantic domain; a state of affairs Fa exists

when an object *a* possesses a property or is involved in a relation *F*, and fails to exist when *a* does not possess or is not involved in relation *F*. *Ramsey* reduction then states that for any proposition φ, φ is true if and only if φ, $\forall\varphi[T\varphi \leftrightarrow \varphi]$. The principle effects what is sometimes known also as the redundancy theory of truth, where to say that φ is true is just to say that φ, and to say that φ is to say that φ is true. The principle allows us to move freely back and forth from true propositions to true metalinguistic propositions that state that the propositions are true.

Validity, *V*, is defined as a relation among the truth conditions of the assumptions and conclusions of an inference, such that it is logically impossible for the assumptions to be true and the conclusions false. The truth of logically necessary propositions is invoked in step (7) as $\Box\varphi \rightarrow \varphi$. The formalism reflects the intuitive reasoning that if (*GV*) is valid, then it is also sound, since its assumption says that it is valid. But, as we have seen, since sound arguments necessarily have true conclusions, it follows in that case that (*GV*), as its conclusion states, is invalid.

PROOF 1 *Validity Horn of the Validity Paradox*

(1)	$\forall x[Vx \leftrightarrow \Box[\forall y[Ayx \wedge Ty] \rightarrow \forall y[Cyx \rightarrow Ty]]]$	*Validity*
(2)	$V[GV]$	*Assumption*
(3)	$\Box[[\forall y[\mathrm{A}y[GV]] \wedge Ty] \rightarrow \forall y[Cy[GV] \rightarrow Ty]]$	(1,2)
(4)	$\forall y[TAy[GV]] \leftrightarrow V[GV]$	(*GV*)
(5)	$\forall y[TCy[GV]] \leftrightarrow \overline{V}[GV]$	(*GV*)
(6)	$\Box[TV[GV] \rightarrow T\overline{V}[GV]]$	(3,4,5)
(7)	$TV[GV] \rightarrow T\overline{V}[GV]$	(6, $\Box\varphi \rightarrow \varphi$)
(8)	$V[GV] \rightarrow \overline{V}[GV]$	(7, *Ramsey*)

The second dilemma horn is more difficult. It is blocked by modal fallacy, except where the accessibility relations defining a strong system of modality like *S*5 or its conservative extensions make it possible to infer necessity from possible necessity. To demonstrate the difference in strengths of modalities in deriving the inference that $\overline{V}[GV] \rightarrow V[GV]$, we first show that the inference fails in weak modal systems, and then offer a formal proof of the second paradox dilemma horn invoking the characteristic axiom of modal *S*5. This is how the proof is blocked in weak systems of modality:

PROOF 2 *Failure of Invalidity Horn of Validity Paradox in Modal Systems Weaker than S*5

(1)	$\forall x[Vx \leftrightarrow \Box[\forall y[Ayx] \wedge Ty \rightarrow [\forall y[Cyx \rightarrow Ty]]]]$	*Validity*
(2)	$\forall x[\overline{V}x \leftrightarrow \Diamond[\forall y[Ayx] \wedge Ty \wedge \exists y[Cyx \wedge \overline{T}y]]]$	(1)
(3)	$\overline{V}[GV]$	*Assumption*
(4)	$\Diamond[\forall y[Ay[GV]] \wedge Ty \wedge \exists y[Cy[GV] \wedge \overline{T}y]]$	(2,3)
(5)	$\forall y[TAy[GV]] \leftrightarrow V[GV]$	(*GV*)
(6)	$\forall y[TCy[GV]] \leftrightarrow \overline{V}[GV]$	(*GV*)
(7)	$\Diamond[TV[GV] \wedge \overline{T}\,\overline{V}[GV]]$	(4,5,6)
(8)	$\Diamond V[GV]$	(7, *Ramsey*)
(9)	$\overline{V}[GV] \rightarrow \Diamond V[GV]$	(3–8)

The conclusion falls short of the second horn of the validity paradox in the categorical form, $\overline{V}[GV] \rightarrow V[GV]$, and thereby of the entire validity paradox, $\overline{V}[GV] \leftrightarrow V[GV]$. The mere logical possibility of the deductive validity of (*GV*) is all that is validly derivable

from the assumption that (GV) is deductively invalid, if the modality of deductive inference is weaker than $S5$.

By contrast, we now see how the proof goes through in modal system $S5$ and its conservative extensions. The proof depends on the principle that for any inference φ, $\Box(\varphi \rightarrow \Box\varphi)$, invoked at step (9), according to which it is logically necessary that if an argument is deductively valid, then it is logically necessarily valid, or valid in every logically possible world. The intuitive justification is that the same abstract set of propositions, true or false, for states of affairs that are realized or unrealized in any logically possible world, is ideally available for combination into all the same arguments, and the same logical laws of valid deductive inference standardly prevail, in every logically possible world. The first unproblematic half of the paradox, that $V[GV] \rightarrow \overline{V}\,[GV]$, is recalled without further ado as the conclusion of *Proof 1*, in step (20). We also appeal to weak standard principles of *Necessitation*, $\Box[\varphi \rightarrow \psi] \rightarrow [\Box\varphi \rightarrow \Box\psi]$, in step (10), and *Duality*, $\Diamond\varphi \leftrightarrow \neg\Box\neg\varphi$, in step (13). The proof hinges essentially on the characteristic axiom of modal $S5$, $\Diamond\Box\varphi \rightarrow \Box\varphi$, introduced in step (16).

PROOF 3 *Invalidity Horn of the Validity Paradox in S5*

(1)	$\forall x[Vx \leftrightarrow \Box[\forall y[Ayx \wedge Ty] \rightarrow [\forall y[Cyx \rightarrow Ty]]]]$	*Validity*
(2)	$\forall x[\overline{V}x \leftrightarrow \Diamond[\forall y[Ayx \wedge Ty] \wedge \exists y[Cyx \wedge \overline{T}\,y]]]$	(1)
(3)	$\overline{V}\,[GV]$	*Assumption*
(4)	$\Diamond[\forall y[Ay[GV] \wedge Ty] \wedge \exists y[Cy[GV] \wedge \overline{T}\,y]]$	(2,3)
(5)	$\Diamond[\exists y[Cy[GV]] \wedge \overline{T}\,y]$	(4)
(6)	$\forall y[TCy[GV]] \leftrightarrow \overline{V}\,[GV]$	(GV)
(7)	$\Diamond\,\overline{T}\,\overline{V}\,[GV]$	(5,6)
(8)	$\Diamond V[GV]$	(7, *Ramsey*)
(9)	$\Box[V[GV] \rightarrow \Box V[GV]]$	$\Box(\varphi \rightarrow \Box\varphi)$
(10)	$\Box[\varphi \rightarrow \psi] \rightarrow [\Box\varphi \rightarrow \Box\psi]$	*Necessitation*
(11)	$\Box[\neg\Box V[GV] \rightarrow \neg V[GV]] \rightarrow [\Box\neg\Box V[GV] \rightarrow \Box\neg V[GV]]$	(10)
(12)	$\Box[V[GV] \rightarrow \Box V[GV]] \rightarrow [\neg\Box\neg V[GV] \rightarrow \neg\Box\neg\Box V[GV]]$	(11)
(13)	$\Box[V[GV] \rightarrow \Box V[GV]] \rightarrow [\Diamond V[GV] \rightarrow \Diamond\Box V[GV]]$	(12, *Duality*)
(14)	$\Diamond V[GV] \rightarrow \Diamond\Box V[GV]$	(9,13)
(15)	$\Diamond\Box V[GV]$	(8,14)
(16)	$\Diamond\Box V[GV] \rightarrow \Box V[GV]$	($S5$)
(17)	$\Box V[GV]$	(15,16)
(18)	$V[GV]$	(17, $\Box\varphi \rightarrow \varphi$)
(19)	$\overline{V}\,[GV] \rightarrow V[GV]$	(3–18)
(20)	$V[GV] \rightarrow \overline{V}\,[GV]$	(*Proof 1*)
(21)	$V[GV] \leftrightarrow \overline{V}\,[GV]$	(19,20)

5 Validity, Necessity, and Deductive Inference

The validity paradox can only be avoided by disallowing formulations of the modality governing the logical necessity of deductively valid inference as strong as or stronger than $S5$. The fact that the validity paradox goes through in modal $S5$ and stronger logics, but not in weaker systems, suggests that the modality of deductive inference, on

pain of contradiction in the derivation of inferences that are deductively valid if and only if they are deductively invalid, must be weaker than *S*5. Needless to say, the status of deductively valid inference in *S*5 is also thereby placed in doubt.

If *S*5 itself is redefined to embody a sufficiently nonstandard model of deductively valid inference that avoids the validity paradox, then it might be possible to interpret the modality of deductively valid inference in terms of such an appropriately nonstandard *S*5. The defender of *S*5 as the modality of deductive validity nevertheless cannot reasonably appeal to the intuition that a deductively valid inference accessible from the actual world ought to be deductively valid in every logically possible world accessible from any logically possible world. An equivalence relation for accessibility provided for the model set theoretical semantics of *S*5, involving reflexivity, symmetry and transitivity, must be adequate even for deductively valid inferences involving modal structures in which not all models contain all the same objects. It must be adequate, indeed, for deductively valid inference in any modal environment weaker than *S*5, and so, by the same reasoning, presumably, weaker than *S*4, and so on, down to the weakest modal logic. The conclusion to which the provability of the validity paradox in *S*5 ultimately points is that the modality of deductively valid inference in general cannot be stronger than that formalized by the weakest modal system interpreted only as reflexive world-accessibility.

References

Bendiek, Johannes (1952) Die Lehre von den Konsequenzen bei Pseudo-Scotus. *Franziskanische Studien*, 34, 205–34.

Bocheński, J. M. (1937) Notes historiques sur les propositiones modales. *Revue des Sciences Philosophiques et Théologiques*, 26, 673–99.

Bocheński, J. M. (1938) De consequentiis scholasticorum earumque origine. *Angelicum*, 15, 92–109.

Gödel, Kurt (1931) On formally undecidable propositions of *Principia Mathematica* and related systems I ["Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme I," *Monatshefte für Mathematik und Physik*, 38, 1931], translated by Jean van Heijenoort (ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Cambridge, MA: Harvard University Press, 1967, 596–617.

Jacquette, Dale (1996) The validity paradox in modal S_5. *Synthese*, 109, 47–62.

Keene, G. B. (1983) Self-referent inference and the liar paradox. *Mind*, 92, 430–3.

Mates, Benson (1965) Pseudo-Scotus on the soundness of *Consequentiae*. In A. T. Tymieniecka (ed.), *Contributions to Logic and Methodology in Honor of J. M. Bocheński*. Amsterdam: North-Holland, 132–41.

McDermott, A. Charlene Senape (1972) Notes on the assertoric and modal propositional logic of the Pseudo-Scotus. *Journal of the History of Philosophy*, 10, 273–306.

Pseudo-Scotus, *In Librum Primum Priorum Analyticorum Aristotelis Quaestiones*, Question 10, Duns Scotus, *Ioannis Duns Scoti Opera Omnia*, ed. Luke Wadding [1639]. Paris: Vives, 1891–5, vol. II.

Ramsey, F. P. (1931) Facts and Propositions. In R. B. Braithwaite (ed.), *The Foundations of Mathematics and Other Logical Essays*. London: Routledge & Kegan Paul, 138–55.

Read, Stephen (1979) Self-reference and validity. *Synthese*, 42, 265–74.

Sorensen, Roy A. (1988) *Blindspots*. Oxford: Clarendon Press.

This page intentionally left blank

Part VI

LOGIC, EXISTENCE, AND ONTOLOGY

This page intentionally left blank

17

Quantifiers, Being, and Canonical Notation

PAUL GOCHET

1 Introduction

Aristotle was the founder of logic and ontology. The first discipline is concerned with the validity of arguments irrespective of their subject-matter. Its foundations were laid down in the *Prior Analytics*. Topic neutrality is achieved by abstracting the *form* of the arguments from their *content*, an operation which presupposes that we draw a distinction between the logical terms which make up the form ('every M is L,' 'some M is L,' 'A possibly belongs to no B' . . .) and the non-logical terms which belong to the content.

The second discipline, called 'first philosophy' by Aristotle (and 'ontologia' by Rudolphus Goclenius in the *Lexicum Philosophicum* (1613)) investigates being in its own right, that is the categorial aspects of entities in general, and the modes and aspects of being. It can be traced back to Aristotle's *Categories* and *Metaphysics*.

The third notion occurring in the title is central both to logic and to ontology. Indeed the question arises whether *existence* should be distinguished from *being*. For example, in *Principles of Mathematics* (1903), Russell claims that such a distinction is in fact presupposed by any denial of existence: "what does not exist must *be* something, or it would be meaningless to deny its existence" (Russell 1903: 450).

The interplay between logic and ontology has inspired major philosophical works of the twentieth century such as Russell's *Philosophy of Logical Atomism* (1918) and Wittgenstein's *Tractatus logico-philosophicus* (1921). Though both works now belong to the history of the subject, the issue they address, that is whether a logical language could be designed which would depict the main ontological structures of reality, remains a live issue.

With Quine's *Word and Object* (1960), a major shift of emphasis occurred. The mirror of the most important traits of reality is no longer to be sought in *language* as such, but in the *theories* about the world which scientists hold to be true, and only derivatively in the language needed to formulate them.

According to Quine, the ontological work incumbent on philosophers consists of the critical scrutiny of the realm of objects introduced into scientific theories by scientists. It is "the task of making explicit what had been tacit, and precise what had been vague; of exposing and resolving paradoxes, smoothing kinds, lopping off vestigial growths, clearing ontological slums" (Quine 1960: 274).

Logic plays a major role in the work of attaining precision and explicitness just described. The time has come to take stock of what has been achieved over the last 40 years by applying logic to ontology. Although my concern is thematic rather than historical, I shall devote much space to a detailed presentation and examination of Quine's views on the interplay between logic, existence, and ontology.

The motivation for my choice lies in the influential and challenging character of Quine's theses. I shall try to isolate what I consider to be of lasting value in his doctrines. I shall also describe and critically examine the arguments offered by opponents to Quine who claim that his logic is too restricted and his ontology too poor.

2 A Methodology for Ontology

For the philosopher who undertakes to clean up the conceptual framework built by the scientist and to purify it of unnecessary ontological excrescences, Ockham's razor, "*Entia non sunt multiplicanda praeter necessitatem*" is the main tool. To apply that precept, however, we have to answer the preliminary question: 'what are unnecessary entities?' One possible answer is: entities are unnecessary if we can abstain from countenancing them without sacrificing *scientific truth*.

That answer is controversial. One might argue that besides preserving the *set of truths* of a given science, we should also be concerned about preserving the *explanatory power* of our theories. One burning issue here is the question raised by the status of natural kinds and natural kind words. Kripke and Putnam have argued that natural kind words are rigid designators (Putnam 1975: 229–35). The very definition to the concept of rigid designator as "term which designates the same entities in our world and in all possible worlds" draws us willy-nilly into possible world semantics.

Quine has also contributed to the *methodology of ontology* by imposing a constraint encapsulated in the motto: "No entity without identity" (see Haack 1978: chapter 4). Such a requirement is fulfilled by *sets*: two sets are identical if and only if they have the same members. It is not fulfilled, however, by the entities of linguistic semantics such as concepts and propositions (for a defense of the latter see Orilia 1999).

The demand for clear *identification criteria* has far-reaching consequences in ontology. It has a bearing on another burning issue under discussion today: that of the status of possible objects. By Quine's standards, possible objects are not eligible as *entities*. They lack criteria of identification. Nobody, Quine complains, can decide whether "the possible fat man in that doorway" and "the possible bald man in that doorway" denote the same individual (Quine 1953, 1961: 4). (For another diagnosis of this puzzle, see Cocchiarella 1987: 126 f.).

Fifteen years after Quine first published "On What There Is," Kripke (1963) laid down a semantics which extends the standard definitions of satisfaction and truth to a first-order logic enriched with modal operators (see also Bayart 1958, 1959). The novelty of this approach lies in the model which contains a set of possible worlds together with an accessibility relation between worlds. The *domains* are allowed to vary from one world to another. An individual a which shows up in the domain D_1 of possible world W_1 may be absent from the domain D_2 of possible world W_2. That individual may also be present, but then the question of identifying a across possible worlds arises.

Quine argues that identifying individuals across possible worlds fundamentally differs from the familiar task of reidentifying an individual across successive moments of time. In the latter case, relevant criteria are available such as, if physical objects are concerned, continuity of displacement, continuity of deformation and continuity of chemical change. These criteria, however, cannot be extended across worlds "because you can change anything to anything by easy stages through some connecting series of possible worlds" (Quine 1981: 127).

Here again the problem is worth reconsidering in the light of recent developments. Several authors (Gupta 1980; Cocchiarella 1984) have provided evidence showing that the contrast between identification across moments of time and identification across possible worlds is not so sharp as Quine contends.

3 The Need for a Criterion of Ontological Commitment

The history of philosophy is replete with discussions about abstract objects. Plato held that *Forms*, such as Beauty, existed independently of the mind which conceived them and of the particular objects in which they were exemplified. For Aristotle, however, species differed from their instances but existed only in so far as they were instantiated by the latter.

In the Middle Ages, the distinction between concrete and abstract objects prompted a lasting discussion known as the *debate on universals*. A broad spectrum of positions were defended, ranging from realism to nominalism. According to the latter, universals are just words. The question has yet to be conclusively resolved. In the twentieth century, Church diagnosed the source of the trouble in these terms: "No discussion of an ontological question . . . can be regarded as intelligible unless it has a definite criterion of ontological commitment" (Church 1958: 1012).

Quine came to grips with the problem and provided a definite criterion: "[i]n general, *entities of a given sort are assumed by a theory if and only if some of them must be counted among the values of the variables in order that the statements affirmed in the theory be true*" (Quine 1953, 1961: 103).

Quine's criterion is informative. It serves to uncover *hidden* ontological commitments. Consider the following sentence due to Geach: "Some people admire only one another" in which the number of mutual admirers remains unspecified. Kaplan, has shown that Geach's sentence implicitly quantifies over classes. Its formulation in first order logic reads as follows (Quine 1982: 293):

$$\exists z(\exists x(x \in z).\&\forall x([x \in z \rightarrow \exists y(x \text{ admires } y)\&\forall y(x \text{ admires } y \rightarrow x \neq y\&y \in z)])).$$

When combined with his views about predicates, Quine's criterion of ontological commitment ceases to be neutral. In *Philosophy of Logic*, Quine writes "Predicates are not names, predicates are the other parties to predication" (Quine 1970: 27–8). This syntactic consideration leads to ban second order logic statements such as $\exists F \forall x\, Fx$ and forces us to rewrite them in first order logic as $\exists \alpha \forall x\, x \in \alpha$. This is not satisfying however. As Boolos observes, the first formula is valid but the second is not (Boolos 1975: 512).

Simons disentangled the two issues. He showed that we can quantify over variables belonging to the syntactical category of predicates without committing ourselves to say that predicates refer to properties. A restriction should be imposed upon Quine's criterion of ontological commitment. Not all quantification is committal: "nominal quantification commits one to things denotable because names denote, while other forms of quantification do not, since it is the office of names, and names alone, to denote, other categories of expression having other offices, the variables of these categories inheriting their offices from potential constants thereof" (Simons 1997: 268).

Cocchiarella criticizes Quine for assuming that being is a *genus*. Quine's criterion does justice to primary substances and complete (saturated) objects but fails to do justice to universals. Universals, Cocchiarella argues, have a *predicable nature* that constitutes their universality. That predicable nature consists of a *mode of being* different from the mode of being of saturated objects. Universals, unlike sets, are not generated by their instances.

According to Cocchiarella, we need *predicate variables* taking universals as their values if we want to represent not only saturated but also unsaturated entities in our formal ontology. If, following Quine, we take predicate variables as *schematic letters* which admit substitution but not quantification, we shall not be able to quantify over unsaturated entities such as natural properties and relations. Yet such a quantification is needed in the construction of a formal ontology for natural science (see Section 7).

To capture the ontological distinction between individuals and universals, we have to give predication precedence over membership and to recognize an ontological import to predicates as such (Cocchiarella 1997).

4 The Role of a Canonical Notation

According to Quine, ontologists should not address the *direct question* 'What objects are there?' Quine proposes a *detour* through existing scientific theories. Ontologists would start with a given theory and ask themselves what objects it is committed to. He coined the locution "semantic ascent" for referring to this shift of attention from the world to theories and their languages.

Positive knowledge about the world is not confined to specialized sciences only. Common sense knowledge expressed in everyday language is also knowledge. If we want to spot the ontological commitments of our knowledge as a whole, a preliminary task need to be performed. We have to *regiment* our language into a *canonical system of logical notation*.

Several sections of *Word and Object* show how constructions of ordinary language can be paraphrased into the artificial language of first-order logic. Some of these regimentation exercises are known to whoever has learned to translate arguments couched in natural language into the inferential schemes of standard first-order logic. For instance, 'Every man is mortal' is paraphrased into 'For every object x (if x is a man then x is mortal).' More drastic changes come next, such as the elimination of proper names and the elimination of definite descriptions. These are specifically Quinean doctrines.

Indirect discourse, however useful it may be for historians, has a major drawback. It violates "the substitutivity of identity: the putting of equals for equals" (Quine 1994b: 145). In the propositional attitude construction: 'Ralph believes that Cicero denounced Catiline,' the substitution of 'Tully' for 'Cicero' may fail to preserve truth. To prevent the unsafe substitution, Quine suggests a radical remedy: replacing indirect quotation by direct quotation.

Far from distorting our picture of the world, such regimentation would help us see the world aright. If we are 'limning the true and ultimate structure of reality,' Quine maintains, the canonical scheme that suits us is "the austere scheme that knows no quotation but direct quotation and no propositional attitudes but only the physical constitution and behavior of organisms" (Quine 1960: 221).

5 The Ontology of Mathematics

Quine's *New Foundations for Mathematical Logic* (1936) contains some technical innovations which are philosophically significant. The first one is the notion of stratification. A formula is called *stratified* if it is possible "to put numerals for the variables in such a way that 'ε' comes to occur only in contexts of the form '$n \varepsilon n+1$'" (Quine 1953, 1961: 91). Stratified formulas satisfy Russell's *type theory* (1908). Unstratified formulas would have to be declared meaningless by Russell's standards.

For Quine, on the contrary, unstratified formulas such as '$y \varepsilon y$' are meaningful, but they are not eligible as instances of F in the comprehension axiom $(\exists x)\,(\forall y)\,(y \varepsilon x \leftrightarrow F)$. Hence a formula can be meaningful *without carrying any ontological commitment*.

Russell's type theory has forbidding ontological consequences: the universal class V gives rise to an *infinite series* of quasi-universal classes. The null class also. The Boolean class algebra "no longer applies to classes in general, but is reproduced within each type" (Quine 1953, 1961: 92). The same is true of arithmetics. All that *ontological inflation* would be cut down in one stroke by adopting the stratification theory of *New Foundations*.

Stratification theory substitutes a *syntactic hierarchy* of formulas for the *ontological hierarchy* of types of entities. It switches from the multilayered universe of objects to a single universe of objects, with a general quantifier ranging over all the objects in the universe. As Vidal-Rosset puts it, the syntactic device of stratification "frees set theory from the realist assumption of types in the same way *free logic* purifies standard first-order logic of its ontological commitments." The claim that the existence of an infinite set is a *theorem*, rather than a *postulate*, is another achievement of *NF*. That startling thesis has been demonstrated later by Specker (1953) and Crabbé (1984).

Let us now move on to *set theory and its logic* and consider the theory of virtual classes. As a preparatory step, the reader should remember here that the grammar of first-order logic admits three basic constructions: (1) predication; (2) infixation or prefixation of connectives; and (3) quantification over individual variables. Predication unites a name with a predicate. Names refer to individuals. Predicates do not *refer* to classes or properties, they are *satisfied* by individuals.

Quine's statement already quoted "Predicates are not names; predicates are the other parties to predication" (Quine 1970: 27–8) might strike the reader as dogmatic.

It should not. A justification of this statement can be found later in the book. (We owe the point to Fernandez de Castro). Quine observes that quantifying over predicate variables leads to an unconstrained principle of comprehension. From the logical triviality $(\forall x)(Fx \leftrightarrow Fx)$ we can derive the unwanted conclusion $(\exists x)(Gx \leftrightarrow Fx)$ (Quine 1970: 68).

If we wish to *refer* to a class we need a class abstract, that is an expression like $\{x: Fx\}$ which can be rendered in natural language by 'the set of x that are F.' Whenever a class-abstract occurs only on the right of 'ε' we can treat the whole combination '$\varepsilon\{x: Fx\}$' as 'F' and say that '$y\ \varepsilon\{x: Fx\}$' reduces to 'Fy'. Conversely we may jointly introduce the membership symbol and the class abstract as *fragments of a predicate*. Most of what is said of classes with the help of the two-place predicate 'ε' can then be considered as *a mere manner of speaking* involving no reference to classes, that is no *ontological commitment to classes*.

Set theory and its logic offers a new definition of natural numbers which again enables the mathematician to reduce his ontological commitment without impoverishing science. Let us start with Frege's definition. Natural numbers are the common members of all classes z such that 0 is a member of z and all successors of members of z are members of z. Notice that the unavoidable quantification over classes makes the virtual theory of classes inapplicable here.

If the Fregean definition of natural numbers is to achieve its purpose, infinite classes are required. Quine, however, succeeded in showing that the need for *infinite classes* can be circumvented. We can define numbers in terms of their *predecessors*. This amounts to describing natural numbers as the members of all classes z which contain 0 if, besides containing their members, they also contain the predecessors of their members. For the new definition to work, "there are going to have to be larger and larger classes without end . . . but they can all be finite" (Quine 1963: 76). This meager basis should be enough for deriving the law of mathematical induction.

When put into an epistemological setting, Quine's ontology for mathematics shades into the structuralist position advocated in *Mathematics as a Science of Patterns* (Resnik 1997). Resnik's position is foreshadowed by Quine in the following statement: "what matters for any objects, concrete or abstract, is not what they are but what they contribute to our overall theory of the world as neutral nodes in its logical structure" (Quine 1995: 74–5).

The adoption of a structuralist ontology in which all that there is to an object is the role that it plays in theory is compatible with realism. As Hylton observes, "there is no issue concerning realism about objects which is separate from the issue of realism about the theory which mentions them: to repeat, ontology is derivative upon truth; hence, if we are realists about truth we are more or less automatically realists about objects too" (Hylton 2000: 298).

6 The Notion of Existence

Non-denoting singular terms such as 'Pegasus' have unwanted consequences for standard logic. From the logical truth '$(\forall x)(x = x)$' we obtain 'Pegasus = Pegasus' by the law of universal instantiation. Applying the rule of existential generalization next, we

derive the statement '$(\exists x)(x$ = Pegasus).' A factual falsity has been inferred from a logical truth. Clearly there is something amiss here.

Three solutions have been put forward. The most drastic one consists of first replacing proper names by definite descriptions ('Pegasus' becomes 'the unique object that pegasizes') which are eliminated by Russell's technique at a later stage. The trouble is that standard description theory, as opposed to *free description theory* (Lambert 1987), has unwanted consequences. It leads to paraphrasing a true sentence such as "Theory T is ontologically committed to the perpetual motion machine" into the false one "There is one and only one perpetual motion machine and theory T is committed to it" (Jacquette 1996: 56–69).

The second solution consists of modifying the laws of first-order logic in such a way that it becomes free of existence assumptions with respect to singular terms. Hintikka (1959) produced a *free logic* by submitting the application of the rule of existential generalization $f(a/x) \vdash (\exists x)fx$ to a condition: the truth of the premise $(\exists x)(x = a)$ which states that a exists.

The third solution consists of treating denotationless singular terms as denoting nonexistent objects and taking bound variables as ranging over objects which are either existent or nonexistent. On that account the use of a bound variable is noncommittal. The task of expressing existence devolves to a special predicate, the predicate 'exists' (see Section 10).

A variant of the third approach can be found in a version of first-order logic which operates with two pairs of quantifiers, viz (1) $\forall_a$ and $\exists_a$ which bind variables ranging over *existent* ('*actual*') *individuals* and (2) $\forall$ and $\exists$ which bind variables ranging over *possible individuals*. Distinct rules apply to possible and actual quantifiers. Whereas the law of universal instantiation $\forall x\phi \rightarrow \phi(\zeta/x)$ is logically true for the possible quantifiers without qualification, it holds for the actual quantifiers only on the proviso that an existential premise is supplied, premise which is false when the singular term is denotationless. For actual quantifiers the law of universal quantification reads as follows: $\exists_a y(\zeta = y) \rightarrow [ax\ \phi \rightarrow \phi(\zeta/x)]$ (Cocchiarella 1990: 245).

7 The Ontology of Natural Sciences

According to Cocchiarella, the ontology of physics requires objects which blur the sharp distinction drawn by Quine between objects located in time and objects located in possible worlds. A first motivation for countenancing objects which transcend the *realia–possibilia* dichotomy arises within the framework of the theory of special relativity. There can be objects, the theory says, that exist only in the past or future of our own local time, but which however "might exist in a causally connected local time at a moment which is simultaneous with our present" (Cocchiarella 1984: 351).

These things are real, even if not presently existing. Hence they are entitled to be called *realia* instead of *possibilia*. They qualify as values of our bound variables. Cocchiarella claims that a canonical notation reduced to standard first-order logic has not enough *expressive power*. We need to enrich the language with two *causal tense operators, viz.* 'P_c' for 'it causally was the case that' and 'F_c' for 'it causally will be the case that,' and to add the axioms and rules of quantified modal logic S_4.

Next, Cocchiarella spells out a semantics in which the *accessibility relation* between possible worlds appears in the guise of a *signal relation* linking together momentary states of the universe. Here again we see that a physicalistic interpretation can be grafted onto the suspect notions of the semantics of modal logic and that the gap between modality and time can be bridged.

Transuranic elements provide us with a second sort of entity which stand on the border between the *possible* and the *real*. When the formation of the earth was completed, "it contained the atoms of only ninety-two chemical elements, with uranium being the heaviest" (Cocchiarella 1986: 119). The question whether the universe outside of the earth contains atoms of transuranic elements is an open question. Whether these atoms exist or not, their elements as natural kinds are known so well that atoms of those elements have been produced in accelerators. We have, therefore, to reckon with transuranic substances that "as a matter of contingent fact, are and will never be realized in nature by any objects whatever, but which, as a matter of natural or causal possibility, could be realized" (Cocchiarella 1996: 45).

Aristotle held the view that universals such as the *ultima species* Man exist only in so far as there are concrete human beings that instantiate them (Moderate Realism). Transuranic substances which are not instantiated in concrete objects nevertheless belong to the *causal matrix* of the universe. 'Belonging to the causal matrix of the universe' has to be understood *analogically*. Just as some modes of being in Aristotle's system of categories must be understood 'analogically' (we owe this point to Cocchiarella).

To accommodate these transuranic substances, we need to relax Aristotle's Moderate Realism a little bit and replace 'instantiate' by '*can* instantiate.' To express this conceptual shift, we have to avail ourselves of the modal operator of *causal realizability*, *viz.* $<>_c$. The fundamental thesis of modal natural realism is stated in this way:

$$(\forall F) <>_c (\exists_1 x) \ldots (\exists_j x) F(x_1 \ldots x_j)$$

The colloquial rendering of the formula reads as follows: 'for all n-place predicates it is causally possible that there exists a n-tuple of concrete objects which exemplifies it.' Quine finds quantification over predicates objectionable. Predicates, he insists, are not referring expressions. However, we can recast Cocchiarella's formal representation of modal realism in a way which complies with Quine's requirement. It suffices to replace the predicate variable by an individual variable (ranging over sets) and to bestow the role of predicate to the *set-membership* predicate.

$$(\forall K) <>_c (\exists <x_1 \ldots x_j>) <x_1 \ldots x_j> \varepsilon K$$

The predicate variable has been replaced by an individual variable K which takes natural kinds as values. The colloquial rendering is now: "for natural kinds K, it is causally possible that there exists a n-tuple of concrete individuals that is member of K." Admittedly Quine has misgivings about natural kinds which he takes to be *vestigial growths*. Yet natural kinds satisfy the requirement of *extensionality*. Kinds "can be seen as sets, determined by their members" (Quine 1969: 118). Hence my departure from Quine's standards is minimal.

The distinction between *natural kinds* and *conventional groupings*, just like the distinction between *lawlike statements* and *accidental generalizations*, however elusive it may be, is an essential ingredient of the standard account of *science*. As Peirce observes, prediction would be impossible and induction baseless if there were no genuine laws; and there would be no law if there were no real kinds (Haack 1992: 25).

8 Do Intensions Belong to the Furniture of the World?

I shall now consider a new argument put forward to support a much more dramatic revision in ontological theory than the latter two. In *Rethinking Identity and Metaphysics*, Hill challenges Quine's extensionalist ontology and writes: "Intensions are part of the ultimate furniture of the universe," and "in limning the true and ultimate structure of reality intensions must be given their due" (Hill 1997: 120). Even the description of the mechanisms at work in a successful transplantation of organs requires that we appeal to intensional notions.

Consider a man who donates a kidney to his twin brother. We can reconstruct the reasoning of the surgeon along the following lines: whenever transplantation occurs between twin brothers, the recipient's immune system 'thinks' the donor's kidney x to be sufficiently like diseased kidney y not to reject x as foreign. Hence "x can be substituted for y, though they are not the same" (Hill 1997: 120).

One might object, however, that the *physical exchange* of kidneys and the *logical substitution* of terms are altogether different things which should be kept separate. One might also question the claim that we are forced to make use of a non-mentalistic use of 'belief' in the description of the behavior of the immune system.

Alternative descriptions are available which do not rest upon the dubious notion of the 'body's belief.' Let us pay heed to the following dissymmetry: although the same causes always have the same effects, the same effects do not always have the same causes. If we bring it to bear on the issue, we can see the immune system's behavior as a case in which *different causes* produce the *same effects*.

In *Matter and Memory* (1929), Bergson considered two rival descriptive accounts of the same chemical process. The first one used psychological terms, the second one used physical terms. Bergson chose the second. Here are the scientific data: hydrochloric acid always acts in the same way upon carbonate of lime – whether in the form of marble or of chalk. We might therefore be tempted to say that the acid *perceives* in the various species (marble, chalk) the characteristic of a *genus*. Bergson took the other option and said that "similarity . . . acts objectively like a force." In a similar vein, I suggest that we should favour the description which does not make use of the notion of 'body' belief.'

9 How to Treat Intensional Contexts without Positing Intensions

Frege holds that when we embed a sentence such as 'Cicero denounced Catiline' into a construction like 'Ralph believes that . . . ,' a *shift of reference* occurs in the embedded sentence. The names now refer to whatever their customary sense was when they

occurred in the independent clause. This shift is meant to explain why substituting 'Tully' for 'Cicero' in a belief construction may fail to preserve truth.

Frege's appeal to semantic deviance prompted Davidson's comment: "If we could recover our pre-Fregean semantic innocence, I think it would seem to us plainly incredible that the words 'The earth moves,' uttered after the words 'Galileo said that,' mean anything different, or refer to anything else, than is their wont when they come in different environments" (Davidson 1968: 144).

Frege's account compels us to say that in the sentence 'Cicero denounced Catiline and Ralph believes that Cicero denounced Catiline,' the first occurrence of 'Cicero' (and 'of Catiline') does not have the same referent as the second one. The *arbitrarily created ambiguity* precludes the derivation of the statement '($\exists x$) (x denounced Catiline and Ralph believes that x denounced Catiline).'

Following Recanati's (2000) lead, I shall argue that most of the facts which Frege tries to account for in *semantic terms*, by positing intensional entities, can be dealt with in *pragmatic terms* by carefully distinguishing the perspective of the *ascriber* of propositional entities from that of the *ascribee*. Making appropriate use of the ascriber–ascribee contrast would require us to shift from what might be described as the *ascriber's 'world'* to the *ascribee's 'world'*, but the *ontology* would remain that of the ascriber all along, that is the singular terms would refer to the same objects, whether we were talking about the actual world or about the ascribee's belief world.

First we should stress that the problems raised by propositional attitudes are much more complex than philosophers thought. As Recanati shows, three preliminary distinctions must be drawn if we want to do justice to the complexity of the data. First we should distinguish between (1a) *descriptive phrases* (such as 'The President') and *quantified phrases* (such as 'someone') on the one hand and (1b) *proper names* (such as 'Cicero') on the other. Definite descriptions and quantifiers induce *scope ambiguities*: 'Someone will be in danger' does not have the same truth-conditions as 'It will be the case that someone is in danger.' Names do not induce scope ambiguities: 'Cicero will be in danger' has the same truth-conditions as 'It will be the case that Cicero is in danger.'

Belief sentences with descriptive or quantified phrases, Recanati observes, are ambiguous in a way that exactly parallels the ambiguities found in temporal sentences with descriptive or quantified phrases. John believes that someone is a spy admits of two readings. If 'someone' takes wide scope, we obtain (2a) the *relational reading* of 'believes,' to use Quine's terminology. The sentence says: 'Someone is such that John believes of him that he is a spy.' If 'someone' takes the narrow scope, we obtain (2b) the *notional reading*. The sentence now reads: 'John believes that there are spies.'

When belief is relational, the ascriber and the ascribee refer to the same singular object. When belief is notional, on the contrary, quantification is internal to the ascribed content and not endorsed by the speaker. The ascriber makes no ontological commitment. Believing in that sense has neither a *converse* nor a *relatum*. Hence the exportation: 'John believes there are spies therefore there are people John believes to be spies' is invalid.

The distinction between *relational* and *notional* readings of the sentences containing propositional attitudes has been mistakenly conflated with a third one: the opposition between two varieties of relational reading: (3a) the *transparent* and (3b) the *opaque*

readings. In opaque readings, replacement of a singular term by a co-referential term may fail to preserve truth.

The failure of the substitutivity principle applied to 'Cicero' in the opaque reading of 'Ralph believes that Cicero denounced Catilina' can be imputed to the *double role* played by 'Cicero.' The name 'Cicero' denotes the same individual for both the ascriber and the ascribee, but the ascribee, as opposed to the ascriber, is ready to use the *name* 'Cicero,' but not necessarily the name '*Tully*' for Cicero. The opaque reading of the belief sentence can thus be paraphrased to read:

Ralph believes of Cicero thought of as 'Cicero', that he denounced Catiline

The co-referentiality of 'Cicero' and 'Tully' licenses the replacement of 'Cicero' by 'Tully' when these names are *used*, but not when they are *mentioned*. Hence we *cannot* obtain *via the substitutivity principle*:

Ralph believes of Tully thought of as 'Tully', that he denounced Catiline

which is the formal paraphrase of the opaque reading of 'Ralph believes that Tully denounced Catiline.'

Existential generalization, however, goes through. As 'believes' is relational we can infer '$(\exists x)$ (Ralph believes that x denounced Catiline)' from 'Ralph believes that Cicero denounced Catiline' whether 'believes' is transparent or opaque.

The *hybrid reasoning* however causes a problem. Consider the inference: 'Cicero denounced Catiline and Ralph believes that Cicero denounced Catiline therefore there is someone who denounced Catiline and who is believed by Ralph to have denounced Catiline.' In the premise, the first occurrence of 'Cicero' refers to Cicero whereas the second refers to Cicero *thought of as 'Cicero'*. Hence we cannot, on pain of equivocation, represent its conclusion by an existential quantifier binding two occurrences of the same variable x.

Hintikka's epistemic logic is equipped to cope with that problem. Hintikka imputes the failure of existential generalization in epistemic contexts to a failure of the *presupposition of uniqueness* if the singular term occurs inside the scope of the belief construction. He imputes it to a failure of both an *existence* and a *uniqueness* presupposition if the singular term occurs inside *and* outside the scope of the belief construction.

On Hintikka's account, an inference of the form 'bRc & B_a bRc therefore $(\exists x)$ (xRc & $B_a xRc$)' in which b occurs both inside and outside the belief operator 'B_a' is valid only if we supply an *auxiliary premise* of the form '$(\exists x)$ ($x = b$ & $B_a x = b$).' Admittedly we have been forced to enlarge our *logic*, but we still do this without bringing *intensions* into our *ontology*.

In the semantics for modal (*viz.* epistemic and doxastic) logic, what one quantifies over is "the totality of those functions that pick out *the same* individual from the domains of the different possible worlds" (Hintikka 1969: 137). The world lines which tie up individuals across possible worlds, however, are human artefacts which do not belong to the *furniture of the world*. Our departure from Quine's ontology is thus reduced to the minimum.

As far as Quine is concerned, he endorsed the purely extensionalistic treatment of *de re* propositional attitudes worked out in Burdick's paper (1982: 185–230; see Quine 1995: 98).

10 Fiction, Intentional Objects and Existence

However different the *ontology* of fiction may be from that of nonfictional prose, its *logic* proves to be the same. Binary relations have a converse both in the real world and in the world of fiction: "[r]eaders will automatically conclude that Gladstone shakes hands with Holmes when reading that Holmes shakes hands with Gladstone" (La Palme Reyes 1994: 312).

Even though 'Sherlock Holmes' denotes nothing in the real world, it refers to something in fiction and even *refers rigidly*, that is it designates the same individual in all the counterfactual situations defined relatively to the situations taken as being actual within the work of fiction. Similarly 'man' is a natural kind.

The quantified phrase 'every man,' however, has a different *domain* in fiction and in standard discourse. Should Conan Doyle ascribe immortality to one of his characters, he would not falsify the sentence: '$(\forall x)$ (x is a man $\rightarrow$ x is mortal).' The domain of fiction does not *intersect* with the domain of science, even if a name like 'Gladstone' may occur both in fiction and in history books. In the novels, 'Gladstone' designates a character.

Can we form the *union* of the two domains? Lauener gives a negative answer: "I do not believe that lumping all the individuals into one huge pool would make sense" (Lauener 1986: 285). Can we lump together *possible worlds*? Hintikka replies that we cannot: "The . . . trouble . . . with Meinong's jungle, is that it has not been zoned, plotted and divided into manageable lots better known as possible worlds" (Hintikka 1989: 40).

Admittedly, if our concern is *ontological*, if we only care about the 'furniture of the world,' then putting actual entities and fictional beings together would blur the distinction between reality and fiction and generate pure obscurantism. There is, however, another approach, as Hintikka observes in *Intentions of Intentionality* (1975). Our concern may be *transcendental*. We may be interested in bringing together all *thinkable* objects (which include existents, inexistents, and even impossible beings).

If we want to quantify over that unified domain, however, we need *neutral quantifiers*. Here we move beyond *free logic*, which remained content with *neutral singular terms*, and we enter into *Meinongian logic* invented by Routley (1966) in '*Some things do not exist*' and developed by several authors. See the recent contributions due to Jacquette (1996) and Pasńiczek (1998).

Far from being a gratuitous exercise, a logic of that kind is indispensable if we want to represent, for example the inference which starts with the assumption that there is a barber who shaves everybody in the village who does not shave himself and which ends with the conclusion that there is not such a barber.

We need a *Meinongian logic* to assess reasoning about inexistents just as we need a *paraconsistent logic* (or Batens's *adaptive dynamic logic*) to assess the reasoning of the scientist confronting an inconsistency. When Clausius discovered a contradiction between Carnot's theory and Joule's ideas, he did not apply the principle *ex falso sequitor quodli-*

bet, nor did he stop reasoning. He "implicitly used a logic that *localizes* the specific contradictions and *adapts* itself to these" (Meheus 1993: 385).

11 Lesniewski's Ontology

Consider the following syllogism:

All horses are animals
Bucephalus is a horse
Bucephalus is an animal.

It contains two types of predication: (1) *generic/generic predication* in the major premise and in the conclusion, (2) *individual/generic predication* in the minor premise. Representing that syllogism within the predicate calculus forces us to alter the purity of logic by introducing a semantic distinction between *singular names* ('Bucephalus') and *general names* ('horse', 'animal'). Such a distinction blurs the fact that singular names are logically and syntactically on a par with general names (Waragai 1999: 15). Next we are led to fuse general names with the copula in front of them and to attribute different meanings to the *copula* 'is,' depending on whether it occurs in a predication of the first or of the second sort.

This has prompted several authors (among them Lejewski 1954) to switch from *first-order predicate logic* to the deductive system that Lesniewski created in 1920, that is to '*ontology*.' The latter is based upon a *single copula* in terms of which the other meanings of 'is' can be defined. No distinction is made in the system between proper and general names. The task of expressing existence can be removed from the quantifier and the identity can be made ontologically noncommittal, as it is the case in Meinongian logic.

Lesniewski's ontology has been recently shown to be interpretable in monadic second-order predicate logic, which shows that its first-order part is *decidable* (Cocchiarella forthcoming).

Acknowledgments

I wish to acknowledge a deep debt to Eric Audureau, Nino Cocchiarella, Gabriella Crocco, Lieven Decock, Susan Haack, and Claire Hill.

References

Ackrill, J. L. (1963) *Aristotle's Categories and De interpretatione*. Oxford: Clarendon Press.

Bayart, A. (1958) La correction de la logique modale du premier et du second ordre S5. *Logique et Analyse*, 1, 28–45.

Bayart, A. (1959) Quasi-adéquation de la logique modale du second ordre S5 et adéquation de la logique modale du premier ordre. *Logique et Analyse*, 6–7, 99–121.

Bergson, H. (1929) *Matter and Memory*. (N. M. Paul and W. Scott Palmer trans.) London: Allen & Unwin. (Original work published 1896.)
Boolos, G. (1975) On second-order logic. *Journal of Philosophy*, 72, 509–27.
Burdick, H. (1982) A logical form for propositional attitudes. *Synthese*, 52, 185–230.
Church, A. (1958) Ontological commitment. *Journal of Philosophy*, 55, 1008–14.
Cocchiarella, N. (1984) Philosophical perspectives in tense and modal logic. In D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic* (vol. 2): *Extensions of Classical Logic* (pp. 309–53). Dordecht: Reidel.
Cocchiarella, N. (1986) *Logical Investigations of Predication Theory and the Problem of Universals*. Napoli: Bibliopolis.
Cocchiarella, N. (1987) *Logical Studies in Early Analytic Philosophy*. Columbus: Ohio State University Press.
Cocchiarella, N. (1990) Quantification, time and necessity. In K. Lambert (ed.), *Philosophical Applications of Free Logic* (pp. 242–56). New York: Oxford University Press.
Cocchiarella, N. (1996) Conceptual realism in formal ontology. In R. Poli and P. Simons (eds.), *Formal Ontology*. (pp. 27–60). Dordrecht: Kluwer Academic.
Cocchiarella, N. (1997) Logic and ontology. In G. Küng, S. Norkowski, J. Wolenski and J. Kozak (eds.), *Logic, Philosophy and Ideology: Essays in Memory of Joseph M. Bochenski*.
Cocchiarella, N. (forthcoming) A conceptualist reduction of Lesniewski's ontology. *History and Philosophy of Logic*.
Crabbé, M. (1984) Typical ambiguity and the axiom of choice. *Journal of Symbolic Logic*, 49, 1074–8.
Davidson, D. (1968) On saying that. *Synthese*, 19, 130–46.
Decock, L. (forthcoming) *Trading Ontology for Ideology*.
Goclenius, R. (1613) *Lexicon philosophicum*.
Gupta, A. (1980) *The Logic of Common Nouns*. New Haven, CT: Yale University Press.
Haack, S. (1978) *Philosophy of Logics*. Cambridge: Cambridge Univesity Press.
Haack, S. (1992) "Extreme scholastic realism": its relevance to philosophy to-day. *Transactions of the Charles S. Peirce Society*, 28, 19–50.
Hill, C. (1997) *Rethinking Identity and Metaphysics*. New Haven, CT: Yale University Press.
Hintikka, J. (1959) Existential presuppositions and existential commitments. *Journal of Philosophy*, 56, 126–37.
Hintikka, J. (1969) *Models for Modalities*. Dordrecht: Reidel.
Hintikka, J. (1975) *The Intentions of Intentionality*. Dordrecht: Reidel.
Hintikka, J. (1989) *The Logic of Epistemology and the Epistemology of Logic*. Dordrecht: Kluwer Academic.
Hylton, P. (2000) Quine. *The Aristotelian Society*, Supplementary volume, 74, 281–99.
Jacquette, D. (1996) *Meinongian Logic*. Berlin: Walter de Gruyter.
Kripke, S. (1963) Semantical considerations on modal logic. In L. Linski (ed.), *Reference and Modality* (pp. 63–72). Oxford: Oxford University Press.
Lambert, K. (1987) On the philosophical foundations of free description theory. *History and Philosophy of Logic*, 8, 57–66.
La Palme Reyes, M. (1994) Reference structure of fictional texts. In J. Macnamara and G. E. Reyes (eds.), *The Logical Foundations of Cognition*. (pp. 309–24). New York: Oxford University Press.
Lauener, H. (1986) The language of fiction. *Bulletin de la Société mathématique de Belgique*, 38, 273–87.
Lejewski, C. (1954) Logic and existence. *British Journal for the Philosophy of Science*, 5, 104–19.
Meheus, J. (1993) Adaptive logic in scientific discovery: the case of Clausius. *Logique et Analyse*, 143–4, 359–91.
Meinong, A. (1960) The theory of objects (I. Levi, D. B. Terrell and R. M. Chisholm, trans.).

In R. M. Chisholm (ed.), *Realism and the Background of Phenomenology*. (pp. 76–117). Glencoe: The Free Press (original work published 1904).
Orilia, F. (1999) *Predication, Analysis and Reference*. Bologna: CLUEB.
Pasniczek, J. (1998) *The Logic of Intentional Objects*. Dordrecht: Kluwer Academic Publishers.
Putnam, H. (1975) *Mind, Meaning and Reality*. Cambridge: Cambirdge University Press.
Quine, W. V. O. (1953, 1961) *From a logical point of view*. New York: Harper & Row.
Quine, W. V. O. (1960) *Word and Object*. Cambridge, MA: MIT Press.
Quine, W. V. O. (1963) *Set Theory and its Logic*. Cambridge, MA: Belknap Press of Harvard University Press.
Quine, W. V. O. (1969) *Ontological relativity and other essays*. New York: Columbia University Press.
Quine, W. V. O. (1970) *Philosophy of Logic*. Englewood Cliffs, NJ: Prentice-Hall.
Quine, W. V. O. (1974) *The Roots of Reference*. La Salle, IL: Open Court.
Quine, W. V. O. (1981) *Theories and Things*. Cambridge, MA: Belknap Press of Harvard University Press.
Quine, W. V. O. (1982) *Methods of Logic*. 4th edn. Cambridge, MA: Harvard University Press.
Quine, W. V. O. (1994a) Assuming objects. *The Journal of Philosophy*, 60, 171–92.
Quine, W. V. O. (1994b) Promoting extensionality. *Synthese*, 98, 143–51.
Quine, W. V. O. (1995) *From stimulus to science*. Cambridge, MA: Havard University Press.
Recanati, F. (2000) Opacity and the attitudes. In P. Kotatko and A. Orenstein (eds.), *Knowledge, Language and Logic: Questions for Quine*. (pp. 367–407). Dordrecht: Kluwer Academic.
Resnik, M. (1997) *Mathematics as a Science of Patterns*. Oxford: Clarendon Press.
Ross, W. D. (1924) *Aristotle's Metaphysics*, 2 vols. Oxford: Clarendon Press.
Ross, W. D. (1949) *Aristotle's Prior and Posterior Analytics: A Revised Text*. Oxford: Clarendon Press.
Routley, R. (alias Sylvan) (1966) Some things do no exist. *Notre Dame Journal of Formal Logic*, 7, 251–76.
Russell, B. (1903) *Principles of Mathematics*. London: W. W. Norton.
Russell, B. (1908) Mathematical logic as based on the theory of types. In B. Russell, *Logic and Knowledge*. (pp. 59–102). London: George Allen & Unwin.
Russell, B. (1918) The philosophy of logical atomism. In B. Russell, *Logic and Knowledge*. (pp. 177–281). London: George Allen & Unwin.
Simons, P. (1997) Higher-order quantification and ontological commitment. *Dialectica*, 51, 255–71.
Specker, E. (1953) The axiom of choice in Quine's New Foundations for mathematical logic. *Proceedings of the National Academy of Science*, 39, 972–5.
Vidal Rosset, J. (forthcoming) "New Foundation": un exemple de la relativité des normes en théorie des ensembles.
Waragai, T. (1999) Aristotle's master argument about primary substance and Lesniewski's logical ontology: a formal aspect of metaphysics. In R. Rashed and J. Biard (eds.), *Les doctrines de la science de l'Antiquité à l'Age classique*. (pp. 9–35). Leuven: Peeters.
Wittgenstein, L. (1961) *Tractatus logico-philosophicus*. (D. F. Pears and B. F. McGuinness trans.). London: Routledge & Kegan Paul (original work published 1921).

Further Reading

Batens, D. (1994) Inconsistency-adaptive logics and the foundations of non-monotonic logics. *Logique et Analyse*, 145, 57–94.
Bencivenga, E. (1986) Free logics. In D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic* (vol. 3): *Alternatives in Classical Logic*. (pp. 373–426). Dordecht: Reidel.

Castaneda, H.-N. (1989) *Thinking, Language and Experience.* Minneapolis: University of Minnesota Press.
Chakrabarti, A. (1997) *Denying Existence.* Dordrecht: Kluwer Academic Publishers.
Gochet, P. (1980) *Outline of a Nominalist Theory of Propositions.* Dordrecht: Reidel.
Gochet, P. (1986) *Ascent to Truth.* Munich: Philosophia Verlag.
Gochet, P., Gribomont, P. and Thayse, A. (2000) *Logique*, vol. 3. Paris: Editions Hermès.
Haack, S. (1996) *Deviant Logic, Fuzzy Logic: beyond the Formalism.* Chicago: Chicago University Press.
Haller, R. (ed.) (1985–6) Non-existence and predication. *Grazer Philosophische Studien*, 25–6.
Haller, R. (ed.) (1995) Meinong and the theory of objects. *Grazer Philosophische Studien*, 50.
Hausser, R. (1999) *Foundations of Computational Linguistics.* Berlin: Springer Verlag.
Lambert, K. (1981) On the philosophical foundations of free logic. *Inquiry*, 24, 147–203.
Ludwig, K. and Ray, G. (1998) Semantics for opaque contexts. *Philosophical Perspectives*, 12, 141–66.
Lycan, W. (1999) The trouble with possible worlds. In M. Tooley (ed.), *Analytical Metaphysics* (vol. 5). (pp. 274–316). New York: Garland Publications.
Marcus, R. B. (1993) *Modalities.* New York: Oxford University Press.
Morscher, Ed., Czermak, J. and Weingartner, P. (eds.) (1977) *Problems in Logic and Ontology.* Graz: Akademische Druk und Verlangsanstalt.
Nef, F. (1999) *L'objet quelconque.* Paris: Vrin.
Orenstein, A. (1977) *Willard Van Orman Quine.* Boston, MA: Twayne.
Orilia, F. (1987) Definite descriptions and existence attributions. *Topoi*, 6, 133–8.
Parsons, T. (1980) *Nonexistent Objects.* New Haven, CT: Yale University Press.
Pasńiczek, J. (1993) The simplest Meinongian logic. *Logique et Analyse*, 143–4, 329–42.
Rahman S., Rückert, H. and Fischmann, M. (forthcoming) On dialogues and ontology: the dialogical approach to free logic. *Logique et Analyse.*
Rouilhan, Ph.de. (1996) *Russell et le cercle des paradoxes.* Paris: Presses Universitaires de France.
Routley, R. (1980) *Exploring Meinong's Jungle and Beyond.* Canberra: Philosophy Department.
Schock, R. (1968) *Logics without Existence Assumptions.* Stockholm: Almqvist & Wiksell.
Shaw, J. L. (1988) Singular existential sentences: contemporary philosophy and the Nyaya. In P. Bilmasia and P. Fennor (eds.), *Religions and Comparative Thought.* (pp. 211–40). Delhi: Indian Books Centre.
Simons, P. (1995) *Philosophy and Logic in Central Europe from Bolzano to Tarski.* Dordrecht: Kluwer Academic.
Waragai, T. (1990) Ontology as a natural extension of predicate calculus with identity equipped with description. *Annals of the Japan Association for Philosophy of Science*, 7, 233–50.
Waragai, T. (1996) IS-A relation, the principle of comprehension and the doctrine of limitation of size. *Annals of the Japan Association for Philosophy of Science*, 9, 23–34.

18

From Logic to Ontology: Some Problems of Predication, Negation, and Possibility

HERBERT HOCHBERG

1 Negation and Nonexistence

Russell contrasted considering things "from a logical point of view" with linguistic and philosophical points of view, where 'philosophical' meant 'ontological.' This fits with his suggesting 'Philosophical Logic' for the title of Wittgenstein's *Tractatus*. While Wittgenstein supposedly responded that the phrase was nonsense, ontological issues raised by logic and 'logical form' are now a basic part of philosophical logic. One of the oldest problems concerns negation. In a passage in *The Sophist* Plato writes: "When we assert *not-being* it should seem, what we assert is not the *contrary* of *being*, but only something *other*." (Taylor 1971: 164). Taken with other passages, this suggests Plato considers 'x is not-F' as (1) '(f)(f ≠ F **if** fx)' or (2) '(y)(y ≠ x **if** Fy).' [I use '**if**' to avoid explicit use of a truth-functional conditional and obvious problems posed by using '⊃' to eliminate '¬'.] While Taylor takes Plato to construe 'not-F' in terms of "being something other than what we call 'F'," he construes 'other' in a restricted sense that involves F belonging to a group of incompatible properties. Owen rejects such an interpretation of Plato and takes 'x is not-F' as 'all attributes of x are different from F', rather than as 'some attribute of x *excludes* F' (Owen 1986: 131, 114–15). Many Plato scholars consider such readings problematic, but Owen, following (1), analyzes not-F in terms of the Platonic forms of *difference* (≠) and *sameness* (=). Forgetting 'types' and questions about how Plato construes relations, obvious problems arise if '=' and '≠' are used to define '¬', as the equivalences – (a) '¬(f = g) **iff** f ≠ g' and (b) '¬(f ≠ g) **iff** f = g' – are not derivable. Consider (b). Using (1), '¬(f ≠ g)' becomes '(R)(≠(R, ≠) **if** R(f, g))'. But we can neither derive 'f = g' from that nor that from 'f ≠ g', though '[(R)(≠(R, ≠) **if** R(f, g))] **iff** f = g' may seem obviously true. The same holds for (a) and '[(R)(≠(R, =) **if** R(f, g))] **iff** f ≠ g', using (2) in place of (1). '¬¬Fa **iff** Fa' poses a related problem.

Bradley and Bosanquet suggested '¬Fx' be construed as '(∃f)(fx & f is incompatible with F),' while Demos took '¬p' as '(∃q)(q is true & q is incompatible with (in opposition to) p).' Russell argued that Demos did not avoid *negative facts*, as incompatibility is a form of negation, being the Scheffer stroke function, and that Demos' view generated a problematic regress, since 'p and q are in opposition' means 'p and q are not both true' (Russell 1918, 1919). But in the 1925 edition of *Principia* he wrote: "Given all true atomic propositions, together with the fact that they are all, every other true

proposition can theoretically be deduced by logical methods" (Whitehead and Russell 1950: xv). Since atomic facts ontologically grounded true atomic propositions, this took a set of atomic facts and the general fact that the set contained *all* the *atomic facts* to avoid negative facts, though he spoke of *atomic propositions*, not atomic facts. A negation of an atomic proposition was true if it followed from the statement of the general fact and the 'list' of true atomic propositions. This followed his rejecting conjunctive facts due to 'p, q ⊨ p & q.'

Russell's theme has been revived in recent years by views appealing to a 'meta-fact' about atomic facts or a *class* or *totality* of atomic facts or both. The simplest version of such a view recognizes a domain of *all* atomic facts, a class, as the ontological ground for true negations of atomic propositions. The class of all atomic facts is taken to suffice as the truth maker for such negative truths, since it is purportedly *not* a fact *that* an excluded atomic fact *is not* in the class. That is a consequence of an ontological analysis of classes taking a class, say {a, b}, to suffice as the truth ground for statements like 'a ∈ {a, b}' and '¬c ∈ {a, b},' as opposed to holding that a relation, ∈, obtains or does not between the term and the class. Moreover, one can argue that classes are presupposed by standard systems of logic, since the logical variables and quantifiers presuppose domains, which are classes, or 'ranges' of application. But recognizing such a 'range' implicitly recognizes a domain (class) comprised of all and only things satisfying some condition. As classes of particulars and properties correspond to individual and predicate variables and quantifiers, respectively, the sentential variables can be taken to correspond to a domain of facts, rather than to the 'truth values' used in the 'evaluations' in logic texts. This requires rejecting the 'substitutional' account of quantification as untenable. On such an account, the quantifier sign '(∃x)', for example, is read in terms of 'There is a name (constant)' rather than 'There is an object (individual).' Supposedly one can then 'semantically ascend' to talk of signs, instead of things, and avoid *ontological commitments* to non-linguistic objects in the domain of a quantifier. Semantic 'ascension' has led to tortured attempts to prove that formal systems can have non-denumerably many proper names. But that is of no import, for one can simply assume that there are sufficiently many individual constants. The real problem is the assumption that for *every* object there is a corresponding sign – a claim that involves quantifying over *objects* as well as signs. The standard response, going back to a 1968 argument of Belnap and Dunn, is the pointless and problematic claim that such a use of an objectual quantifier can be construed substitutionally, involving a further quantification over objects that is then treated substitutionally, and so on *ad infinitum*.

Conjunctive and disjunctive facts may be avoided, as true conjunctions and disjunctions are so in virtue of the truth or falsity of component atomic sentences. But negation raises a unique problem. The difference is reflected, first, by there being no standard logical rule for negation corresponding to 'p, q ⊨ p & q,' and, second, by an evaluation assigning *one of* T or F, but *not both*, to the atomic sentences in a standard bivalent logic. This latter point can be taken to reflect the traditional logical laws of excluded middle and non-contradiction and their special status, though all tautologies, being logically equivalent, are 'equal.' Such laws provide a basis for the use of truth tables, as the 'law of identity' is presumed by any coherent system of signs. But the truth table for negation does not explicate the meaning of '¬'. Nor does it resolve

questions about negative facts and the ontological correlate of the negation sign. What a standard truth table shows is: (1) that '¬' is taken as the sign for negation; (2) that every sentence of the schema is taken to be true *or* false; and (3) that 'or' in (2) is used in the exclusive sense since no sentence is both true and false. Some logical signs, and concepts, are basic, and so-called *elimination* and *introduction* rules neither provide analyses of them nor resolve ontological issues raised by them. Recent purported explications of the *meaning* of the quantifier signs, stemming from Wittgenstein's Tractarian views, by means of such rules also do not do what they purport to (Celluci 1995; Martin-Löf 1996). Such rules merely codify the *interpretation* of the quantifier *signs*, as truth tables do for truth functional *signs*, linking them to *generality* and *existence.*

Russell appealed to a general fact about all true atomic facts *and*, implicitly, a class ('list') of atomic facts to ground the truth of true negations. Later, others took such a class of atomic facts to suffice while some held that a general fact alone sufficed. All such attempts fail to resolve the issue of negative facts. We can see why by returning to the attempt to take the truth ground of '¬ Fa' to be a class, D, of atomic facts *and* a general fact, (p)(p ≠ Fa), with D giving the range of 'p.' '(p)(p ≠ Fa)' states that no fact is a's being F. But such a general fact involves an apparent negation. Limiting the discussion to a miniature world (model) with D = {Ga, Fb}, we can take Russell's *list* in terms of '(p)(p = Ga ∨ p = Fb),' stating that Ga and Fb are all the atomic facts, without a negation. But that is still problematic. As '(∃x)(∃y)(x ≠ y & (z)(z = x ∨ z = y))' states that there are only two particulars, we can state that a and b are the only particulars by '(x)(x = a ∨ x = b).' That entails, with an additional name 'c,' that '(∃x)(x = c)' entails '(c = a ∨ c = b).' But as '(x)(x = a ∨ x = b)' does not entail '¬(∃x)(x = c),' '(p)(p = Ga ∨ p = Fb)' does not entail either '¬(∃p)(p = Fa)' or' '¬ Fa' is true.' All that follows that is relevant is '(∃p)(p = Fa) ⊨ (Fa = Ga ∨ Fa = Fb),' which entails '(Fa ≠ Ga & Fa ≠ Fb) ⊨ ¬(∃p)(p = Fa),' *assuming* that we can instantiate to 'Fa,' as we assumed about 'c' above. Stating that the nonexistence of Fa grounds the truth of '¬Fa' thus involves 'Fa ≠ Ga & Fa ≠ Fb' or '(p)(p ≠ Fa),' as well as the apparent implicit use of 'Fa' to represent a nonexistent fact.

The issue raised by '(p)(p ≠ Fa)', or instantiating to 'Fa' from '(p)(p = Ga ∨ p = Fb),' recalls Meinong's nonexistent objects and nonsubsistent *objectives*, since Fa does not exist. The correspondence theory of truth, taking facts as truth grounds for sentences (propositions), that Moore set forth in lectures of 1910–11 raised the issue that was put cryptically by Wittgenstein (1961: 4.022). An atomic statement, or a 'thought' that a is F, represents a situation – *shows* its sense – whether or not it is true, and *states* that it obtains. As showing or *representing* is a relation, between a statement or thought and a *situation*, that obtains whether or not the represented situation does, since the thought must have the same sense whether it is true or not, a problem arises. Moore avoided the issue by saying that his talk of the 'non-being' of a fact was merely an unavoidable way of speaking, while taking the *being* of the fact that-p to *directly prove* the truth of 'the belief that-p.' But holding, like Russell, that "Fa" is true **iff** the fact that-a is F exists', his use of the clause 'that-a is F' pointed to the implicit recognition of facts as *possibilities* (situations) which may *obtain* (exist) or *not*. Thus *correspondence* was an ambiguous concept. In one sense 'Fa', whether true or not, corresponded to a *possibility*; in another sense, if true, it corresponded to an *existent* fact.

2 Designation and Existence

Carnap (1942: 24, 50–2) considered the issues of truth and reference in terms of the semantics of 'designation'. Consider (1) 'a' designates Theaetetus; (2) 'F' designates the property of flying; (3) 'Fa' designates the state of affairs that Theaetetus is flying. Carnap took (1)–(3) as semantical 'rules' for a schema. With *designates* as a semantical relation, (3) is true even if 'Fa' is false. (1)–(3), as semantical rules, do not express matters of fact. That such rules are rules of a particular schema is a matter of fact. The same sort of distinction applies to ordinary language variants of (1)–(3) – 'Theaetetus' designates Theaetetus, etc. Considered as statements about the usage of terms, they express matters of fact, but, properly understood, they are semantic rules. Taking the signs as interpreted signs – symbols, in the sense of Wittgenstein's *Tractarian* distinction between a sign and a symbol, there is, in a clear sense, an *internal* or *logical* relation involved in such rules. (1)–(3) express *formal* or *logical truths*, since the *symbols*, not signs, would not be the symbols they are without representing what they represent. This incorporates a 'direct reference' account of proper names and the direct representation of properties and relations by primitive predicates. This was involved in Russell's notion of a "logically proper name" or label that functioned like a demonstrative, as opposed to a definite description that 'denoted' indirectly, via the predicates in the descriptive phrase. In the last decades of the century, with the decline of interest in and knowledge of the work of major early twentieth-century figures, petty debates have erupted about priority. One of the most absurd concerns whether Barcan or Kripke originated Russell's account, which was set out in the first decade of the century and adopted by many since. The absurdity has been compounded by the misleading linking of Russell with Frege in what some speak of as the 'Frege–Russell' account of proper names, which ignores Russell's attack on Frege's account in the classic "On Denoting" (Russell 1956a; Hochberg 1984). The direct reference account was ontologically significant for Russell and others who took the primitive nonlogical constants (logically proper names and predicates), representing particulars and properties (relations) respectively, to provide the ontological commitments of the schema (Bergmann 1947; Hochberg 1957). This contrasted with Quine's taking quantification as the key to ontological commitment – "to be is to be the value of a variable" – which allows a schema limited to first order logic to contain primitive predicates while avoiding properties, by fiat. That fits Quine's replacing proper names by definite descriptions, involving either primitive or defined predicates. For one only then makes ontological *claims* by means of variables and quantifiers, and predicates retain ontological innocence (Quine, 1939, 1953). If primitive predicates involve ontological commitments, as in Carnap's (2), attempting to eliminate all directly referring signs via descriptions faces an obvious vicious regress, aside from employing an *ad hoc* and arbitrary criterion.

Wittgenstein simply ignored the problem about (3) by giving (1) and (2) the role of (3), as Russell was to do in the 1920s under his influence. This was covered over by his speaking of the 'possibilities' of combination being 'internal' or 'essential' properties of the 'objects' that were combined. Carnap's (3), which articulates Moore's view, makes explicit reference to a possible fact or situation. Russell had suggested using his theory

of descriptions to avoid reference to possible facts, as well as to nonexistent objects (Russell 1905). He developed that idea in 1913 (Russell 1984; Hochberg 2000), but abandoned the book, partly due to Wittgenstein's influence. Russell replaced (3) by "'Fa' is true **iff** the fact *consisting of* Theaetetus and the property of flying exists," thereby avoiding a designation relation connecting a sentence to a purported state of affairs. What he suggested is more explicitly rendered by:

(3R) 'Fa' is true ≡ Fa ≡ E! (p)(T(a, p) & A(F, p) & f(φx, p)),

with 'T,' 'A,' and 'f' for 'is a term in,' 'is attributed in,' and 'is the form of' and Ωx as the form of monadic first-order exemplification. (3R) is a tripartite biconditional that is an interpretation, but *not* designation, rule *and* a 'rule of truth,' specifying a truth maker, that avoids *possibilities* and Meinongian nonsubsistent objectives. The relations T, A, and f do not raise the same problem, since atomic sentences, unlike names and predicates, are not *designators*, as they are in Carnap's (3). Since they do not designate atomic sentences are not taken as names of *situations* in (3R), as Wittgenstein does take them in the *Tractatus*, despite his claim to the contrary. We can now express the nonexistence of the *purported* fact that-a is F by:

(3N) '¬Fa' is true ≡ ¬Fa ≡ ¬E! (ι p)(T(a, p) & A(F, p) & f(φx, p)).

The question that arises is whether recognizing the class of atomic facts allows for specifying a truth maker in terms of (3N) without recognizing negative facts. One might argue we can do so since an ontological ground for taking such statements to be true is acknowledged: the class or domain of atomic facts taken as the correlate of the sentential variables. It is tempting to argue that it is no more a further fact that no such member of the class exists than it is a further fact that such a fact does exist, if the sentence 'Fa' is true. As there is no need to hold that when an atomic fact exists there is an additional fact, the fact that the atomic fact exists, there is no need to recognize the fact that an atomic fact does not exist, a negative fact, when the atomic fact does not exist. This is supposedly reinforced by recognizing that what makes a statement of class membership true or false is not a relational fact involving the relation of class membership, but simply the class itself. One can apply the same idea in the case of true negations, by taking 'Fa' to be false *given* the class of atomic facts. If the appeal to a set of facts, taken as the domain of atomic facts, is viable we can avoid negative facts. But there is a simple argument against such a view. We cannot say, where '¬Fa' is true, that *the fact* Fa does not belong to the totality or *is not* or that the fact Fa is not identical with any of the atomic facts by using 'Fa' or the expressions 'that-Fa' or 'the fact Fa' to *designate* a nonexistent fact. Rather, we can only *describe* such a fact and purport to *denote* it by a definite description to make such a claim.

The claim that the fact Fa does not belong to the class of atomic facts thus involves a description of that fact and a statement of the form '¬((ι p)(T(a, p) & A(F, p) & f(φx, p)) ∈ D),' and not one like '¬c ∈ {a, b}.' We cannot simply appeal to a class or domain as the truth ground for either 'The fact Fa does not exist' or '¬Fa.' For such attempts to dispense with negative facts involve implicit claims that amount to:

(N′) $(q)(q \neq (\iota\, p)(T(a, p) \,\&\, A(F, p) \,\&\, f(\Omega x, p)))$,

where the variables 'q' and 'p,' as earlier, range over existent atomic facts. (N′) serves the purpose of a list or corresponding universal disjunction or reference to the domain D, while avoiding problems raised by infinite lists or disjunctions. If '≠' is a primitive sign, as *diversity* is taken by some to be *phenomenologically* basic, as opposed to *identity*, then (N′) becomes:

(N″) $(\exists_u p)((T(a, p) \,\&\, A(F, p) \,\&\, f(\Omega x, p)) \,\&\, (q)(q \neq p))$,

using the subscripted 'u' for 'uniqueness.' This will obviously not do as an expression of the truth ground for '¬Fa,' since it states that *there exists* a fact, a's being F, that is diverse from every fact. Making sense of (N″) requires accepting both existent (actual) and merely possible facts, different senses of 'exists' and different variables to range over such respective domains, as in some 'free' intensional logics. This is not acceptable to one seeking to avoid negative facts by appealing to classes or totalities. Yet, to reject diversity as basic, and treat '≠' in terms of '¬' and '=', treats (N′) as '$(q)\neg(q = (\iota\, p)(T(a, p) \,\&\, A(F, p) \,\&\, f(\Omega x, p)))$', and hence as:

$(q)\neg(\exists_u p)((T(a, p) \,\&\, A(F, p) \,\&\, f(\varphi x, p)) \,\&\, (q = p))$.

This returns us to the problematic use of an embedded negated existential claim that either repeats what we must account for or leaves us with the issue of negative facts.

The rejection of negative facts by simple appeals to classes or totalities is not viable. Accepting them, however, poses a problem as to their analysis (Hochberg 1999: 193f). But there is an alternative. Consider the following derivation, with 'p' and 'q' ranging over monadic atomic facts (thus simplifying matters by omitting reference to the form φx):

(DN) 1 $a \neq b$
2 $F \neq G$
3 $(q)(q = (\iota\, p)(T(b, p) \,\&\, A(F, p)) \vee q = (\iota\, p)(T(a, p) \,\&\, A(G, p)))$
4 $(p)(q)[\{(x)(y)(f)(g)((x = y) \,\&\, (f = g) \,\&\, T(x, p) \,\&\, A(f, p) \,\&\, T(y, q) \,\&\, A(g, q)) \equiv p = q\}]$
5 $(\exists x, y, f, g)(x = a \,\&\, y = b \,\&\, F = f \,\&\, G = g)$

6 $\neg E!\ (\iota\, p)(T(a, p) \,\&\, A(F, p))$.

Since (4) states that monadic (first order) atomic facts are the same **iff** their constituents are the same, (DN) is a valid argument. Hence, as '$\neg E!\ (\iota\, p)(T(a, p) \,\&\, A(F, p))$' is taken to be equivalent to, or a transcription of '¬Fa,' we have derived the latter. For, assuming '$E!\ (\iota\, p)(T(a, p) \,\&\, A(F, p))$,' we can instantiate (3) to (6) '$Fa = Fb \vee Fa = Ga$,' using the atomic sentences to abbreviate the corresponding descriptions of the relevant purported facts. But, by (4), (6) is false, so we arrive at '¬Fa', that is '$\neg E!\ (\iota\, p)(T(a, p) \,\&\, A(F, p))$.' We thus 'ground' the truth of '¬Fa' *without* appealing to a negative fact by the use of '$\neg E!(\iota\, p)(T(a, p) \,\&\, A(F, p))$' as an implicit premise. (DN) differs in this crucial way

from using a generalization like '(q)¬(q = (ιp)(T(a, p) & A(F, p))),' as a premise, to arrive at '¬E! (ιp)(T(a, p) & A(F, p)).' In the latter case, since the premise and conclusion are trivially equivalent, we merely assume the negation to be derived and thereby acknowledge, rather than avoid, negative facts. But (DN) requires (1) and (2), which can be taken as recognizing basic and specific facts of diversity. This raises two issues: Is diversity or identity the fundamental concept? Are facts of diversity, or denials of identity, negative facts? In any case, (DN) can be seen as illustrating a sense in which Plato was right.

3 Logical Truth, Modality, and Ontology

We avoid conjunctive facts since 'p, q ⊨ p & q' justifies taking the facts that ground the truth of the conjuncts as the truth makers for a true conjunction. But what ontologically grounds logical entailments and logical truths? To hold there is no ground can lead one to follow logical positivists and rule out the question as a pseudo-question, along with other 'metaphysical' questions, and to viewing logic as a matter of 'convention' or as involving only 'internal' questions relative to a system. The conventionalist move has variants other than the Viennese one. There is the French fashion that includes 'being responsible for your own birth,' as *we* impose our concept of 'birth' on Being (Sartre), and "your child not being your child without language" (Lacan), and the anglo-American variants emphasizing 'world making,' 'ways of life,' 'webs of belief,' 'rules,' 'normative aspects' and 'social contexts.' All of them, linked one way or another to holism and German Idealism, have a hollow ring, as does Hume's speaking of the 'necessity' of logical entailment in terms of a psychological *determination* to proceed from one idea to another. Employing model theory (set theory) to provide 'semantics' for logical systems does not change the basic issue, despite familiar problems that lead some to believe that logic rests on axiomatic systems that require 'arbitrary' (hence conventional) restrictions to avoid paradoxes. Neither positivism nor conventionalism fits the obvious fact that coherent discussion of the issues assumes fundamental and familiar logical truths and rules. In a different context, Moore expressed the basic theme behind the logical realist's rejection of the three-headed Hydra of conventionalism–idealism–psychologism: the task is not to prove the obvious but to clarify the grounds for it being so.

Ontologically grounding logical truth is traditionally linked with the explication of 'necessity' and 'possibility' and the question of whether there are necessities other than logical ones. Concern with modalities dates from Aristotle through the medieval period to the present. The logical positivists, following a theme in Russell and the early Wittgenstein, sought to explicate 'necessary truth' in terms of *logical truth*. The latter notion was sometimes considered in purely formal or 'syntactical' terms. Logical and mathematical truths were taken to be so since they were theorems of certain calculi. This led to Carnap's distinguishing 'external' from 'internal' questions and declaring the former 'pseudo-questions.' One could only consider questions about logical and mathematical truth as questions about formulae being theorems of some system. Aside from the inadequacy of such a view, given Gödel's incompleteness result (Gödel 1986; Lindström 2000), it is philosophically inadequate in a basic sense. Consider standard

propositional logic, which is complete. It is a system of logical truths in virtue of the concepts of truth, falsity, and negation and the logical 'laws' that truth tables are based on. Speaking of logical or mathematical truth solely in terms of theorems of some formal system takes one nowhere.

Carnap subsequently sought to explicate the notions of logical truth, necessity, and possibility, by extending his 1942 system of semantics to modal logic. The development of modern modal logic is taken to begin with Lewis' and Langford's work on propositional modal logic – their addition of modal signs ('◊' for 'possible') to propositional calculi and employing a modal conditional of *strict implication* 'p ɜ q.' But Carnap took the modal concepts to be 'unclear' and 'vague' requiring an explication of the notion of 'logical necessity.' He sought to provide one in terms of *logical truth*, taken as a meta-linguistic 'semantical concept' (Carnap 1947: 174). It is fashionable, but more myth than fact, to date quantified modal logic from Barcan's March, 1946 paper (Hughes and Cresswell 1996: 255; Boolos 1993: 225) that was received on September 28, 1945, having been extracted from a doctoral dissertation in progress. The same journal published Carnap's paper (1946) on quantification and modalities in June, having received it November 26, 1945. Carnap's paper was based on *Meaning and Necessity*, a book he had worked on in 1942, completed a first version of it in 1943 and, after an extensive correspondence with Quine and Church, published in 1947 (Carnap 1947: vi). It was the third of a series of books on logic and semantics done in the 1940s. In both earlier works of the trilogy he mentioned his work on a system of quantified modal logic in the 1943 manuscript (Carnap 1942: 85, 92; Carnap 1943: xiv).

Carnap's 1946 paper contains one of the earliest semantics for a system of modal logic that he altered and developed in the 1947 book. Barcan's paper consists of simple derivations from assumptions. One *assumption*, the *Barcan formula*, was among the *theorems* for which Carnap offered semantical proofs in 1946 and 1947. Semantics for modal systems, of the general kind now associated with Kripke's name, occur earlier in the work of Kanger (1957) and Bayart, while Carnap is often said to have 'anticipated' them. Based on a theme in the *Tractatus*, reflecting the idea that a necessary truth is true in all possible worlds, Carnap introduced 'state descriptions,' as sets or lists of *sentences*, which we can consider, in terms of the miniature model we used for discussing negation, as sets of *possibilities*. In our simple case there are the sets: {Fa, Gb, ¬Fb, ¬Ga}, {Fa, ¬Gb, Fb, ¬Ga}, etc. While 'Fa' would be true for only some such sets or 'worlds,' 'Fa ∨ ¬Fa' would be true in all, and hence *necessarily* true. Thus '**N**(p ∨ ¬p)' is a theorem by a *rule* for '**N**' ('necessary'). Carnap's system is, in effect, the one known as S_5, obtained by the addition of '◊p ⊃ **N**◊p' to S_4, 'characterized' by '**N**p ⊃ **N**(**N**p)'. One adds these to a system, often called 'T', usually obtained by taking '**N**(p ⊃ q) ⊃ (**N**p ⊃ **N**q)', '**N**p ⊃ p' and all valid formulae of standard propositional logic as axioms, along with rules like substitution, *modus ponens*, and 'necessitation' (if a formula is a theorem then the formula preceded by '**N**' also is).

In S5 the modal concepts are not relativized to a possible world. The essential conceptual change made after Carnap was that the modal concepts were relativized, so certain 'things' (including 'worlds') were possible *relative to* some possible worlds but not others. For example, a 'world,' w, with domain {a, b} can be said not to be 'acces-

sible from' one, w*, with domain {b}, and Fa is not then 'possible' relative to w* (whether 'Fa' is then rejected as a formula, taken to be without a truth value, etc. is irrelevant). This illustrates the simple idea behind the later modifications of Carnap's semantics that led to formally characterizing different modal systems in terms of logical characteristics (transitivity, symmetry, etc.) of a relation, on the set of worlds, and constructing models for alternatives to S5, such as S4, T, etc. But, from an ontological point of view, we merely have various axiom systems about unexplicated and ungrounded modal concepts or overly rich ontologies (if one speaks literally of 'worlds'), though different systems appear to fit, more or less, different uses of 'necessity' and 'possibility' – logical necessity, causal necessity, etc. One problematic mutation has been the construal of causal notions in terms of a primitive counter-factual relation, $p\Box\!\!\rightarrow q$ (had p occurred q would have), and a triadic *similarity* relation, S, between possible worlds – w is more similar to (closer to) w′ than w″ is – where conditions for S provide a 'semantics' for '$\Box\!\!\rightarrow$'. Such attempts are notoriously vague, either turning in transparent circles by illicitly employing causal notions or introducing arbitrary stipulations (conditions) relativizing S (closer in what way?). The appeal to possible worlds as entities is often denied by claiming that talk of such worlds is merely a way of speaking, as Moore once said about his referring to nonexistent facts. But philosophical honesty requires literal talk or the admission that one merely speaks fancifully about linguistic structures, models, and connections among them. Recent revivals of Carnap's construal of state descriptions as sets of sentences take the form of construing possible worlds as sets of sentences. This leads some to think that, as sets are 'abstract entitites,' they deal with 'metaphysics' and ontology. Such pointless patterns invariably involve problematic uses of the term 'sentence' and ignore the fact that atomic sentences require interpretation rules like Carnap's (3). Such rules introduce the basic problems posed by possible facts and possible worlds that are obviously not resolved by talking about sets of sentences. *Modes* of facts (*possibility, actuality*) and possible facts can ontologically ground talk of possible worlds, taken as sets of facts where at least one element is a *possible* but not *actual* fact. But such modes, as modalities, are neither clarified nor codified by the various modal logics. The same is true of (1) the use of 'possible' involved in considering the possibility of further objects, properties, and worlds; (2) the sense of 'possible' in the phrase 'possible world'; and (3) the 'modality' involved in categorial *necessities* – F necessarily being a property, etc. – based on *exemplification* and presupposed by standard systems of logic as well as modal systems. The philosophical problems posed by modal logics might have led to their demise but for connections to intuitionistic logic (Gödel 1933) and 'reinterpretations' of '**N**' and '◊' in terms of *provability* and *consistency*, in the development of a logic of *provability* related to Gödel's incompleteness results (Boolos 1993; Lindström 1996).

Predication and the categorially *necessary* distinction between terms and attributes are at the core of two logical paradoxes – the Bradley–Frege paradox and the Russell paradox. Both stem from mistaken ontological analyses of *exemplification* and confusing properties with propositional functions. The first results from taking exemplification as a relation connecting a term (terms) and a property (relation) to form a fact (or proposition). It then seems that a further relation must connect the exemplification relation itself to the term(s) and property (relation), and so on *ad infinitum*. This led

Frege, and later Russell in the case of relations (sometimes properties as well), to insist that no such relation was needed as 'concepts' were 'incomplete' or 'unsaturated' functions that required being completed rather than being connected to terms (arguments). It led Bradley to hold that exemplification was paradoxical and Sellars to argue that realism about properties and relations was thereby refuted. The problem, which bears on an earlier discussion, disappears on the analysis employing (3R). Consider the fact (ɪ p)(T(a, p) & A(F, p) & f(φx, p)). No *connection* of exemplification is involved. Monadic exemplification is a *logical form*, Øx. We also have recognized *logical relations*, T, A and f, between a fact and its term, attribute, and logical form, *but they are not exemplified* and do not give rise to *further facts* or *possibilities*. A relation of exemplification is not illicitly used in clauses like 'T(a, p)', as there is no further fact that a and the described fact *exemplify* T. For, by Russell's theory of descriptions, 'T(a, (ɪ p)(T(a, p) & A(F, p) & f(φx, p)))' – a stands in T to the fact that a is F – reduces to 'E! (ɪ p)(T(a, p) & A(F, p) & f(φx, p))' – the fact that a is F exists. This is why T, A, and f can be said to be *logical relations* without simply dismissing the problem. They are not relations that combine terms into further facts. Hence no regresses arise. The point can be reinforced. Assume the fact exists and we 'name' it '[Fa]'. 'T(a, [Fa])' is true since [Fa] exists, and not in virtue of a further relational fact, as 'a ∈ {a, b}' is true given {a, b}. The point also applies to *part-whole* relations and mereological calculi.

The Russell paradox for properties arises from taking ¬φφ (non-self-exemplification) as a property, the so-called Russell property, and deriving a paradox from asserting that such a property exists. Thus one avoids it by avoiding the existential claim. But there is a further point. One need not even consider such an existential claim, since '¬φφ' is a dyadic *abstract*. Thus the attempt at self-predication involves the purported sentence '¬φφ(¬φφ)'. Since '¬φφ' is dyadic, as are 'φx' and 'x = x', even when the terms are the *same*, '¬φφ(¬φφ)' is not even well-formed, being like 'R(R)' where 'R' is a dyadic predicate. To consider '¬φφ(¬φφ, ¬φφ)' is pointless. For, to arrive at the familiar paradox, one must employ a conversion rule that allows replacing both occurrences of the variable 'φ' in the *predicate* '¬φφ' by the *subject* sign '¬φφ' to arrive at '¬¬φφ(¬φφ)'. But that involves replacing a monadic predicate variable by a dyadic predicate, which mixes logical 'types,' in one of Russell's unproblematic uses of 'type.' This is so irrespective of illicitly assuming that 'φφ' and '¬φφ' represent relations or properties or forms – as some pointlessly and problematically take 'x = x' and '¬x = x' to represent dyadic relations or monadic properties – of *self-identity* and *non-self-identity*. In the latter case we have, at most, the *dyadic relations* of identity and diversity. Such philosophically problematic moves are aided by the formal device of forming 'abstracts' – as in the case of lambda-abstracts. Thus, using '(λx,y)(x = y)' for the identity relation or 'function,' one forms '(λx)(x = x)' to represent the monadic function of self-identity. One then easily moves to '(λφ)φφ' and '(λφ)¬φφ' to arrive at the purported Russell 'property' or function. The device is misleading, first, as such functions cannot be confused with properties, and, second, as forming such signs has no ontological significance whatsoever, unless one postulates that corresponding, and problematic, entities exist. Such issues aside, the basic distinction between monadic and dyadic predicates prevents the Russell paradox for properties without resorting to a hierarchy of types or similar restriction, which removes a ground for claiming arbitrary restrictions are required to avoid logical paradoxes.

References

Barcan, R. (1946) A functional calculus of first order based on strict implication. *Journal of Symbolic Logic*, 11, 1–16.

Bergmann, G. (1947) Undefined descriptive predicates. *Philosophy and Phenomenological Research*, 8, 55–82.

Boolos, G. (1993) *The Logic of Provability*. Cambridge: Cambridge University Press.

Carnap, R. (1942) *Introduction to Semantics*. Cambridge: Harvard University Press.

Carnap, R. (1943) *Formalization of Logic*. Cambridge: Harvard University Press.

Carnap, R. (1946) Modalities and quantification. *Journal of Symbolic Logic*, 11, 33–64.

Carnap, R. (1947) *Meaning and Necessity*. Chicago: University of Chicago Press.

Cellucci, C. (1995) Wittgenstein on the meaning of logical symbols. In R. Egidi (ed.), *Wittgenstein, Mind, and Language*. (pp. 83–91.) Dordrecht: Kluwer.

Gödel, K. (1933) Eine interpretation des intuitionistischen aussagenkalküls. *Ergebnisse eines mathematischen kolloquiums*, 4, 34–40.

Gödel, K. (1986/1995) *Collected Works*, vol. III. S. Feferman *et al.* (eds.) Oxford: Oxford University Press. pp. 334–62.

Hochberg, H. (1957) On pegasizing. *Philosophy and Phenomenological Research*, 17, 551–4.

Hochberg, H. (1984) Russell's attack on Frege's theory of meaning. In H. Hochberg, *Logic, Ontology and Language*. (pp. 60–85.) Munich: Philosophia Verlag (original work published 1976).

Hochberg, H. (1999) *Complexes and Consciousness*. Stockholm. Thales.

Hochberg, H. (2000) Propositions, truth and belief: the Russell–Wittgenstein dispute. *Theoria*, 66, 13–50.

Hughes, G. E. and Cresswell, M. J. (1996) *A New Introduction to Modal Logic*. London: Routledge.

Kanger, S. (1957) *Provability in Logic*. Stockholm: Almqvist & Wiksell.

Lindström, P. (1996) Provability logic – a short introduction. *Theoria*, 62, 19–61.

Lindström, P. (2000) Quasi-realism in mathematics. *The Monist*, 83, 122–49.

Martin-Löf, P. (1996) On the meanings of the logical constants and the justification of the logical laws. *Nordic Journal of Philosophical Logic*, 1, 11–60.

Owen, G. E. L. (1986) Plato on not-being. In G. E. L. Owen, *Logic, Science and Dialectic: Collected Papers in Greek Philosophy*. Ithaca, NY: Cornell University Press.

Quine, W. V. O. (1939) Designation and existence. *Journal of Philosophy*, 36, 701–9.

Quine, W. V. O. (1953) *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

Russell, B. A. W. (1956a) On denoting. In B. A. W. Russell, *Logic and Knowledge*, R. Marsh (ed.) London: Allen & Unwin (original work published 1905).

Russell, B. A. W. (1956b) The philosophy of logical atomism. In B. A. W. Russell, *Logic and Knowledge*, R. Marsh (ed.) London: Allen & Unwin (original work published 1918).

Russell, B. A. W. (1956c) On propositions: what they are and how they mean. In B. A. W. Russell, *Logic and Knowledge*, R. Marsh (ed.) London: Allen & Unwin (original work published 1919).

Russell, B. A. W. (1984) *Theory of Knowledge: The Collected Papers of Bertrand Russell*, vol. 7, E. Eames *et al.* (eds.) London: Allen & Unwin.

Taylor, A. E. (1971) *Plato: The Sophist and the Statesman*, R. Klibansky and E. Anscombe (eds.). London: Dawsons.

Whitehead, A. N. and Russell, B. A. W. (1950) *Principia mathematica*, vol. 1. Cambridge: Cambridge University Press.

Wittgenstein, L. (1961) *Tractatus Logico-Philosophicus*. D. F. Pears and B. F. McGuinness (trans.). London: Routledge & Kegan Paul.

Further Reading

Bergmann, G. (1960) The philosophical significance of modal logic. *Mind*, 69, 466–85.
Bosanquet, B. (1948) *The Essentials of Logic*. London: Macmillan.
Carnap, R. (1937) *Logical Syntax of Language*. London: Routledge & Kegan Paul.
Hintikka, K. J. J. (1963) The modes of modality. *Acta philosophica fennica*, 65–81.
Kripke, S. (1980) *Naming and Necessity*. Oxford: Blackwell.
Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.
Sainsbury, M. (1991) *Logical Forms*. Oxford: Oxford University Press.

19

Putting Language First: The 'Liberation' of Logic from Ontology

ERMANNO BENCIVENGA

There are two ways of conceiving the relation between language and the world: they differ by making opposite choices about which of them is to be assigned priority, and which is to be dependent on the other. The priority and dependence in question here are conceptual, not causal: at the causal level everyone agrees that certain portions of the nonlinguistic world (intelligent entities, say) must be in place before meaningful expressions come to pass, so what we are concerned with is how *the notion* of meaningful is to be understood – whether 'meaningful' is defined as something that *means* some portion of the world or rather as something that belongs to a self-sufficient structure of analogies and oppositions. For example, taking for granted that there would be no meaningful expression 'John' unless some intelligent entity came up with it, is 'John' a meaningful expression *because* there is a John that it means or rather *because* it is a certified component of the English language, categorized as a name and clearly distinct from 'Paul' – though somewhat analogous to 'Jack'? If you go the first route, I will say that you are a *realist* at the conceptual (or transcendental) level; if you go the second one, I will call you a conceptual (or transcendental) *idealist*. 'Realist' is a transparent term, since 'res' is 'thing,' 'object,' in Latin and clearly this kind of realist puts things (conceptually) first, considers them basic in her logical space; 'idealist' is more controversial, since the 'idea' in it recalls a psychologistic jargon that is not as popular today as it once was, so one might think that some other root, more clearly expressive of the semantical, logico/linguistic character of the current analysis, should be preferred. And yet, once we are clear about its implications, 'idealist' remains a better choice because it lets us see the connections of this contemporary debate with other, classical ones; later I will explore some such connections. Before I do that, however, I have to explain what the contemporary debate looks like.

My example of a meaningful expression above was not chosen at random: in the case of names there is more agreement than with any other part of speech concerning *what* they mean. 'John' means a (male) human being, 'Lassie' means a dog, 'the Queen Mary' means a ship, and in general a name that means anything means an object – or, as people say, *denotes* it or *refers to* it (the terminology is highly unstable: 'reference' and 'denotation' are used as translations of the Fregean 'Bedeutung,' but 'meaning' is also used for the same purpose, consistently with Frege's own suggestion, and indeed it is the most natural English counterpart of this perfectly ordinary German word). There

are complications here, since names may be ambiguous and the objects meant may be past or future as well as present ones, but none of that touches the essence of names' favored condition: what *kind* or *category* the meaning of a name belongs to is hardly ever an issue, much less so than, say, with predicates or connectives. Probably because of this (and of the great importance that concrete, middle-sized objects like human beings, dogs, and ships have in our form of life), it is around names that the realism/idealism controversy has surfaced in the clearest form within contemporary logic. And *free logics* have been its most conspicuous outcome.

Standard quantification theory (SQT) lets us prove theorems like

(1) $Pa \supset \exists x Px$,

one of whose substitution instances would seem to be

(2) If Pegasus is a winged horse then there exists a winged horse.

And that would seem to be a problem, because it would seem that the antecedent of (2) is true and its consequent is false. As it turns out, no such problem arises if you are a conceptual realist; for then you are faced by a simple, unproblematic dilemma. Either, that is, Pegasus exists, and hence there does exist a winged horse after all (it is not for logic to say what the criteria for existence are, you might add: it may well be that being described in a story is a sufficient condition for existing), or Pegasus does not exist, *and then (2) is not a substitution instance of (1).* For the individual constant a in (1) (I will take symbols to be autonymous, that is, names of themselves) is supposed to stand for an arbitrary name, and in a realist framework nothing can be a name unless there is an object it names. Consistently with this view, in SQT a must receive an interpretation in all models; in every possible world there must be an object it refers to – an object *existing in that world.* If identity is added to the language, we can give direct expression to this semantical condition by proving the theorem

(3) $\exists x(x = a)$.

As for the *deceptive appearance* that 'Pegasus' is a name (and (2) is a substitution instance of (1)), that will have to be dispelled, which can be done in one of two major ways. You can either assign to 'Pegasus' a *conventional* reference, thus *making it* a name (Frege (1893) used this strategy for his formal language, and Carnap (1947) generalized it to natural languages), or agree with Russell that (a) 'Pegasus' is a disguised definite description, (b) definite descriptions (disguised or otherwise) are *incomplete symbols*, that is, have no meaning in isolation but only in context, and (c) once the appropriate contextual *definiens* is provided for the sentence that has the 'grammatical form' (2), no apparent (and confusing) reference to Pegasus occurs in it (none could, because *there is* no Pegasus for anything – or anyone – to refer to).

When the models of SQT graduate into the possible worlds of modal semantics (that is, bring out explicitly what they were to begin with), one might be displeased by the provability of stronger variants of (3). For the same unobjectionable

(4) $a = a$

which lets us infer (3) by existential generalization would now (by existential generalization and necessitation) let us infer both

(5) $\Box\exists x(x = a)$

and

(6) $\exists x\Box(x = a)$,

that is, both that necessarily there is an a (*whatever* name a might be) and that there is an object which is necessarily (or *essentially*) a – that a refers to the very same object in all worlds (and, again, a could be *any* name). But there are those who are perfectly happy to live with these extreme consequences of realism: they accept the Barcan formula

(7) $\Box\forall xA \rightarrow \forall x\Box A$

and hence are prepared to admit that all possible worlds have the same domain of objects – that a 'possible world' is a particular distribution of accidental properties over *these* objects, the objects existing in *this* world, or more simply *the* objects.

If, on the other hand, you are an idealist, then you do not think that something is a name because it names an object; your definition of a name is not object-based (but language-based: a name is an expression belonging to a certain grammatical category), and it is perfectly possible that an expression satisfy this definition while there is no object that it names. So, if 'Pegasus' is taken to be one such *nondenoting name*, (2) *will* be a substitution instance of (1) and its falsity will be enough to disprove the logical truth of (1). Hence you will be forced to modify (specifically, to weaken) SQT so as to rule out that (1) be a theorem; a free logic is what results from this modification.

A (further) schematization of (1), that is,

(8) $A(a/x) \supset \exists xA$

(the Principle of Particularization) is often an axiom of SQT. When it isn't, some equivalent basic assumption is present – most typically, either the Principle of Specification

(9) $\forall xA \supset A(a/x)$

or the already cited Rule of Existential Generalization

(10) $\dfrac{A(a/x)}{\exists xA}$

or the Rule of Universal Instantiation

$$(11)\quad \frac{\forall x A}{A(a/x)}\,.$$

For definiteness, I will assume a system SL containing axiom (9).

So the very first thing we need to do, to transform SL into a free logic, is to drop (9). But that is hardly enough: though from an idealist point of view it is illegitimate to infer

(12) Pegasus is not a winged horse ($\sim Pa$)

from

(13) Nothing is a winged horse ($\forall x \sim Px$),

it is perfectly legitimate to infer

(14) The Queen Mary is not a winged horse ($\sim Pb$)

from it. For, remember, this idealism is the conceptual variety; so it is not supposed to impact your ordinary, empirical sense of what does exist. The Queen Mary exists, hence whatever is true of everything existing is true of it. Therefore, (9) cannot *just* be dropped: it must be *replaced* by a weaker variant of it, saying in effect

(15) $(\forall x A \wedge a \text{ exists}) \supset A(a/x)$.

If identity is part of the language, (15) can be expressed as

(16) $(\forall x A \wedge \exists y(y = a)) \supset A(a/x)$;

if it isn't, E! is introduced as a symbol for existence and (15) becomes

(17) $(\forall x A \wedge \mathrm{E}!a) \supset A(a/x)$.

Nor is that enough either: just as you don't want your new understanding of what a name is to force you to deny that 'the Queen Mary' does name an object, and, in general, to force you to admit fewer objects than the conceptual realist does, you also don't want to be forced to admit more. Your objects – though differently conceptualized – will be the very same ones as before, the *existing* ones; your idealism will not lead you to introducing into the world any creatures of fancy. It will continue to be true for you that all objects exist, that is,

(18) $\forall x \mathrm{E}!x$;

and, since (18) is not provable on the basis of the axioms you have allowed so far, it (or some equivalent principle) will also have to be adopted as an axiom.

During the first phase of their history, from the mid-1950s to the mid-1960s, free logics were developed much as I did so far, at a purely proof-theoretical level. The formal systems were justified by intuitive semantical considerations, which sounded convincing to some and gratuitous to others (Church (1965), for example, pointed out that, formally, one of these systems was a simple exercise in restricted SQT and hence claimed that there was nothing 'distinctive' about it), and, because the correspondence of standard and free logics, respectively, to the realist and idealist, largely incommensurable paradigms was not perceived, inarticulate invocations of 'natural' and 'unnatural' consequences provided what little ground there was for philosophical discussion. But, eventually, the task of defining a formal semantics could no longer be postponed; and here is where the conceptual issues started (however slowly) to come to the fore.

The semantics for SQT allows for no immediate extension that would assign a truth-value to a sentence like

(19) Pegasus is white.

For, in that semantics, (19) is true if the object Pegasus belongs to the set W of white things, and false if Pegasus does not belong to W; but, if there is no Pegasus, we are stuck. Formally, a model for SQT is an ordered pair $\langle D, f \rangle$, where D (the *domain*) is a nonempty set and f (the *interpretation function*) maps the individual constants of the language into D and the n-place predicates into D^n, and an atomic sentence

(20) $Pa_1 \ldots a_n$

is true if

(21) $\langle f(a_1), \ldots, f(a_n) \rangle \in f(P)$

and false if

(22) $\langle f(a_1), \ldots, f(a_n) \rangle \notin f(P)$.

But, if some of the $f(a_i)$ are not defined, the expression preceding the membership sign is not defined either, nor is a truth-value for (20).

A simple solution for this problem (adopted, for example, by Schock (1964) and Burge (1974)) would consist of saying that, when Pegasus does not exist, it is automatically not the case that Pegasus belongs to W; hence (19) is false. Formally, one would consider (20) true if (21) is the case, *and false otherwise* – which includes all cases in which some of the relevant expressions are undefined. As a result, all atomic sentences containing nondenoting names would be false, and truth-values for complex sentences could then be computed in a straightforward manner. But, as an indication of how quickly dealing with nonexistents gets us into deep and controversial philosophical issues, this solution (call it S) forces upon us a very definite (and, to some, objectionable) view of the relation between natural and formal languages. For consider

(23) Pegasus is not white.

If (23) is paraphrased as

(24) $\sim Pa$,

then according to *S* (23) is true; but is there any reason why (23) could not *also* be paraphrased as

(25) Pa,

where *P* stands for the predicate 'not white'?

Some authors (the ones Cocchiarella (1974) calls *logical atomists*) would answer 'No': they would claim that the paraphrase of an ordinary sentence into a formal language must bring out the 'logical form' of that sentence, hence the paraphrase of (23) into the formal language of quantification theory *must* be (24) – since (23) *is* a negation. But others (the ones that in Bencivenga (1981a) I call *logical pragmatists*) would agree with van Fraassen (1969: 90) that "the symbolization of a sentence as . . . [atomic] indicates only the extent to which its internal structure has been analyzed, and the depth of the analysis need only be sufficient unto the purpose thereof"; hence they would claim that (23) can be alternatively paraphrased as (24) or (25) depending on what the context requires. If, for example, we are trying to decide on the validity of the argument

(26) Pegasus is not white.
Therefore, something is not white.

there is no reason for the relevant paraphrase to use a negation sign. Since, on the other hand, there is also no compelling reason why this paraphrase should *not* use a negation sign, the unacceptable consequence follows that (26) is valid or invalid depending on how we *decide* to paraphrase (23) – if we paraphrase it as (25) then the premise of (26) is false when there is no Pegasus, hence the conclusion follows vacuously from that premise (and of course the conclusion always follows when Pegasus exists); if we paraphrase it as (24) then the premise is true when there is no Pegasus and the conclusion is false in a world where *W* is empty. Thus the 'straightforward' solution *S* of a specific problem concerning nondenoting names would end up 'resolving' also such a general philosophical issue as the debate between logical atomism and pragmatism!

Those who find the outcome above disturbingly close to magic might be attracted by another simple way out. We have a problem here because the interpretation function of a model is not always total: it may be undefined for some individual constants. Let us agree then to make it always total, by adding to each model an additional domain (called the *outer domain*) and interpreting there the constants that remain undefined in the ordinary domain of quantification (the *inner domain*). Formally, a model in *outer domain semantics* (whose classical text is Leblanc and Thomason (1968)) is an ordered triple $\langle D, D', f \rangle$, where D and D' are disjoint sets and f maps the individual

constants into $D \cup D'$ and the n-place predicates into $(D \cup D')^n$. The most obvious reading of the outer domain makes it the set of nonexistent objects; after all, consistency requires that

(27) $E!a$

be false whenever $f(a) \in D'$. And here again we encounter serious philosophical problems. For whether 'there are' (in whatever sense of that phrase) any nonexistent objects would seem to be a metaphysical issue, and one that logic should remain neutral about; otherwise, it is hard to see how people holding opposite positions on this issue could even reason with one another. But the current 'simple' approach to our problem decides the issue, making nonexistent objects a logical necessity. The realism/idealism controversy I placed in the background of free logics lets us diagnose the confusion from which this new bit of magic originates. Free logics only make sense from an idealist point of view; but outer domain semantics continues to think of objects as conceptually primary – it continues to think of the notions of truth, validity, and the like as dependent upon the notion of an object. So, when objects are not available, it brings them in anyway: nonexistent objects, which is as much as saying nonobjective objects, an oxymoron demanded by the awkward, brutal superposition of two distinct *and conflicting* conceptual schemes. (We will see shortly that a more careful and discriminating operation relating the two schemes has much better chances of improving our understanding.)

It begins to look as if 'simple' and 'straightforward' in this context have a tendency to keep dangerous company with 'conceptually confused.' There is a good reason for that. Formal semantics as we know it is already biased toward realism: its starting point is indeed a domain, a set of *objects*, on which nothing is said and no conditions are imposed – 'object' is a primitive notion in the realist's logical space, hence one that is going to enter into the definition of (all) others but that itself cannot be defined. Being directly expressive of the idealist's standpoint would require conceiving of semantics in an entirely new way, or maybe even challenging the very enterprise of semantics – insofar as the latter is seen as the project of accounting for central logical notions in terms of an *objective* interpretation of the various parts of speech. But, despite some vague gestures in this direction (as when Meyer and Lambert (1968) say that 'Pegasus is a horse' is true not because Pegasus is a horse but because 'Pegasus' is a horse-word), no such revolutionary work is in sight. What has happened, instead, is that a number of authors have painstakingly translated portions of the idealist framework into the realist one – which may be a more immediately useful job, given the prominence of realism in contemporary philosophy of logic, hence the expediency of gradually forcing that prominent position away from itself, as opposed to just building an alternative structure next door, where few are likely to look for it. Needless to say, the results of these translations have been anything but straightforward, which some kibitzers have considered a good ground for criticism.

The translation job was initiated (not under that description) by van Fraassen (1966a, 1966b) with his semantics of *supervaluations*. Given a partial model $M = \langle D, f \rangle$ (that is, a model whose interpretation function is partial on the individual constants), M will leave all atomic sentences containing nondenoting names truth-valueless. A

classical valuation over M is any result of 'completing' M by assigning arbitrary truth-values to all those atomic sentences (van Fraassen thinks of each such valuation as the result of adopting a specific 'convention'). A classical valuation can be routinely extended to a valuation of all sentences (and we can continue to call this extension a classical valuation); the *supervaluation over M* is the logical product of all classical valuations over M. Thus, for example, assume that $f(a)$ is defined, $f(b)$ is undefined, and $f(a) \in f(P)$, and consider the following sentences:

(28) Pb
(29) $\sim Pb$
(30) $Pa \wedge Pb$
(31) $Pa \vee Pb$
(32) $Pb \vee \sim Pb$
(33) $Pb \wedge \sim Pb$.

In some classical valuations over M (28) will be true and in some it will be false, hence in the logical product of all these valuations (28) will remain truth-valueless. The same can be said about (29) and (30); but when it comes to (31)–(33) the situation is different. Whatever truth-value the second disjunct of (31) might have in a classical valuation over M, its first disjunct will always be true there, hence the disjunction will always be true, hence the logical product of all these valuations will make (31) true. Also, whatever truth-value the first disjunct of (32) might have in a classical valuation (over any model), its second disjunct will have the *opposite* truth-value, hence the disjunction will always be true and will remain true in every supervaluation. For similar reasons, every supervaluation will make (33) false.

The problem with van Fraassen's approach is that it is formulated at a sentential level of analysis. 'Conventions' assign arbitrary truth-values to unanalyzed atomic sentences, hence have no way of bringing out the logical structure of those sentences – no way of accounting for the specifics of *predicate* (let alone *identity*) logic. Since van Fraassen wants principles like (16) and (17) – or, for that matter, like the uncontroversial

(34) $\forall x(A \supset B) \supset (\forall x A \supset \forall x B)$

and

(35) $a = a$

to turn out logically true, his only chance is to *make* them so by imposing (*meta*conventional?) requirements on classical valuations. One can do better by developing this approach at a quantificational level of analysis; that is, by thinking not so much of arbitrary assignments of truth-values to truth-valueless atomic sentences as of arbitrary assignments of *denotations* to nondenoting names. One will then talk not of conventions and classical valuations over a model M, but rather of *extensions* of M which provide an interpretation for all the individual constants undefined in M; and the logical

product which constitutes the final valuation for M (and which can still be called the supervaluation over M) will be defined over these extensions. Then (34) and (35) will be (logically) true for the same reason for which (31) and (32) are.

In this reformulation, the translation character of supervaluational semantics can be made clearer. For the reformulation starts, in true realist fashion, by accounting for the truth-values of (previously truth-valueless) sentences in terms of objects; but then, by playing all possible characterizations of these objects against one another, it cancels what is specific to any of them, and what finally emerges as true or false does so simply as an expression of the rules constituting *the language itself*. (31) and (32) are true and (33) is false because of how negation, disjunction, and conjunction behave, not because of what b is; since we are working in a realist framework, we must mobilize a reference for b to activate the logical behavior of negation, disjunction, and conjunction, but once we have activated it that reference can (in effect) be discounted and forgotten. To use terminology coming from a different quarter, the reference of b is only an *intentional object*: not the realist's object, that is, but still enough of a presence for him/her to apply familiar conceptual moves, to allow him- or herself enough maneuvering space, as it were – it is enough for concepts that he/she takes to be dependent ones (negation and the like) to be called in play. After they *are* called in play, the rug can be pulled from under his/her feet (this object by courtesy can be recognized for what it is – that is, not an object at all) and (hopefully!) the structure thus built will remain standing.

But there is a complication. Return to (1) – one of the characteristic schemes marking the distinction between standard and free quantification theories. You would expect (1) to be falsified in some partial model; so, to test your expectation, consider a model $M = \langle D, f \rangle$ where $f(a)$ is undefined and $f(P)$ is empty, and an extension $M' = \langle D', f' \rangle$ of M where $f'(a) \in f'(P)$ – intuitively, in terms of the substitution instance (2), a world in which there exist no winged horses and an extension of that world in which Pegasus exists and is a winged horse. To begin with, (1) is truth-valueless in M, but once we move to M' it becomes true (Pegasus is a winged horse there but also exists there, hence *there exists a winged horse*). And, since it is easy to see that the same holds for every extension of M (M' was selected as the most representative case – if Pegasus is not a(n existing) winged horse then the conditional is automatically true), the conclusion is that (1) is true in the supervaluation over M. And, again, *this* result can be generalized; so (1) ends up being *logically* true, and we are back in standard quantification theory.

What has gone wrong? That we have taken our maneuvering space too seriously, we have treated the objects by courtesy as fully-fledged objects. As I explain in Bencivenga (1987), the same danger arises in Kant's transcendental philosophy. Because that philosophy is a delicate combination of transcendental idealism and empirical realism, it too needs to translate from one framework into the other – specifically, to translate realist criteria of objectivity into transcendental idealism. Intentional objects are a tool that can be used in carrying out the translation; but it is important to realize that, when this task is completed, they are not to be counted as objects at all. Intentional objects are not more a *kind* of objects than *alleged objects* are; they are only a manner of speaking. The only objects are the existing ones, the objects *simpliciter*; who thinks otherwise

is going to fall into the trap of assigning objective status to God, the soul, the world, and other fictional entities. The (conceptual) 'construction' of objects *simpliciter* takes time, and during this time intentional objects play a role; but by the end they are supposed to disappear. (In fact, Kant's case is more complicated: the end is never reached and all we are ever left with are objects by courtesy – *phenomena*. But here we can disregard these additional complications.)

The situation we are facing in supervaluational semantics is analogous. When evaluating a sentence containing nondenoting names in a model M, we (in effect) perform a mental experiment over M: we imagine, that is, that those names be denoting – that the relevant objects exist. We do not *believe* they do; they have only an instrumental role to play, and eventually we want to get rid of them. But (and here is where the problem arises), during the mental experiment these 'objects' will not sit quietly and let us exploit them: they will often *contradict* what real information we have, about objects in whose existence we *do* believe. If we let them do that, by the time we are ready to drop them they will have already done enough damage; hence we must prevent any such interference – never 'revise' what we already know on the basis of a mental experiment. To return to the example above, we do know that there exist no winged horses; so when we move to M' we don't expect this (imaginary) model to tell us anything new *about that*. It is Pegasus that we need to 'know' about – more precisely, it is Pegasus that we need to bring in so that our rules for conditionals can operate appropriately. We need a truth-value for the antecedent of (1), and that is why we are moving to M'; we do not need a truth-value for the consequent because we already have one, a *real* one, so whatever truth-value M' might assign it we are not interested in it. How can we formalize this distinction of levels? How can we provide a rigorous articulation of the purely instrumental role of the mental experiment – which is supposed to be canceled out at the end and in the meantime is not supposed to challenge what is the case for objects *simpliciter*, the only objects there really are? The answer is provided in Bencivenga (1980, 1981, 1986), where the main formal tool is the *valuation for M' from the point of view of M*, $v_{M'(M)}$, defined as follows (I will only give the base of the recursive definition, for an atomic A; the rest is routine; v_M and $v_{M'}$ are ordinary valuations for M and M', respectively, defined as usual):

(36) $v_{M'(M)}(A) = v_M(A)$ if $v_M(A)$ is defined;
$v_{M'(M)}(A) = v_{M'}(A)$ otherwise.

That is, this valuation gathers all the truth-values available in the real world and adds fictional ones to them *only where there are gaps*, thus implicitly resolving all conflicts in favor of reality and leaving only an instrumental role to fiction. As a result, the antecedent of (1) is true in it (because it is undefined in M and true in M') and its consequent is false (because it is false in M, hence we need not look any further). And the supervaluation is constructed over all $v_{M'(M)}$, where M' is an extension of M, thus making (1) *not* logically true. I call the fundamental assumption embodied in (36) the Principle of the Prevalence of Reality: real information is always to prevail over fictional 'data.' That such a principle be proffered in a context of transcendental idealism will be found surprising only by those who forgot Kant's relevant advice: it is precisely realism at the conceptual, transcendental level that is responsible for the *empirical* ide-

alism of people like Berkeley – that is, for the absurd thesis that the world *is constituted by* ideas.

The deep and subtle connections between free logics (and especially free semantics) and idealism should be apparent by now; in closing, I want to bring out another sense in which free logics evoke fundamental philosophical tensions – and in this respect, too, side with Kant, though here the opposition includes not only realists but also another major brand of idealists, the Hegelian, 'absolute idealism,' variety. When motivating the enterprise of 'liberating' logic from ontological commitments, people typically voiced (more or less explicitly) such pragmatic criteria as I have myself hinted at above: logic must be a neutral tool, one must be able to use the same logic when reasoning across opposite metaphysical positions, one needs to leave the denotational character of a name (say, 'God') open while one (for example) proceeds to *prove* that there exists a denotation for it. All of that makes sense; but the 'liberation' in question involves much more than pragmatics – indeed it is liberation of a much more literal kind than one might expect.

That something which is not the case be logically possible is, as Kant (1964) noted, a main instrument of defense against the strictures of reality; and, one might add, it creates an arena for the imaginative exploration of alternatives to reality, maybe even (who knows?) for the overcoming of those strictures. The opposite tendency is expressed by those who want to make reality itself – as much as possible of it – a matter of logical necessity. Requiring that all worlds have the same domain of objects is only the first step in that direction; but the first step is also the most important one, and the one at which the most vigorous resistance must be mounted – if one is going to be unhappy with the final outcome. If no resistance is forthcoming, the fixity of 'natural kinds' and other predicates will come next; one will begin to assert that "[w]ith possibilities, less is more" (as does Almog 1991: 622, summarizing the conventional wisdom of the Kaplan school), and will find oneself agreeing in fact (if not, God forbid!, in principle) with the Hegelian reduction of history to logic – except that the old German master did a much more thorough job of it.

In light of the above, it is a sobering exercise to remind oneself of these famous words contained in the preface to Hegel (1991: 21–2): "To comprehend *what is* is the task of philosophy, for *what is* is reason. . . . If . . . [a philosopher's] theory builds itself a world *as it ought to be*, then it certainly has an existence, but only within his opinions – a pliant medium in which the imagination can construct anything it pleases." For then one realizes that the space made available by free logics to a coherent thinking and talking about nonexistents is not only of value for reflection and argument concerning mythological winged horses; the very plausibility of normative and utopian claims requires that breathing room. What sense would it make to blame some behavior as morally wrong if this were the only possible world? How could we legitimately long for a state based on 'people's rational choice' if that nondenoting singular term (a nondenoting definite description, for which free logics provide as much of an account as they do for nondenoting names) were destined to remain nondenoting? Which of the objects existing here – *the* objects, that is – do we expect to *become* a denotation for it? *Could* any of *these* objects do so? When we appreciate the point implied by such rhetorical questions, we will see that free logics are the first, still quite tentative, but also absolutely crucial step toward a logic *of the free*.

References

Almog, Joseph (1991) The plenitude of structures and scarcity of possibilities. *Journal of Philosophy*, 88, 620–22.

Bencivenga, Ermanno (1980) *Una logica dei termini singolari*. Torino: Boringhieri.

Bencivenga, Ermanno (1981a) On secondary semantics for logical modalities. *Pacific Philosophical Quarterly*, 62, 88–94.

Bencivenga, Ermanno (1981b) Free semantics. *Boston Studies in the Philosophy of Science*, 47, 31–48.

Bencivenga, Ermanno (1986) Free logics. In Dov Gabbay and Franz Guenthner (eds.), *Handbook of Philosophical Logic*, vol. III (pp. 373–426). Dordrecht: Reidel.

Bencivenga, Ermanno (1987) *Kant's Copernican Revolution*. New York: Oxford University Press.

Burge, Tyler (1974) Truth and singular terms. *Noûs*, 8, 309–25.

Carnap, Rudolf (1947) *Meaning and Necessity*. Chicago: University of Chicago Press.

Church, Alonzo (1965) Review of "existential import revisited," by Karel Lambert. *Journal of Symbolic Logic*, 30, 103–4.

Cocchiarella, Nino (1974) Logical atomism and modal logic. *Philosophia*, 4, 41–66.

Frege, Gottlob (1893) *Grundgesetze der Arithmetik*, vol. I. Jena: Verlag Hermann Pohle.

Hegel, Georg (1991) *Elements of the Philosophy of Right*. (H. B. Nisbet, trans.). Cambridge: Cambridge University Press (original work published 1821).

Kant, Immanuel (1964) *Groundwork of the Metaphysic of Morals*. (H. J. Paton, trans.). New York: Harper & Row (original work published 1785).

Leblanc, Hugues, and Thomason, Richmond (1968) Completeness theorems for some presupposition-free logics. *Fundamenta Mathematicae*, 62, 125–64.

Meyer, Robert, and Lambert, Karel (1968) Universally free logic and standard quantification theory. *Journal of Symbolic Logic*, 33, 8–26.

Schock, Rolf (1964) Contributions to syntax, semantics, and the philosophy of science. *Notre Dame Journal of Formal Logic*, 5, 241–89.

van Fraassen, Bas (1966a) The completeness of free logic. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 12, 219–34.

van Fraassen, Bas (1966b) Singular terms, truth-value gaps, and free logic. *Journal of Philosophy*, 67, 481–95.

van Fraassen, Bas (1969) Presuppositions, supervaluations, and free logic. In Karel Lambert (ed.), *The Logical way of Doing Things* (pp. 67–91). New Haven, CT: Yale University Press.

Part VII

METATHEORY AND THE SCOPE AND LIMITS OF LOGIC

This page intentionally left blank

20

Metatheory

ALASDAIR URQUHART

1 Introduction

The older tradition in mathematical logic, represented by the foundational work of Frege, Whitehead, and Russell, was one in which the aim was to develop as large a part of mathematics as possible within a fixed axiomatic system. In general, questions that fell outside the basic system (such as the system of type theory on which *Principia Mathematica* is based), were ignored.

Under the influence of the great German mathematician David Hilbert, a new approach became influential in the 1920s, sometimes called 'metamathematics' or 'metalogic.' This new approach, in contrast with the earlier, can be described as critical in spirit, both in the sense that the underlying ideas showed a strong Kantian influence, but also in that the trend was towards analysing logical systems from the outside, rather than working within a fixed system of axioms. As a consequence of this change in direction, logic became a much more mathematical subject than formerly, a trend that continues to this day. The results that emerged from the research program of the Hilbert school remain among the most striking in all of logic.

In the present chapter, we describe these results in non-technical language, and indicate their philosophical significance. They are in many cases of a negative character, showing that the optimistic goals of Hilbert's foundational program could not be achieved. Nevertheless, a central concept emerged from this research activity, that of computability. The truly remarkable fact that this concept, in contrast to notions like that of provability and definability, does not depend on the system with respect to which it is defined, but is in a certain sense absolute, is fundamental to modern computer science and technology.

We begin with an outline of Hilbert's program in the foundation of mathematics, the achievements of Gödel that contributed positively to Hilbert's aims (the completeness theorem) and results like his incompletenesss theorem that showed the original aims to the program to be untenable, and led to its demise, at least in its original form. The essay also discusses the concept of computability that emerged in the 1930s in the wake of the incompleteness theorems, and the resulting clarification of the extent to which logic can be considered a purely formal subject. It concludes

with a discussion of the philosophical bearings of the basic results, in particular the question of the absolute or relative nature of logical concepts.

2 Hilbert's Program

David Hilbert (1862–1943) dominated German mathematics in the first half of the twentieth century. His formalist program, which attained its classical formulation in the 1920s, was intended to provide a final solution to the foundational problems that had arisen in the wake of the debates in the foundations of set theory, and the constructivist criticisms of the Dutch intuitionist L. E. J. Brouwer (1881–1966).

Brouwer had severely criticized the free use of classical logic as applied to infinite structures, and in particular the law of excluded middle. Given a constructive reading of the logical particles, the law of excluded middle can be read as asserting the universal solubility of all mathematical problems (that is to say, 'P or not P' asserts that we have either a constructive proof of the proposition P, or a constructive refutation of P). Since there is no warrant for this belief, Brouwer rejects the applicability of classical logic in general.

Hilbert presented himself as the champion of classical methods in mathematics, making such ringing assertions as the following from an address of 1927:

> Taking the principle of excluded middle from the mathematician would be the same, say, as proscribing the telescope to the astronomer or to the boxer the use of his fists.
>
> (van Heijenoort 1967: 476)

Hilbert spelled out his program in detail in a series of addresses from the 1920s that can be found in translation in van Heijenoort's collection of basic logical texts. He accepted (in a sense) the constructivist criticism of classical logic, since he denies the existence of the actual infinite. However, he wished to keep the powerful deductive tools of classical logic and set theory, and so was forced to adopt an indirect strategy of justification.

The essentials of Hilbert's formalist program are as follows. Classical mathematics is to be given a complete and fully rigorous formulation by employing the resources of mathematical logic (making use of the work already done in this area by Frege, Whitehead, and Russell). However, not all the statements occurring in such systems are held to be directly meaningful. In particular, purely existential statements are to be read as infinite disjunctions, and so we cannot directly attribute a constructive meaning to them. The part of mathematics that is directly meaningful for Hilbert he describes as the part consisting of finitary inferences, operating on concrete objects consisting of strings of symbols.

> If logical inference is to be reliable, it must be possible to survey these objects completely in all their parts, and the fact that they occur, that they differ from one another, and that they follow each other, or are concatenated, is immediately given intuitively, together with the objects, as something that neither can be reduced to anything else nor requires

> reduction. This is the basic philosophical position that I consider requisite for mathematics and, in general, for all scientific thinking, understanding and communication.
>
> (van Heijenoort 1967: 376)

Thus the main goal of Hilbert's program can be stated as the solution of the *consistency problem*. We can regard the symbol sequences constituting the formalized version of mathematical assertions as purely formal objects. It is then a mathematically well-defined problem to show that a sequence of such objects satisfying the (purely formal) conditions for being a correct proof cannot end with an obvious contradiction like '0 = 1.' To be a fully convincing demonstration, and avoid the charge of circularity, the proof must itself be based only on finitary reasoning. Hilbert hoped that by completing such a consistency proof he would achieve a final vindication of classical mathematics.

Hilbert was a congenital optimist, and in particular believed strongly in the solvability of all mathematical problems, a faith that expressed itself in the phrase he used as the conclusion of his last major public address: "We must know. We shall know" (Ewald 1996: 1165). This credo forms the background to another major problem of the Hilbert school, the *Entscheidungsproblem* or decision problem. The problem here is to decide by a mechanical, algorithmic procedure for a given formula of first-order predicate logic, whether it is logically valid or not.

If there were a positive solution to this problem, this would have extraordinarily far-reaching consequences. In particular, all known mathematical theories can be formalized in terms of finite sets of axioms in first-order logic. If the decision problem were solvable, then it would be possible for any such theory to decide whether a given sentence is a theorem simply by forming the implication that has the conjunction of the axioms as the antecedent, and the sentence as the consequent, and testing this implication for validity. Hence, all mathematical problems would be solvable in principle by a purely mechanical procedure. Thus Hilbert's belief in the solvability of all problems would be true, and what is more, in an extremely strong sense, since arbitrary problems could be solved without the intervention of human ingenuity.

A final problem that is of a somewhat subsidiary character, but fits naturally into Hilbert's formalist viewpoint, is the problem of completeness for first-order predicate logic. The problem was originally posed by Hilbert and Ackermann in their textbook of logic of 1928. We can define validity in predicate logic in two different ways, syntactically or semantically. The first definition of validity defines it in terms of derivability from a fixed set of axioms or rules, such as those originally proposed by Frege. The second definition defines it as truth in all possible interpretations. The second notion is not a finitistically meaningful notion, since it refers to the infinite totality of all possible interpretations. The question is nevertheless a natural one for Hilbert to ask, since it equates an infinitistic notion with a purely finitary, combinatorial notion.

In the 1930s, decisive progress was made on all three problems described above. As an unexpected bonus, there emerged for the first time, a completely precise and absolutely general notion of a mechanical or algorithmic procedure. In the following sections, we shall describe the dramatic developments of this decade.

3 Gödel's Theorems

In 1930 and 1931, Kurt Gödel solved the completeness problem for predicate logic, and made a basic contribution to the consistency problem. It was in the course of trying to solve the latter problem that he made the unexpected discovery that any axiomatic system containing a certain minimal part of number theory was necessarily incomplete.

Gödel's proof that the formal rules originally given by Frege are complete for the semantical concept of logical validity in first-order logic was published in 1930 (Gödel 1986: 102–23). Since then, many different versions of the proof have been given. Perhaps the most intuitively understandable form of the proof is to consider it as arising from the systematic search for a counter-model. This is the approach adopted when using the well-known formalism of semantic tableaux – currently employed in many introductory texts.

In the approach to completeness using analytic tableaux, the basic formalism consists of a tree labeled with sentences. We label the root of the tree with the negation of the formula in which we are interested. Each branch in the tree can then be considered as part of a search for a model that makes the negated formula labeling the root true. For example, if a branch in the tree contains an existential sentence $\exists x\, Fx$, then we extend the branch by adding an instance Fa. Similarly, if a branch contains a disjunction $A \vee B$, then we split the branch into two branches, one containing A, the other B. This search for a model must be carried out in a systematic way – we omit the details here. Provided the search is in fact systematic, then completeness can be seen to hold in the following sense. Either the search ends in failure, so that each branch terminates with an explicit contradiction, or this does not happen, in which case a model can be seen to exist. In the first case, the labeled tree is a proof of validity of the starting formula. For the details of this version of Gödel's completeness theorem, the reader is referred to Smullyan's elegant monograph of 1968.

The completeness theorem constitutes a vindication of Hilbert's formalist program, since it gives a purely syntactical, formal equivalent for the non-constructive concept of semantical validity. Gödel's next result, his great incompleteness theorem, threw in doubt most of Hilbert's formalist tenets.

In his original conception of the formalist program, Hilbert seems to have assumed implicitly the completeness of the axiomatic systems from which he began. The empirical evidence for this assumption was overwhelming. The axiomatic systems for number theory and analysis employed by the Hilbert school were more than adequate for formalizing all of the basic mathematics of the day, and more abstract topics such as functional analysis, the theory of transfinite cardinals and point set topology were all easily accommodated in the formal system of set theory created by Hilbert's colleague Ernst Zermelo. It was a shock, then, when Gödel showed that even elementary number theory is essentially incompletable.

Gödel's famous first incompleteness theorem can be stated as follows. Let T be a formal system of number theory so that all its theorems are true, and in which the predicate "s is a sequence of formulas constituting a proof of the formula A in the system T" is decidable, that is, there is an algorithm to decide for a given sequence s and formula A whether or not the relation holds. Then if T contains a certain minimum amount of

elementary number theory, it is incomplete, which is to say, there is a sentence G of T so that neither G nor its negation is a theorem.

Gödel's proof of the theorem (Gödel 1986: 144–95) involves the construction of a self-referential sentence akin to the Liar paradox. Gödel's basic insight was that by a system of encoding ("Gödel numbering"), formulas of the number-theoretical language could be considered as themselves being numbers (more precisely, Gödel's encoding produces an isomorphic image of the logical system in the natural numbers). In particular, we can express in the system T a number-theoretical relation $P(x, y)$ expressing the fact that x is the code number of a sequence of formulas that is a proof in the system T of the formula with code number y. Furthermore, since we have assumed that the proof predicate of T is decidable, the relation P is decidable in T, that is to say, if for particular numbers m, n, the relation $P(m, n)$ holds, then the sentence $P(\mathbf{m}, \mathbf{n})$ is provable in T, where '$\mathbf{m}$' is the numeral denoting the number m, and if it does not hold, then $\neg P(\mathbf{m}, \mathbf{n})$ is provable in T. Gödel can hence express the predicate 'x is the code number of a formula provable in T' as the existential formula $\mathrm{Prov}(x) \leftrightarrow \exists y P(y, x)$.

Gödel completes the proof of his first incompleteness theorem by making use of a clever diagonal construction to construct a self-referential sentence G that (interpreted via the coding devices) says of itself that it is not provable. More formally, we have as a theorem of T:

$$G \leftrightarrow \neg \mathrm{Prov}(\ulcorner G \urcorner),$$

where $\ulcorner G \urcorner$ stands for the code number of the sentence G itself. Gödel can now show that neither G nor its negation is a theorem of T by an argument resembling the reasoning leading to the contradiction in the Liar paradox. In this case, though, the paradoxical argument leads to incompleteness, not a contradiction.

We assumed in the above sketch of Gödel's argument that all of the theorems of T were true. However, an examination of the details of the proof shows that in demonstrating G itself to be unprovable, it is sufficient to assume that T is consistent. (A few years after Gödel's result appeared, Rosser using a more complicated self-referential sentence showed that the assumption of simple consistency was sufficient for incompleteness.) What is more, the argument showing this has a constructive, in fact finitary, character. It can therefore be formalized in T itself (since we assumed that T is adequate for elementary number theory). Thus the implication

$$\mathrm{Con}(T) \rightarrow \neg \mathrm{Prov}(\ulcorner G \urcorner)$$

is provable in T, where $\mathrm{Con}(T)$ is the statement formalizing the consistency of T. However, since the consequent of this implication is provably equivalent to G itself in T, it follows that if T itself is consistent, then $\mathrm{Con}(T)$ is unprovable in T.

This last result, known as Gödel's second incompleteness theorem, clearly has strong negative implications for Hilbert's consistency problem. If we start from a system of mathematics that encompasses the usual elementary forms of reasoning in number theory, then it presumably should include all of finitary reasoning (there is some uncertainty here, since Hilbert's notion of 'finitary reasoning' is not completely clear). But then if the system is consistent, it cannot prove its own consistency, and so a proof of

its consistency by finitary means is impossible. Hence it appears that Gödel's second incompleteness theorem precludes mathematics pulling itself up by its own bootstraps, as Hilbert had hoped.

Within the space of two years, Gödel had answered two of the fundamental problems of the Hilbert school. There remained the decision problem, though after the negative results of the incompleteness paper, it seemed most unlikely that this problem could have a positive solution. Gödel himself, in fact, came very close to providing a negative solution in the later sections of his incompleteness paper, a part that is rarely cited, since it was overshadowed by the later results of Church and Turing. However, these results are of considerable philosophical interest.

It was pointed out above that the completeness problem is not formulable in finitary terms, since it contains the non-constructive concept of semantic validity. However, it is possible to imagine constructive analogues of the completeness problem. To be more specific, let us imagine a formula of first-order predicate logic containing a certain number of predicate and relation symbols. We might then ask whether a constructive analogue of Gödel's completeness theorem holds in the sense that for any such formula, either it is provable by the usual axioms and rules for predicate logic, or, if it is not provable, we can find mathematical predicates (say, relations and predicates definable in number theory) so that when they are substituted for the atomic predicates in the formula, the resulting mathematical formula is refutable in some fixed axiom system for mathematics.

Gödel showed by analysing his unprovable formula G that for any consistent formal system S for mathematics, there are unprovable formulas of predicate logic that cannot be shown to be invalid by the substitution method in S. Looked at from the foundational point of view, this shows that the attempted constructive reformulation of the completeness problem fails. It also shows that it is highly unlikely that there could be an algorithm for the decision problem that could be proved to be correct in a standard system for mathematics.

The techniques that Gödel employed in the proof just described are essentially the same as those used a few years later by Church and Turing in showing the decision problem unsolvable. However, Gödel himself did not draw this conclusion. The difficulty lay in the fact that there was at that time no accepted precise definition delineating the class of mechanical procedures or algorithms. The creation of this definition was the next great step forward in logic, and is described in the next section.

4 Computability

Hilbert expected a positive solution to the decision problem, so that he was content to formulate the problem in terms of the intuitive mathematical notion of an algorithm. Gödel's incompleteness results, though, clearly pointed towards the conclusion that the problem was in fact unsolvable. To prove a negative result, however, it was essential to give a precise mathematical delineation of the concept of a mechanical procedure, or algorithm. This was first achieved by Alonzo Church and Alan Turing in 1936–37. Although Church was first in proposing a precise definition of computability (so that the identification of the intuitive with the mathematical concept is usually called

'Church's thesis'), Turing's conceptual analysis is usually held to be superior, and we shall follow Turing here. The reader is referred to an article by Wilfried Sieg for a penetrating account of the historical background to the work of Church and Turing (Sieg 1997).

Turing proceeded by giving a conceptual analysis of mechanical computation; the intended notion is that of a human carrying out the steps of an algorithm (recall that when Turing was writing, digital computers did not yet exist). His analysis can be explained in terms of two basic types of conditions, *boundedness conditions* and *locality conditions* (the terminology is that of Sieg). The computer (in the 1930s, when Turing was writing, this was always taken to denote a human being) operates in discrete time steps in a discrete symbol space – one can imagine a two-dimensional space, like a sheet of paper, or a one-dimensional space, like the paper tape of a Turing machine. The computer can perform the elementary actions of changing observed symbols and changing the set of observed symbols (moving in the symbol space). The boundedness conditions are these: the computer can recognize only finitely many distinct symbols, and has only a finite number of internal mental states (these are computational states, and need not be taken as mental states in a broader sense). The locality conditions are these: at each step, only finitely many symbols are observed, and in a single step, the computer can only move to a new symbol that is within a bounded distance of a previously observed symbol.

Turing adds to this model a deterministic condition: the elementary actions performed at each time step are uniquely determined by the current internal state, and the currently observed symbol configuration. To specify the functions computed by such a device, we need to add some conventions on input and output. With this, we have a complete model for mechanical computation.

Turing argued that a mechanical model for computation such as we have described in general terms above, is equivalent to the special case where the symbol space is one-dimensional, and at each step, exactly one symbol in this space is being observed. This is the well-known model of the *Turing machine*. The conceptual analysis sketched above is convincing evidence that this model is in fact a *universal* model for computation, in the sense that any mathematical function computed by an algorithm can be given in the form of a Turing machine.

Assuming this analysis of computation, we can now give a completely general definition of formal system, a concept that underlies Hilbert's conception of his program. A formal axiomatic system, then, is one in which there is a mechanical procedure to determine whether a string of symbols represents a meaningful assertion, and there is a set of axioms and rules that are also mechanically checkable (that is to say, there is an algorithm to determine whether or not a given string of symbols is or is not a proof in the system). With this definition, it is possible to state and prove a completely general version of Gödel's incompleteness theorem.

We can define the function $f_M(n)$ computed by a machine M as follows. We shall say that M *halts* if in the course of a computation it reaches a combination of internal state and input symbol for which it has no instruction. We shall suppose that the input and output of the machine consist of numerals in decimal notation. If M is given a number n as input, then if M eventually halts with the decimal notation for a number o written on the tape, then we say that o is the value of f_M for the input n. Notice that in general,

f_M is only a *partial* function, since there may be numerical inputs for which M goes into an infinite loop, for example, and never halts. We can define computable functions with two or more inputs in the same way.

It is now relatively easy to prove the undecidability of the decision problem. Every Turing machine can be specified by a finite list of instructions having a form such as: 'If you are in state 3, and are looking at the symbol 1, then change it to a 0, and go left one square.' These can be encoded as single numbers, using the technique of Gödel numbering, so that we can speak of the Turing machine M_k with code number k. We define the *halting problem* to be the problem of deciding, given two numbers k and n, whether the machine M_k eventually halts, when given input n.

We can prove the halting problem unsolvable by a straightforward diagonal argument. Let us suppose that the halting problem is solvable. Then there must be a Turing machine M that when given the pair of numbers k and n, outputs 1 if the machine M_k eventually halts, when given input n, otherwise 0. We can then use M to construct a new machine M' that when given the single input k, halts if M_k given k as input fails to halt, and otherwise fails to halt. (The details of the construction of M' from M are an exercise in Turing machine programming that we leave to the reader.) But now let h be the code number of the machine M', so that $M' = M_h$. Then on input h, M_h halts if and only if it does not halt, a contradiction.

Given the basic undecidability result we have just proved, we can now show the decision problem unsolvable. The proof is essentially a large scale exercise in formalizing assertions in first order logic. That is to say, given a machine M and input k, we can write down a formula $F(M, k)$ of first order logic that is valid if and only if M halts on input k. It follows that the decision problem must be unsolvable, since any algorithm solving it would lead to an algorithm solving the halting problem.

Our proof of unsolvability of the halting problem has another welcome corollary; another proof of Gödel's theorem. Let S be a standard formal system of number theory. We can formalize the encoding of Turing machines in S, so that we can, for example, write down a formula of S that expresses the fact that a number is a code number of a Turing machine. Using methods similar to those used in proving the decision problem unsolvable, we can find a formula $H(x, y)$ so that $\mathrm{H}(\mathbf{k}, \mathbf{n})$ is true if and only if the machine M_k halts on input n. Now we claim that S, if consistent, cannot prove all true statements of the form $H(\mathbf{k}, \mathbf{n})$ or $\neg H(\mathbf{k}, \mathbf{n})$. For suppose that it did; then we could solve the halting problem as follows. We can write a programme to print out one by one all the theorems of S (this is because we assumed that S is a formal system). Then to decide whether or not machine M_k halts on input n, we simply have to wait to see whether $H(\mathbf{k}, \mathbf{n})$ or $\neg H(\mathbf{k}, \mathbf{n})$ emerges as a theorem. This is impossible, so S is necessarily incomplete with respect to this class of statements. In fact, we can be a little more specific; there must be a particular machine M_k and input n so that M in fact does not halt on input n, but S cannot prove the statement $\neg H(\mathbf{k}, \mathbf{n})$.

A striking property of Turing's definition is that it is absolute, that is to say, it does not depend on the details of formalism used to define it. This aspect was stressed by Gödel in remarks at Princeton in 1946 commenting on an address by Tarski:

> Tarski has stressed in his lecture (and I think justly) the great importance of general recursiveness (or Turing's computability). It seems to me that this importance is largely due to

> the fact that with this concept one has for the first time succeeded in giving an absolute definition of an interesting epistemological notion, i.e., one not depending on the formalism chosen. (Gödel 1990: 150)

The evidence for Church's thesis is overwhelming, both in the sense that all known functions that are intuitively computable are computable in the sense of Turing, but also in the sense that many other proposed definitions (such as general recursiveness, computability by Markov algorithms, computability by register machines and so on) are equivalent to Turing's definition. One might object, of course, that (as was pointed out above) in the late 1930s the empirical evidence that the currently accepted formal systems for mathematics were complete was also overwhelming. In the case of Church's thesis, however, we have available Turing's conceptual analysis sketched above showing that any alternative concept of computability must drop the boundedness and locality conditions. A robust notion of quantum computation has recently emerged, that involves dropping (in a sense) the locality condition. This new notion does not lead to quantum computable functions that are not Turing computable, but it does open the door to possibly large gains in efficiency based on the exploitation of new features due to non-classical quantum effects.

5 Absolute and Relative in Logic

Hilbert's program was based on a belief that all mathematical concepts and constructions can be fully mirrored by formal, syntactical methods. Most of the results we have discussed above show that such mirroring is in fact impossible. For example, the concept of truth in number theory cannot be fully represented by provability in any formal system. In a similar way, we can show that many other mathematical concepts, such as the concept of a definable object, share the same essential incompletability with the notion of mathematical truth. In all of these cases, the incompleteness is a manifestation of the diagonal method. Any attempt to characterize the concept in a fixed formal framework leads by diagonalization to the construction of an object falling outside the formal characterization.

It may be asked whether one could not recover the absolute character of logical concepts by loosening the stringent finitistic character of Hilbert's requirements. A strategy of this sort was considered by Gödel in his 1948 Princeton lecture quoted above. In the case of definability, the argument of Richard's paradox of the least indefinable ordinal number, makes it clear that any absolute notion of definability must take all ordinal numbers as definable. Gödel's suggestion was to take definability in terms of ordinals as a possible definition of absolute definability. That is to say, a set is said to be ordinal definable if there is a sentence of the extended language of set theory in which all ordinals are primitive constants that uniquely defines it in the universe of all sets. This definition has the required property that it is impossible to apply the diagonal argument to find a set that is not definable; trivially, all ordinals are definable, so that the argument of Richard's paradox does not apply. On the other hand, the concept of definable object is obviously highly non-constructive, about as far from Hilbert's finitistic ideas as one can imagine.

It may seem surprising that an absolute concept, that of computability, emerged in the 1930s, a time when most of the concepts of logic, such as provability, were shown to have a relative, not absolute character. In fact, the absoluteness of this concept rests on the assumed absoluteness of another concept, namely the concept of truth for statements of number theory. This can be seen if we look at the definition of what it means for a Turing machine M to compute a function of natural numbers. This can be stated formally as: 'For every input n, there is a computation of M that terminates with a number o as output.' This can be encoded as a universal-existential sentence of elementary number theory. However, we cannot replace the notion of arithmetical truth here with a weaker notion such as the provability of the formalized version of the statement in an axiomatic system for number theory. By arguments similar to those used above in connection with the halting problem, we can show that no such system can prove all and only the true statements of this type. This is yet another manifestation of the incompleteness phenomenon.

Since we do seem to have a 'clear and distinct perception' of the notion of truth in number theory, it has often been argued that this demonstrates a clear superiority of humans over machines. More exactly, the incompleteness and undecidability results of Gödel, Church, and Turing have been held to show that humans have an absolute advantage over machines in that they are able to surpass any fixed machine in their insight into mathematical truths. The best known arguments for this conclusion are due to Lucas (1961) and Penrose (1989).

The Lucas/Penrose argument runs as follows. Let us suppose that we have programmed a computer to print out the theorems of a formal system of number theory one by one (the fact that we can program a computer to do this can be taken as an alternative definition of 'formal axiomatic system'). Gödel's incompleteness theorem applies to the formal system in question, so that there is for any such system a sentence G (the Gödel sentence for the system), that must be unprovable, provided the system is consistent. However, *we*, standing outside the formal system, and using our mathematical insight, can see that the sentence G is true, and so we can surpass the capacity of any fixed machine. This, according to Lucas and Penrose, proves that mechanical models of the mind are impossible, in short, that our minds cannot be machines.

The problem with the Lucas/Penrose argument presented above is that the key premise asserting that we can see the Gödel sentence to be true, remains undemonstrated. In fact, there are good reasons for thinking it to be false. The Gödel incompleteness theorem asserts a hypothetical proposition, namely that *if* the system in question is consistent, then the sentence G is unprovable. However, this hypothetical is provable in the system itself, under quite weak assumptions – in fact, this is the key idea of Gödel's second incompleteness theorem. For Lucas and Penrose to prove their case, they have to argue that we can see G itself to be true. This entails that we are able to show the system consistent.

There is no good reason to think that this last assumption is true. There are unsolved problems of mathematics (the Riemann hypothesis is perhaps the best known case) that have the property that if they are false, then this can be demonstrated by a simple counterexample. It follows from this that if we add such assumptions to a formal axiomatic system of mathematics, then the system is consistent if and only if the conjecture is true. This means that proving the consistency of a system based on, say, a version of

analytic number theory together with the Riemann hypothesis would be equivalent to proving the Riemann hypothesis. The Riemann hypothesis, though, is one of the most famous unsolved problems of mathematics, and it is unclear whether or not it will be solved in the near future. Lucas's and Penrose's assertion of an absolute superiority of minds over machines, then, seems to be without foundation.

Gödel himself tried to draw philosophical consequences from his incompleteness theorem, but was well aware that the simple argument of Lucas and Penrose was inadequate, since it rests on the unsupported assertion that human mathematicians can resolve all mathematical problems of a certain type. His most extended attempt at spelling out the philosophical implications of his theorem is to be found in his Gibbs lecture, delivered in 1951, but first published in the third volume of his collected works (Gödel 1995). Gödel's conclusion takes the form of a disjunction. If we make the assumption that humans can indeed resolve all consistency questions about formal systems of number theory, then an absolute superiority of humans over machines follows by the Lucas/Penrose argument. However, if this assumption is in fact false, then it follows that there must be mathematical assertions of a fairly simple type (since consistency assertions can be expressed, through the device of Gödel numbering as problems of number theory) that are absolutely unsolvable. In Gödel's own words:

> Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified.
>
> (Gödel 1995: 310)

Gödel's own philosophical argument is not open to the simple objection made above to the Lucas/Penrose argument. However, one might still object that it involves an unjustified idealization of the human capacity for proving theorems. In particular, Gödel presupposes that humanly provable mathematical propositions form a well-defined set. However, one could argue that the totality of humanly provable propositions is a very ill-defined collection, with vague boundaries, quite unlike the set of theorems of a formal system.

The philosophical consequences of the incompleteness theorems in the broad sense remain obscure and controversial. In the narrower sense, though, Gödel's results provide a fairly conclusive refutation of Hilbert's formalist program in the foundations of mathematics. This is a rare and very unusual instance of decisive progress in the foundations of mathematics and logic.

References

Ewald, W. B. (ed.) (1996) *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, vol. 2. Oxford: Clarendon Press.

Gödel, K. (1986) *Collected Works*, vol. 1: *Publications 1929–1936*, Solomon Feferman, John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay, and Jean van Heijenoort (eds.). Oxford: Oxford University Press.

Gödel, K. (1990) *Collected Works*, vol. 2: *Publications 1938–1974*, Solomon Feferman, John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay, and Jean van Heijenoort (eds.). Oxford: Oxford University Press.
Gödel, K. (1995) *Collected Works*, vol. 3: *Unpublished Essays and Lectures*, Solomon Feferman, John W. Dawson, Jr., Warren Goldfarb, Charles Parsons and Robert M. Solovay (eds.). Oxford: Oxford University Press.
Lucas, J. R. (1961) Minds, machines and Gödel. *Philosophy*, 36, 112–27.
Penrose, R. (1989) *The Emperor's New Mind*. Oxford: Oxford University Press.
Sieg, W. (1997) Step by recursive step: Church's analysis of effective calculability. *Bulletin of Symbolic Logic*, 3, 154–80.
Smullyan, R. (1968) *First-Order Logic*. Berlin: Springer (reprinted by Dover Publications 1995).
Van Heijenoort, J. (1967) *From Frege to Gödel*. Cambridge, MA: Harvard University Press.

Further Reading

Boolos, G. and Jeffrey, R. C. (1989) *Computability and Logic*, 3rd edn. Cambridge: Cambridge University Press.
Kleene, S. C. (1952) *Introduction to Metamathematics*. New York: Van Nostrand.
Reid, C. (1970) *Hilbert*. Berlin: Springer-Verlag.

21

Metatheory of Logics and the Characterization Problem

JAN WOLEŃSKI

1 Introduction

The word 'metatheory' denotes or perhaps suggests a theory of theories. Metascientific studies in the twentieth century used the term 'metatheory' to refer to investigations of theories in a variety of disciplines, for example, logic, sociology, psychology, history, etc. However, the philosophers of the Vienna Circle who made metatheoretical studies of science the main concern of their philosophy restricted metatheory to the logic of science modeled on developments in the foundations of mathematics. More specifically, the logic of science was intended to play a role similar to metamathematics in Hilbert's sense; that is, it was projected as formal analysis of scientific theories understood as well-defined linguistic items. The word 'metamathematics' was used before Hilbert, but with a different meaning from his (see Ritter et al. 1980: 1175–8). In the early nineteenth century, mathematicians, like Gauss, spoke about metamathematics in an explicitly pejorative sense. It was for them a speculative way of looking at mathematics – a sort of metaphysics of mathematics. A negative attitude to metaphysics was at that time inherited from Kant and early positivists. The only one serious use of 'metamathematics' was restricted to metageometry, and that was due to the fact that the invention of different geometries in the nineteenth century stimulated comparative studies. For example, investigations were undertaken of particular axiomatizations, their mutual relations, models of various geometrical systems, and attempts to prove their consistency. The prefix 'meta' presently suggests two things. First, it indicates that metatheoretical considerations appear 'after' (in the genetic sense) theories are formulated. Secondly, the prefix 'meta' suggests that every metatheory is 'above' a theory which is the subject of its investigations. It is important to see that 'above' does not function as an evaluation but only indicates the fact that metatheories operate on another level than theories do. A simple mark of this fact consists in the fact that theories are formulated in an object language, and metatheories are expressed in a related metalanguage.

It is probably not accidental that Hilbert passed to metamathematics through his famous study of geometry and its foundations. Hilbert projected metamathematics as a rigorous study of mathematical theories by mathematical methods. Moreover, the Hilbertian metamathemics, due to his views in the philosophy of mathematics

(formalism) was restricted to finitary methods. If we reject this limitation, metamathematics can be described as the study of mathematical systems by mathematical methods; they cover those that are admitted in ordinary mathematics, including infinitistic or infinitary – for instance, the axiom of choice or transfinite induction. However, this description is still too narrow. Hilbert's position in metamathematics can be described as follows: only syntactic combinatorial methods are admissible in metatheoretical studies. However, the semantics of mathematical systems is another branch of the methatheory of mathematics. It is interesting that the borderline between syntax and semantics corresponds to some extent with the division between finitary and infinitary methods. I say 'to some extent' because we have also systems with infinitely long formulas (infinitary logic). It is clear that the syntax of infinitary logics must be investigated by methods going beyond finitary tools. It was also not accidental that systematic formal semantics (model theory) which requires infinitistic methods appeared in works by Alfred Tarski who, due to the scientific ideology of the Polish mathematical school, was not restricted to the dogma that only finite combinatorial methods are admissible in metamathematics. Today, metamathematics can be divided into three wide areas: proof theory (roughly speaking, it corresponds to metamathematics in Hilbert's sense if proof-methods are restricted to finitary tools, or it is an extension of Hilbert's position if the above-mentioned restriction is ignored), recursion theory (which is closely related to the decision problem, that is, the problem of the existence of combinatorial procedure providing a method of deciding whether a given formula is or is not a theorem) and model theory, that is, studies of relations between formal systems and structures which are their realizations; model theory has many affinities with universal algebra.

The metatheory of logics (plural is proper, because we have many competing logical systems) is understood here as a part of metamathematics restricted to logical systems. We can also use the word 'metalogic' and say that it refers to studies of logical systems by mathematical methods. This word also appeared in the nineteenth century (see Ritter et al. 1980: 1172–4), although its roots go back to the Middle Ages (*Metalogicus* of John of Salisbury). Philosophers, mainly neo-Kantians, understood metalogic to be concerned with general considerations about logic. The term 'metalogic' in its modern sense was used for the first time in Poland (by Jan Łukasiewicz and Alfred Tarski) as a label for the metamathematics of the propositional calculus. Thus, metalogic is metamathematics restricted to logic, and it covers proof theory, investigations concerning the decidability problem, and model theory with respect to logic.

When we say that metalogic is a part of metamathematics, it can suggest that the borderline between logic and mathematics can be sharply outlined. However, questions like 'What is logic?' or 'What is the scope of logic?' have no uniformly determined answer. We can distinguish at least three relevant subproblems that throw light on debates about the nature of logic and its scope. The first issue focuses on the so called first-order thesis. According to this standpoint, logic should be restricted to standard first-order logic. The opposite view contends that the scope of logic should be extended to a variety of other systems, including, for instance, higher-order logic or infinitary logic. The second issue focuses on the question of rivalry between various logics. The typical way of discussing the issue consists in the following question: Can we or should we replace classical logic by some other system, for instance, intuitionistic, many-

valued, relevant or paraconsistent logic? This way of stating the problem distinguishes classical logic as the system which serves as the point of reference. Thus, alternative or rival logics are identified as non-classical. There are two reasons to regard classical logics as having a special status. One reason is that classical logic appeared as the first stage in the development of logic; it is a historical and purely descriptive circumstance. The second motive is clearly evaluative in its character and consists in saying that classical logic has the most 'elegant' properties or that its service for science, in particular, for mathematics, is 'the best.' For example, it is said that abandoning the principle of excluded middle (intuitionistic logic), introducing more than two logical values (many-valued logic), changing the meaning of implication (relevant logic) or tolerating inconsistencies (paraconsistent logic) is something wrong. It is also said that some non-classical logics, for example, intuitionistic or many-valued logics, considerably restrict the applicability of logic to mathematics. It is perhaps most dramatic in the case of intuitionistic logic, because it or other constructivistic logics lead to eliminating a considerable part of classical mathematics. Thus, this argument says that only classical (bivalent or two-valued) logic adequately displays the proof methods of ordinary mathematics. While the discussion is conducted in descriptive language, it appeals to intuitions and evaluations of what is good or wrong in mathematics. The situation is similar as far as the matter concerns metalogical properties of particular systems such as completeness, decidability or the like, because it is not always obvious what it means to say that a logic possesses them 'more elegantly' than a rival system. The priority of classical logic is sometimes explained by pointing out that some properties of non-classical logic are provable only classically. This is particularly well-illustrated by the case of the completeness of intuitionistic logic: Is the completeness theorem for this logic intuitionistically provable? The answer is not clear, because the stock of intuitionistically or constructively admissible methods is not univocally determined, and they vary from one author to another. Finally, our main problem (what is logic and what is its scope?) is also connected with extensions of logics. If we construct modal logics, deontic logics, epistemic logics, etc., we usually start with some basic (propositional or predicate) logic. We have modal propositional or predicate systems which are based on classical, intuitionistic, many-valued or some other basic logic. Does any given extension (roughly speaking, an extension of a logic arises when we add new concepts, for example necessity, to old ones in such a way that all theorems of the system before extension are theorems the new system) of a chosen basic logic preserve its classification as a genuine logic or does it produce an extralogical theory? The *a priori* answer is not clear, even when we decide that this or that basic system is *the* logic. The problem of the status of extensions of logic is particularly important for philosophical logic because it consists mainly of systems of this sort.

The three issues concerning the question 'What is logic?' are mutually interconnected. The choice between first-order logic or higher-order logic automatically leads to the two other issues, because it equally arises with respect to any alternative logic and any extension of a preferred basic logic. Thus, we have a fairly complex situation. Yet the above division into three issues does not exhaust all problems. Usually it is assumed that first-order logic (classical or not) is based on the assumption that its universe it not empty. However, as Bertrand Russell once remarked, that it is a defect of logical purity, if one can infer from the picture of logic that something exists. This is

perhaps the main motivation for so-called free logic, that is, logic without existential assumptions (logic admitting empty domains). Is it classical or not? The described situation suggests a pessimism as far as the matter concerns a natural and purely descriptive characterization of logic; it seems that an element of a convention is unavoidable here. A further reason that the domain of metalogic cannot be sharply delimited is that several metalogical or metamathematical results distinguish logical (even in a wider sense) from other formal systems. Assume that we decide to stay with the first-order thesis. The second Gödel theorem (the unprovability of the consistency of elementary arithmetic) clearly separates pure quantification logic from formal number theory. It is one reason that metamathematical results are of interest for metalogic. Metalogical investigations also use several concepts that are defined in general metamathematics, for example formal system, axiomatizability, consistency, completeness, provability, etc. Fortunately, we are not forced to answer the borderline question in a final manner. My aim in this essay is to review the most essential metalogical concepts. Classical first-order logic is taken as the paradigm. The treatment is rather elementary. Although I assume some familiarity with syntax and semantics of first-order logic as well as with several simple concepts of set theory, most employed concepts are explained. However, some important concepts of metalogic, for instance that of recursive function, do not allow a brief and elementary clarification. On the other hand, it would be difficult and not reasonable to resign from them. These concepts are marked by * and the reader is asked to consult textbooks listed in the references. In particular, I recommend Hunter (1971) and Grzegorczyk (1974); moreover, Pogorzelski (1994) is suggested as the fullest survey of metalogic (I follow this book in many matters). Special attention will be given to relations between syntactic and semantic concepts that are most strikingly displayed by (semantic) completeness theorems. I do not enter into historical details, although it seem to be proper to include dates when some fundamental theorems were proved (references to original papers are easily to be found in works listed in the Bibliography).

The characterization problem is a special metalogical issue. It consists in giving sufficient and necessary conditions which determine particular logics or classes of logics. These conditions can be syntactic, semantic, or mixed. Let me explain the problem in the case of the propositional calculus. It has been axiomatized in various ways. However, one axiomatic base, rather long, is particularly convenient here. The axioms are these (I use the Hilbert-style formalization use of axiom-schemata. Thus, the letters A, B, C are metalinguistic variables referring to arbitrary formulas of the propositional calculus and modus ponens (B is derivable from A and $A \to B$) is the only inference rule:

(A1) $A \to (B \to A)$

(A2) $(A \to (A \to B)) \to (A \to B)$

(A3) $(A \to B) \to ((B \to C) \to (A \to C))$

(A4) $A \wedge B \to B$

(A5) $A \wedge B \to B$

(A6) $(A \to B) \to ((A \to C) \to (A \to B \wedge C))$

(A7) $A \to A \vee B$

(A8) $B \to A \vee B$

(A9) $(A \to C) \to ((B \to C) \to (A \vee B \to C))$

(A10) $(A \leftrightarrow B) \to (A \to B)$

(A11) $(A \leftrightarrow B) \to (B \to A)$

(A12) $(A \to B) \to ((B \to A) \to (A \leftrightarrow B))$

(A13) $(A \to B) \to (\neg B \to \neg A)$

(A14) $A \to \neg\neg A$

(A15) $\neg\neg A \to A$

A nice feature of this set of axioms is that we can easily distinguish subsets related to particular connectives. (A1)–(A3) characterize implication, (A4)–(A6) conjunction, (A7)–(A9) disjunction, (A10)–(A12) equivalence, and (A13)–(A15) negation. Now if we eliminate (A15), we obtain the axiom set for intuitionistic logic. Thus, we can say that (A1)–(A15) solve the characterization problem for classical propositional logic, but (A1)–(A14) do the same job for intuitionistic propositional logic, provided that the characterization problem is to be solved by axiomatic methods. Other ways of characterizing logical systems proceed by matrices (truth-tables), semantic tableaux*, trees*, semantic games* or Hintikka sets*, but all provide conditions which separate various more or less alternative logics. One characterization result recently became particularly famous. It is the celebrated Lindström theorem which establishes very general conditions for first-order logic. This theorem and will be presented in a separate section below.

Formal metalogical results are interesting in themselves as well as being philosophically important. The problem of the nature of logic has a decisively philosophical character. Several accepted intuitions about logic have gained widespread acceptance: that logic is formal, universal or topic-neutral, and provides sound (leading always from truths to other truths) rules of inference. It is interesting to look at metalogical results as capturing old intuitions; for example, that expressed in the following words of Petrus Hispanus: *dialectica est art artium et scientia scientiarum ad omnium scientiarum methodorum principia viam habent* (dialectics (that is, logic) is the art of arts and the science of sciences that provides methodological principles for all sciences). Another illustration of the philosophical importance of formal results is that, according to intuitionism, intuitionistic and classical logic are simply incomparable. It is sometimes maintained that differences between alternative logics consist in the assignment of different meanings to logical constants and facts, and that these systems are not intertranslatable. As the characterization problem displayed by axioms for the propositional calculus shows, however, at least from the classical point of view, classical logic and all weaker systems are perfectly comparable.

2 Logic via Consequence Operation and Semantics

Intuitively speaking, logic provides manuals for proving some propositions on the basis of some assumptions. These manuals consist of inference rules; for example, *modus ponens* instructs us that we may logically pass from A and $A \to B$ as premises to B as

conclusion. Assume that **R** is a set of inference rules. The notation $X \vdash^{R} A$ expresses the fact that a formula A is provable (derivable) from the set X of assumptions, relative to rules of inference from **R** (I will omit the superscript indexing the provability sign in due course). We define

(DCn) $A \in Cn(X) \Leftrightarrow X \vdash A$.

Although Cn (the consequence operation) and $\vdash$ (the consequence operator) are mutually interdefinable, there is a categorial difference between them. Let **L** be a language understood as a set of formulas. Cn is a mapping from 2^{L} to 2^{L} that transforms sets of formulas into sets of formulas, and the consequence operator maps 2^{L} into **L**, that is, sets of formulas are transformed into single formulas.

The analysis of logic *via* the consequence operator is much more common than that using Cn (see Segerberg (1982) for the first approach). It is also more closely related to codifications of logic *via* natural deduction techniques or sequents which are also used (see Hacking 1979) in analyzing the concept of logic. I will take another route, however, and concentrate on the consequence operation (I follow Surma (1981); see also Surma (1994)). The first question that arises here is this: how many consequence operations do have we? The answer is that there are infinitely many Cn's. Thus, we need to establish some constraints selecting a 'reasonable' consequence operation (or operations). Tarski characterized the classical axiomatically Cn (in fact, Tarski axiomatized the consequence operation associated with the propositional calculus; the axioms given below concern the consequence operation suitable for first order logic). The axioms are these (explanations of symbols: $\varnothing$, the empty set; **L**, language; $\mathcal{N}_0$, the cardinality of the set of natural numbers; $\subseteq$, inclusion between sets; $\in$, the membership relation (being an element of a set); **FIN**, the class of all finite sets; $\cup$, union of sets; $\{A\}$, the set consisting of A as the sole element; $\cap$, product of sets; /, the operation of substitution for terms):

(C1) $\varnothing \leq \mathbf{L} \geq \mathcal{N}_0$

(C2) $X \subseteq CnX$

(C3) $X \subseteq Y \Rightarrow CnX \subseteq CnY$

(C4) $CnCnX = CnX$

(C5) $A \in CnX \Rightarrow \exists Y \subseteq X \wedge Y \in \mathbf{FIN}(A \in CnY)$

(C6) $B \in Cn(X \cup \{A\}) \Rightarrow (A \rightarrow B) \in CnX$

(C7) $(A \rightarrow B) \in CnX \Rightarrow B \in Cn(X \cup \{A\})$

(C8) $Cn\{A, \neg A\} = \mathbf{L}$

(C9) $Cn\{A\} \cap Cn\{\neg A\} = \varnothing$

(C10) $A(v/t) \in Cn\{\forall v A(v)\}$, if the term t is substitutable for v.

(C11) $A \in CnX \Rightarrow \forall v A(v) \in CnX$, if v is not free* in X, for every $B \in X$.

We can divide the axioms (C1–C11) into three groups. The first group includes (C1–C5) as general axioms for Cn. (C1) says that the cardinality of **L** is at most denumerably

(denumerably – finitely or so many as natural numbers) infinite, (C2) that any set is a subset of the set of its consequences, (C3) established the monotonicity of Cn (in general, a function f is monotonic if and only if $x \leq y$ entails $fx \leq fy$; in fact, inclusion is a kind of the $\leq$-operation), (C4) its idempotency (a function f is idempotent if and only if $ffx = fx$), (C5) states the finiteness condition which means that if something belongs to $Cn(X)$, it may be derived from a finite subset of X. In other words: every inference is finitary, that is, performable on the base of a finite set of premises and, according to the character of rules, finitely long. It is an important property, because there are also infinitary logical rules, for example the ω-rule which leads (roughly speaking) from the infinite sequence of premises $P(1)$, $P(2)$, $P(3)$, ... to the conclusion $\forall n P(n)$, but it is commonly recognized that human beings cannot effectively use such rules. (C1–C5) do not provide any logic in its usual sense. The logical machinery is encapsulated by the rest of axioms (related to logic based on negation, implication, and the universal quantifier): (C6) is modus ponens (it shows that modus ponens is the inverse of the deduction theorem), (C7) the deduction theorem (if B is derivable from the set X plus A, then the implication $A \rightarrow B$ is derivable from X; if it is to be applied to predicate logic, we must assume that A and B are closed formulas, that is formulas without free variables), (C8)–(C9) characterize negation, and (C10–C11) characterize the universal quantifier. We can also add axioms suitable for identity or introduce the consequence operation for intuitionistic logic.

Logic (more precisely: classical first-order logic) can be defined as $Cn\varnothing$. More formally we have:

(DL1) $A \in \mathbf{LOG} \Leftrightarrow A \in Cn\varnothing$, or, equivalently $\mathbf{LOG} = Cn\varnothing$.

Of course, modifications of Cn in accord with the ideas of alternative logics lead to their related definitions. For example, intuitionistic logic is given by the equality $\mathbf{LOG_i} = Cn_i\varnothing$. (DL1) looks artificial at first sight, because it is clear that the logical content is related to axioms imposed on Cn; clearly, the empty set here is a convenient metaphor: we can derive something from the empty set only because of the logical machinery already built into Cn. Hence, we have the problem of deciding what stipulations about the consequence operation are proper for logic. This question concerns general as well as special axioms. Worries concerning which logic, classical or some alternative, is the 'logic' also remain on this approach; for example, we can consider this question with respect to modal extensions or formal systems which contain rules related to axioms of arithmetic. Are $Cn_m\varnothing$ (the set of modal consequences, relatively to a system of modal logic, of the empty set) or $Cn_{ar}\varnothing$ (the set of arithmetical consequences of the empty set) logics)? As far as the general axioms are concerned, we can, for instance, drop the requirement of monotonicity (it leads to non-monotonic logics used in computer science) or finiteness (infinitary logic). Hence, any definition of logic *via* the consequence operation needs an additional justification. I will present a motivation for classical logic which can be easily applied to other systems.

First of all, let us observe that (DL1) is equivalent to two other statements, namely (an explanation concerning (DL3): an operation o closes the set X if and only if $oX \subseteq X$, that is, applications of o to X do not produce elements which do not belong to X):

(DL2) $A \in$ **LOG** if and only if $\neg A$ is inconsistent.

(DL3) **LOG** is the only non-empty product of all deductive systems (theories), that is, sets which satisfy the condition: $Cn X \subseteq X$ (are closed under Cn).

Now, (DL2) and (DL3) surely define properties which we expected to be possessed by any logic. We agree that negations of logical principles are inconsistencies and that logic is the common part of all, even mutually, inconsistent theories. Additionally, (DL3) entails that logical laws are derivable from arbitrary premises. Thus, we have the equivalence: $A \in Cn\emptyset$ if and only if $A \in Cn X$, for any X, and the equality **LOG** = $Cn\emptyset = Cn X$, for any X. These considerations show that (DL1) and its equivalents express an important intuition, namely that logic is universal in the sense that it does not require any premises, or is deducible from arbitrary assumptions.

Yet one might argue that such a construction of logic is circular because it defines logic by means of the prior assumption that something is logical. This objection can be easily met by pointing out that our definitions are inductive, that is, selects logical axioms as so called initial conditions and then shows how inductive conditions (in fact, the rules of inference coded by Cn) lead step by step to new logical elements. On the other hand, it is perhaps important for philosophical reasons to look at an independent characterization of logic. This is provided by semantics and it is expressed by (a model of a set X of sentences is a structure consisting of a universe of objects and a collection of relations defined on the universe such that all sentences belonging to X are true; if we admit open formulas, that is, formulas with free variables, a model of a set X of formulas is a structure in which all formulas belonging to X are satisfied):

(DL4) $A \in$ **LOG** if and only if for every model **M**., A is true in **M**.

This last definition describes logic as universal in the sense that logical laws are true in every model (domain). It is related to the old intuition that logic is topic neutral, that is, true or valid with respect to any particular subject matter. Intuitively, there is an obvious link between (DL1)–(DL3) and (DL4). However, we have no formal tools that prove all these definitions are equivalent. Since (DL1)–(DL3) are syntactical descriptions of logic (they use the concepts of consequence operation or consistency which are just syntactic), but (DL4) is semantic in its essence (it defines logic *via* the concept of a model), any comparison of the two approaches requires a rigorous investigation of how syntax and semantics are related. In fact, it consists in a comparison of the set of theorems (the set of provable formulas) of a system under investigation with the set of its validities (truths, tautologies).

3 Metalogic, Syntax and Semantics

Although we basically intend to achieve a precise comparison of syntax and semantics in logic, this section provides an opportunity to introduce several important metalogical concepts and properties (others will be defined in the next section). Let **S** be an arbitrary formal system formulated in a language **L**. The most important metalogical concepts are summarized by the following list:

- **S** is consistent if and only if $Cn\mathbf{S} \neq \mathbf{L}$; if **S** contains the negation sign this definition is equivalent to the more standard: **S** is consistent if and only if no inconsistent pair (that is, consisting of A and $\neg A$) of formulas belongs to the consequences of **S**.
- **S** is post-complete (the name honors of Emil Post, an American logician who defined the property in question) if and only if $Cn(\mathbf{S} \cup \{A\}) = \mathbf{L}$, for any formula A which is not a theorem of **S**.
- **S** is syntactically complete if and only if for any A, either $A \in Cn\mathbf{S}$ or $\neg A \in Cn\mathbf{S}$.
- **S** is semantically complete if and only if every provable formula of **S** is true in every model of **S** and every validity (truth) of **S** is provable in it.
- **S** is decidable if and only if the set of its theorems is recursive*.
- **S** is axiomatizable if and only if there is a set $\mathbf{Ax} \subseteq \mathbf{S}$ such that $\mathbf{S} = Cn\mathbf{Ax}$; if **Ax** is finite (recursive) we say that **S** is finitely (recursively) axiomatizable.

Some comments are in order. Various labels for particular properties are employed by various authors. For example, syntactical completeness is sometimes called negation-completeness. Semantic completeness has in fact two ingredients. The direction from provability to validity (every truth is provable) is considered as soundness (correctness, adequacy) and semantic completeness proper, so to speak, is expressed by the reverse implication (every truth is provable). The given definition of decidability is related to the Church thesis*, that is, the proposal to identify intuitively calculable functions (calculable in the finite mechanically performable steps) with recursive functions. Finally, the definition of axiomatizability does not exclude the situation that **S** forms its own axiomatic base.

We are mainly interested in properties of logic. The propositional calculus is consistent, post-complete, syntactically incomplete (it is enough to consider a single variable; neither p nor $\neg p$ are theorems of propositional logic), sematically complete, decidable (by the truth-table method) and finitely axiomatizable (by concrete formulas) or recursively axiomatizable (by schemata). One qualification is needed with respect to the concept of post-completeness. This property holds for the propositional calculus with axioms as concrete formulas and the rule of substitution. Fortunately, we can define another property, parallel to post-completeness which is possessed by the propositional calculus when it is formalized by axiom-schemata. Now, first-order predicate logic is consistent, not post-complete (if we add, for example, the sentence 'there are exactly two objects' which is not a logical theorem as a new axiom, the resulting system is not inconsistent), not syntactically complete, semantically complete (proved by Kurt Gödel in 1929), undecidable (proved by Alonzo Church in 1936), and finitely or recursively axiomatizable. All these facts apply to first-order predicate logic with identity. Gödel proved in 1931 two famous theorems (both of which assume that arithmetic is consistent): (1) every formal system strong enough for the elementary arithmetic of natural numbers is syntactically incomplete; (2) the consistency of arithmetic is unprovable in arithmetic; both theorems assume that arithmetic is consistent. The first theorem implies that arithmetic is not recursively axiomatizable. Tarski showed in 1933 that the set of arithmetical truths is not definable arithmetically*. Finally, Church proved in 1936 that arithmetic is undecidable. These four theorems are usually called limitative theorems, because they point out limitations inherent to any formalism sufficiently rich to cover the arithmetic of natural numbers.

For our aims, semantic completeness is the most important. In its most general form, the completeness theorem (in its strong form) says (the symbol $\models$ stands for validity):

(CT) **S** is semantically complete if and only if: $\mathbf{S} \vdash A \Leftrightarrow \mathbf{S} \models A$.

(CT) is equivalent to the Gödel–Malcev theorem:

(GM) **S** is consistent if and only if it has a model.

The proof of (GM) requires the axiom of choice* (or its equivalents) which means that it is not a constructive theorem. The most popular proof of (CT) uses the Lindenbaum lemma: every consistent set of sentences has a maximal consistent extension (maximality means here that adding any sentence to a maximally coinsistent set leads to inconsistency); this lemma is also not constructive. If we put $\emptyset$ instead **S** in (CT), we obtain $\emptyset \vdash A$ if and only if $\emptyset \models A$. By (DCn), it gives the weak completeness theorem

(CT1) $A \in Cn\emptyset \Leftrightarrow \emptyset \models A$.

Since the right part of (CT1) expresses the fact that A is true in all models, it legitimizes the equivalence of (DL1) and (DL4) for first-order predicate logic with identity. It should be clearly noted that the completeness theorem, although it establishes the parity of syntax and semantics in semantically complete systems, it does not provide in itself any definition of logic. However, if we agree that universality is its characteristic property, (CT1) shows that universality in the syntactic sense (provability from the empty set of premises) is exactly equivalent to universality in the semantic sense (truth in all models or logical validity). Moreover, this part of (CT) (or (CT1)) which expresses the soundness property (if a formula is provable, it is also true) justifies the intuition that logical rules are infallible: they never lead from truths to falsehoods.

The universality property is also displayed by another theorem, the neutrality theorem, which asserts that first-order predicate logic with identity does not distinguish any extralogical concept, that is, any individual constant or predicate parameter (c_i, c_j are individual constants, P_k, P_n are predicate parameters, the notation $A(c)$ and $A(P)$ means that a constant c (predicate parameter P) occurs in A):

(N) (a) $A(c_i) \in \mathbf{LOG} \Rightarrow A(_{c_j}/_{c_i}) \in \mathbf{LOG}$; (b) $A(P_k) \in \mathbf{LOG} \Rightarrow A(P_n/P_k) \in \mathbf{LOG}$.

This theorem says that if something can be provable in logic about an object or its property, the same can be also proved about any other object or property. It is of course another aspect of the topic-neutrality of logic.

4 The Characterization Problem for First-order Logic

The strong completeness theorem motivates a stronger understanding of logic. Let **T** be an extralogical theory (axiomatized by the axioms belonging to the set **Ax**). Thus **T** is the ordered triple (see Rasiowa and Sikorski n.d.: 187).

$\langle \mathbf{L}, Cn, \mathbf{Ax} \rangle$.

Now the consequence operation Cn operating on $\mathbf{L}$ and $\mathbf{Ax}$ generates the logic of $\mathbf{T}$. Denote logic in this extended sense by $\mathbf{LOG_T}$ and logic given by (DL1) by $\mathbf{LOG}_\varnothing$ (it operates on the empty set). Of course, $\mathbf{LOG}_\varnothing \subseteq \mathbf{LOG_T}$. The modification is not essential for logic in this sense: the stock of logical rules, given by $Cn\varnothing$, is the same. However, this extended concept of logic, which focuses on its applications, leads to a more general formulation of the characterization problem.

Call a logic regular if its logical symbols obey classical (Boolean) principles (it is practically restricted to negation; roughly speaking, 'Boolean' means that our logic is perfectly two-valued); we also assume that $\mathbf{L}$ is countable, that is, contains at most denumerably many sentences. A logic satisfies the compactness property (Com) if and only if it has a model if its every finite subset has a model. A logic satisfies the Löwenheim–Skolem property (LS) if and only if it has a countable model if it has an infinite model. The Lindström theorem (proved by Per Lindström in 1969) is the statement:

(L) First-order predicate logic is the strongest logic which satisfies (Com) or (CT), and (LS).

For example, second-order logic (first-order logic has quantifiers ranging over individuals; second-order logic also admits quantification over properties – the sentence 'for any object x, there is a property P such that x has P' is an example of a second-order sentence) satisfies neither (Com) nor (LS), but (CT) holds for it, if we admit second-order quantification over special entities, and logic with the quantifier 'there are uncountably many' is complete, but then it does not obey (LS). Of course, (L) holds also for logic defined by (DL1), that is, for $Cn\varnothing$. Let me add that no counterparts of (L) are known with respect to non-classical logics, in particular, intutionistic or many-valued logics. The reason is that they are not regular.

There is a considerable debate concerning the interpretation and consequences of (L) (see Barwise 1985). All parties agree that (L) asserts the limitations on the expressive power of first-order predicate logic. In particular, several mathematical concepts, like finiteness, cannot be defined in its language. However, it is a matter of controversy whether (L) determines that only first-order predicate logic deserves to be counted as *the* logic. The first-order thesis, previously explained, restricts the scope of logic to first-order logic, but the opposite standpoint maintains that if logic is to serve mathematics, its expressive power must be much greater than that of first-order languages. It is now clear why this problem becomes central when an extended concept of logic is assumed. Since definability is traditionally regarded as a logical issue, its limitations are perceived as limitations of the power of logic. I will come back to these questions in the next (and final) section.

5 Final Remarks

In this section, I come back to philosophical issues concerning the concept of logic. Let me start with the first-order thesis. Its opponents argue that it restricts the application

of logic in science, in particular, in mathematics, which requires that logic should have a considerable expressive or defining power in order to capture various mathematical concepts. On the other hand, the first-order thesis focuses on the universality property of logic and the infallibility of its inferential machinery (see Woleński 1999). Thus, we have to do here with a conflict between two different expectations concerning logic. The postulate that logic should have great expressive power recalls the ambitious projects of *logica magna* or *lingua characteristica* proposed by Leibniz, Frege, or Russell and intended as languages which are able to cover the whole of science or at least mathematics. The first-order thesis motivated by (L) and (N) sees logic as providing universally valid theorems, being the common part of all deductive systems, always generating a perfectly sound inference machinery. The issue is serious because either we can have strict universality or languages with a great expressive power, but not both virtues together. We can assume that $Cn\emptyset$ always provides secure rules of inference. Thus, the point is what should be regarded as logical: only propositional connectives, quantifiers, and identity, or perhaps also other concepts, like finiteness. It is not surprising that (CT) contributes to our understanding of the universality of logic. However, it was not expected that (Com) and (LS) do too, though if first-order predicate logic does not distinguish any extralogical concepts, it also should be neutral with respect to the cardinality of models, that is, the number of elements in their universes. It is interesting that there are also problems when we consider identity as a logical concept. The argument for its status as a logical constant stems from the fact that first-order logic with identity relation satisfies (CT), (N), and (L). On the other hand, identity enables us to define numerical quantifiers, for example, 'there are exactly two objects', but there are doubts whether such phrases deserve to be called logical. Thus we have reasons to say that the prospects for an answer to the question 'What is logic?' that is unconditional and free of at least some degree or arbitrariness, are not encouraging. The problem becomes still more complicated when non-classical logics are taken into account.

New problems arise when extensions of a basic logic are analyzed. It may be demonstrated by modal logic. Since modal systems are more closely treated in a separate chapter in this *Companion*, I limit myself to a very sketchy remarks. We can and even should ask whether $\Box$ (necessity) and $\Diamond$ (possibility) are logical constants? One might argue that since special conditions, related to particular modal systems, are imposed on modal models, especially on so-called accessibility relations (for example, deontic logic requires that this relation is irreflexive, the system **T** is associated with the condition of symmetry, etc.), modal logics are not universal. On the other hand, the system **K** does not require any particular constraint. Yet we can say that its characteristic formula $\Box(A \to B) \to (\Box A \to \Box B)$ is a modal translation of the theorem of first-order logic $\forall x(A \to B) \to (\forall xA \to \forall xB)$. However, **K** is a very weak system and does not display all traditional intuitions concerning logical relations between modalities. Thus, we perhaps should decide: either universality (no special provisos on modal models) or more content, like in the case of the controversy over the first-order thesis.

How then does metalogic contribute to our understanding of logic? The answer seems to be this. Although metalogical theorems do not provide answers which are free of conventional elements, they precisely show those points where intuitions go beyond formal results.

References

Barwise, J. (1985) Model-theoretic logics: background and aims. In J. Barwise and S. Feferman (eds.), *Model-Theoretic Logics* (pp. 3–23). Berlin: Springer-Verlag.

Hacking, I. (1979) What is logic? *Journal of Philosophy*, 76, 285–319.

Pogorzelski, W. A. (1994) *Notions and Theorems of Elementary Formal Logic*. Białystok: Warsaw University – Białystok Branch.

Rasiowa, H. and Sikorski, R. (n.d.) *The Mathematics of Metamathematics*. Warszawa: PWN – Polish Scientific Publishers.

Ritter, J. et al. (1980) *Historisches Wörterbuch der Philosophie*, vol. 5. Basel: Benno Schwabe.

Segerberg, K. (1982) *Classical Propositional Operators*. Oxford: Clarendon Press.

Surma, S. J. (1981) The growth of logic out of the foundational research in mathematics. In E. Agazzi (ed.), *Modern Logic – A Survey. Historical Philosophical, and Mathematical Aspects of Modern Logic and its Applications* (pp. 15–33). Dordrecht: Reidel.

Surma, S. J. (1994) Alternatives to the consequence-theoretic approach to metalogic. *Notas de Logica Matematica*, 39, 1–30.

Woleński, J. (1999) Logic from a metalogical point of view. In E. Orłowska (ed.), *Logic at Work: Essays Dedicated to the Memory of Helena Rasiowa* (25–35). Berlin: Physica-Verlag.

Further Reading

Cleave, J. P. (1991) *A Study of logics*. Oxford: Clarendon Press.

Ebinnghaus, H.-D. (1985) Extended logic: the general framework. In J. Barwise and S. Feferman, *Model-Theoretic Logics* (pp. 25–76). Berlin: Springer-Verlag.

Flum, J. (1985) Characterizing logics. In J. Barwise and S. Feferman, *Model-Theoretic Logics* (pp. 77–120). Berlin: Springer Verlag.

Grzegorczyk, A. (1974) *An Outline of Mathematical Logic: Fundamental Results and Notions Explained in all Details*. Dordrecht: Reidel.

Hunter, G. (1971) *Metalogic: An Introduction to the Metatheory of Standard First Order Logic*. London: Macmillan.

Kleene, S. C. (1952) *Introduction of Metamathematics*. Groningen: P. Noordhoff.

Manzano, M. (1996) *Extensions of First-Order Logic*. Cambridge: Cambridge University Press.

Shapiro, S. (1991) *Foundations without Foundationalism: The Case for Second-order Logic*. Oxford: Clarendon Press.

Shapiro, S. (ed.), (1996) *The Limits of Logic: Higher-Order Logic and the Löwenheim–Skolem Theorem*. Aldershot: Dartmouth.

22

Logic in Finite Structures: Definability, Complexity, and Randomness

SCOTT WEINSTEIN

1 Validity in the Finite

Is it simpler to reason about finite structures or about arbitrary structures? Some of the major results of logic in the twentieth century provide a clear and surprising answer to one precise version of this question. Suppose first that we restrict our reasonings to propositions which are expressible in first-order logic. We may then understand the question as asking for a comparison between the complexity of

1. determining whether a first order sentence is valid, that is, true under every interpretation whatsoever, and
2. determining whether a first-order sentence is valid in the finite, that is, true under every interpretation with a finite universe of discourse.

This question can be formulated more concisely and concretely in terms of Val, the set of valid sentences of L, the first order language with identity and a single binary relation symbol E, and Fval, the set of sentences of L which are valid in the finite, namely: is the decision problem for Fval simpler than the decision problem for Val?

Let's begin by analyzing the complexity of the decision problem for Fval. It is easy to see that we can make an effective list $A_1, A_2, \ldots$ of finite structures for L which contains every such structure up to isomorphism. We may now subject a sentence $\varphi \in L$ to the following effective procedure: successively test whether A_1 satisfies φ, A_2 satisfies $\varphi, \ldots$; at the first stage where the outcome is negative, halt the procedure and return the answer 'no.' Clearly, this procedure yields the correct answer to the query 'is φ valid in the finite,' if the answer is negative, and yields no answer, otherwise. That is, the complement of Fval is recursively enumerable, or in other words, Fval is co-r.e.

If we attempt such a direct approach to analyzing the complexity of Val, we are stymied at the outset. There is no possibility of effectively generating a list of all structures up to isomorphism, since there are structures of every infinite cardinality; moreover, there is, in general, no effective way to test whether a given infinite structure A satisfies a sentence $\varphi \in L$. Reflection on the apparent complexity of the notion of validity provides the proper context in which to appreciate the extraordinary depth of Gödel's Completeness Theorem for first-order logic: there is a sound and complete

effective proof procedure for first-order validity. In other words, Val is recursively enumerable – in order to discover that a first-order sentence is valid, if it is, we need only look through an effectively generated list of finite objects and check that one is its proof.

So far so good: Val is r.e.; Fval is co-r.e. To complete the picture we need to invoke two more fundamental results of twentieth-century logic. Church's Theorem tells us that Val is undecidable, from which it follows that Val is not co-r.e. On the other hand, Trakhtenbrot's Theorem (see Trakhtenbrot 1950) tells us that Fval is undecidable, from which it follows that Fval is not r.e., that is, there is no sound and complete proof procedure for the first-order sentences which are valid in the finite. This suggests one answer to the question with which we began: reasoning about finite structures is no simpler than reasoning about arbitrary structures – there is an effective proof procedure for validity, but no effective proof procedure for validity in the finite. Indeed, there is a good sense in which we can say that the complexity of the decision problems for Val and Fval are identical, namely, Val and Fval are Turing reducible to one another. That is, there is a Turing machine which will decide membership in Val given an oracle for Fval and there is a Turing machine which will decide membership in Fval given an oracle for Val. Remarkably, Val and Fval turn out to have effectively the same information content.

2 Model Theory in the Finite?

The last section suggests that, in a sense, there can be no proof theory for first-order logic in the finite, since there can be no effective proof procedure for validity in the finite. How about model theory? At the outset, there are disappointments. One of the central results in the model theory of first-order logic, the Compactness Theorem, does not extend to the finite case. Recall the Compactness Theorem: if every finite subset of a set of first order sentences Γ is satisfiable, then Γ itself is satisfiable. Call a set of sentences Γ *satisfiable in the finite*, if and only if, there is a finite structure A which satisfies every sentence in Γ. It is easy to construct a set of first order sentences Γ such that every finite subset of Γ is satisfiable in the finite, whereas Γ itself is not satisfiable in the finite. For example, let $\Gamma = \{\lambda_n \mid n > 0\}$, where λ_n is a first order sentence in the pure language of identity which is true in a structure A, if and only if, the size of A is at least n. Virtually all the finite analogs of well-known consequences of the Compactness Theorem fail as well, for example, the Beth Definability Theorem, the Craig Interpolation Theorem, most all 'preservation theorems,' etc. (See Gurevich (1984) for a compendium of such results; a notable exception is van Benthem's preservation theorem for the modal fragment of first-order logic, see Rosen (1997).)

Further contrasts between the finite model theory of first order logic and classical model theory abound. A central phenomenon of first order model theory is that no infinite structure can be characterized up to isomorphism by a set of first order sentences. Recall that structures A and B are elementarily equivalent, if and only if, they satisfy the same first-order sentences. It is a corollary of the Compactness Theorem that for every infinite structure A, there is a structure B (indeed, a proper class of pairwise non-isomorphic structures B) such that A is elementarily equivalent to B, but A is not isomorphic to B. In contrast, it is easy to show that for all structures A and B, if A is finite

and B is elementarily equivalent to A, then B is isomorphic to A. Indeed, for every finite structure A whose signature is finite, there is a single first-order sentence φ such that for every structure B, B satisfies φ, if and only if, B is isomorphic to A.

3 Definability and Complexity

In light of all these contrasts, one might legitimately wonder what finite model theory could be. The following sections attempt to answer this question by giving a feeling for some of the techniques, results, and open problems of the subject. For the most part, we will pursue questions in definability theory, that is, we will inquire into the expressive power of various logical languages in the context of finite structures. We will see that this study has close connections with the theory of computational complexity.

We collect together here some notions and notations that will ease our progress. A structure A, for us, consists of a universe of discourse $|A|$ and interpretations for a finite set of relation symbols and constant symbols; this set of symbols is called the signature of A. Whenever we mention two structures in the same breath, they are of the same signature; whenever we speak of a collection of structures, they are of the same signature. Let $\mathcal{K}$ be a class of structures. A collection of structures $\mathcal{Q} \subseteq \mathcal{K}$ is a *query relative to* $\mathcal{K}$, if and only if, $\mathcal{Q}$ is isomorphism closed in $\mathcal{K}$, that is,

$$\forall A,B \in \mathcal{K}((A \in \mathcal{Q} \wedge A \cong B) \rightarrow B \in \mathcal{Q}).$$

We will drop the qualification 'relative to $\mathcal{K}$' when the background class is clear from the context. Queries are the proper object of study in our investigation of definability and complexity, since logical languages do not distinguish between isomorphic structures.

We think of a logical language L as consisting of a set of sentences S_L and a satisfaction relation $\models_L$. We will suppress the subscript to $\models$ as it will generally be clear from the context. Given a class of structures $\mathcal{K}$ and a sentence $\varphi \in S_L$, we write $\varphi(\mathcal{K})$ for the *query defined by* φ *relative to* $\mathcal{K}$, that is,

$$\varphi(\mathcal{K}) = \{A \in \mathcal{K} \mid A \models \varphi\}.$$

We write $L(\mathcal{K})$ for $\{\varphi(\mathcal{K}) \mid \varphi \in S_L\}$, the set of queries which are L-definable relative to $\mathcal{K}$.

In what follows, we will analyze and compare the logical and computational complexity of queries relative to classes of finite structures. It will be convenient to introduce, for each signature σ, a canonical countable set of finite structures $\mathcal{F}_\sigma$ which contains, up to isomorphism, every finite structure of signature σ. We let $\mathcal{F}_\sigma$ be the set of structures of signature σ with universe of discourse $[n](= \{1, \ldots, n\})$ for some $n \geq 1$. Unless otherwise indicated, all collections of finite structures we mention are understood to be subsets of $\mathcal{F}_\sigma$ for some σ. We write $\mathcal{D}$ for $\mathcal{F}_{\{E\}}$ where E is a binary relation symbol; $\mathcal{D}$ is, for us, the class of finite directed graphs. For simplicity and concreteness, our discussion will often focus on queries relative to $\mathcal{D}$.

In the following sections, we will address questions concerning the logical resources that are required to define a given query $\mathcal{Q} \subseteq \mathcal{D}$. For example, we will consider whether $\mathcal{Q}$ is definable in second-order, but not in first-order, logic; or whether $\mathcal{Q}$ is definable by an existential second-order sentence, but not by the negation of such a sentence, etc. We can think of this study as yielding information about the complexity of $\mathcal{Q}$ – for example, if $\mathcal{Q}$ is not first-order definable, while $\mathcal{Q}'$ is, we might want to say that the definitional, or descriptive, complexity of $\mathcal{Q}'$ is no greater than that of $\mathcal{Q}$. In this way, we can think of the classes of queries $L(\mathcal{D})$, for various languages L, as descriptive complexity classes, in analogy with the resource complexity classes studied in the theory of computation (see Papadimitriou (1994) for background on the theory of computational complexity). Let us pursue this analogy.

Consider a query $\mathcal{Q} \subseteq \mathcal{D}$. We have been thinking of $\mathcal{Q}$ under the guise of definability. We can, on the other hand, think of $\mathcal{Q}$ as a decision problem: given an $A \in \mathcal{D}$ answer the question whether or not A is a member of $\mathcal{Q}$. Rather than asking what logical resources are required to specify $\mathcal{Q}$, we can ask instead, what computational resources are required to decide membership in $\mathcal{Q}$. To make this precise, we can easily encode each $A \in \mathcal{D}$ as a bit string, thereby making it a suitable input to a Turing machine. If A is of size n, the adjacency matrix of A is the $n \times n$ matrix whose i, j-entry is a 1, if $\langle i, j \rangle \in E^A$, and is a 0, otherwise. We encode A as the bit string $c(A)$ which consists of the concatenation of the rows of the adjacency matrix of A, and for $\mathcal{Q} \subseteq \mathcal{D}$, we let $c(\mathcal{Q}) = \{c(A) \mid A \subseteq \mathcal{Q}\}$. If Y is a resource complexity class, then we write $\mathsf{Y}(\mathcal{D})$ for the collection of queries $\mathcal{Q} \subseteq \mathcal{D}$ such that $c(\mathcal{Q}) \in \mathsf{Y}$. (In a similar fashion, we may define $\mathsf{Y}(\mathcal{F}_\sigma)$ for any signature σ.) We are now in a position to make direct comparisons between resource and descriptive complexity classes. In the following sections, we will see that many important resource complexity classes, for example, P and NP, have natural logical characterizations relative to various sets of finite structures.

4 First-Order Definability

One of the main tools for establishing limits on the expressive power of first-order logic over arbitrary structures is the Compactness Theorem. As noted earlier, we are deprived of the use of this tool in the context of finite structures, so we will need to rely on other techniques. We begin with an exemplary application of the Compactness Theorem, so we can appreciate what we are missing; the example will reappear throughout the following sections.

Let $\mathcal{D}^*$ be the collection of arbitrary structures A of signature $\{E\}$; each $A \in \mathcal{D}^*$ is a, perhaps infinite, directed graph. We call such a graph A *simple*, if and only if, E^A is irreflexive and symmetric, and we let $\mathcal{G}^*$ be the collection of arbitrary simple graphs. A simple graph may be visualized as a loop-free, undirected graph. Note that $\mathcal{G}^*$ is first-order definable relative to $\mathcal{D}^*$. Now let $\mathcal{D}^*_{st}$ (resp., $\mathcal{G}^*_{st}$) be the collection of expansions of structures in $\mathcal{D}^*$ (resp., $\mathcal{G}^*$) to the signature with two additional constant symbols s and t – this is the collection of directed (resp., simple) source–target graphs. A graph $A \in \mathcal{D}^*_{st}$ is *reachable*, if and only if, there is a path from s^A to t^A in A, that is, sequence $a_1, \ldots, a_n$ of nodes of A such that $a_1 = s^A$, $a_n = t^A$, and for every $1 \leq i < n$, $\langle a_i, a_{i+1} \rangle \in E^A$.

Let S^* be the collection of $A \in \mathcal{G}^*_{st}$ such that A is reachable. Is S^* first order definable relative to $\mathcal{G}^*_{st}$? An application of the Compactness Theorem provides a negative answer. For suppose that there is a first-order sentence φ with $\varphi(\mathcal{G}^*_{st}) = S^*$. Let Γ be the set consisting of the following sentences:

$\psi_0 \quad \neg s = t$

$\psi_1 \quad \neg Est$

$\psi_2 \quad \neg\exists x(Esx \wedge Ext)$

$\vdots \quad \vdots$

Notice that a graph A satisfies the conjunction $\psi_0 \wedge \ldots \wedge \psi_n$, if and only if, there is no path in A of length $\leq n$ from s^A to t^A. Therefore, the simple chain of length $n+1$ with end nodes labeled s and t satisfies $\psi_0 \wedge \ldots \wedge \psi_n$, from which it follows that every finite subset of $\Gamma \cup \{\varphi\}$ is satisfiable. Therefore, by the Compactness Theorem, $\Gamma \cup \{\varphi\}$ is satisfiable. On the other hand, it is clear that if a graph A satisfies Γ, then A is not reachable. But, this contradicts the hypothesis that φ defines S^*.

Now, let $S \subset S^*$ be the set of finite reachable simple source–target graphs. The question whether S is first-order definable is no longer immediately accessible to an application of the Compactness Theorem of the sort sketched above. The Compactness Theorem can be pressed into service to answer the question by exploiting 'pseudofinite' structures, that is, infinite structures which satisfy every first-order sentence which is valid in the finite (see Gaifman and Vardi (1985) for details); but, we will follow a different approach, due to Gurevich (1984), which proceeds via Ehrenfeucht games and yields additional information. The approach involves a reduction from a query on linear orders.

Let $\mathcal{L}_{st} \subseteq \mathcal{F}_{\{<,s,t\}}$ be the set of finite linear orders with minimal element s and maximal element t. The conjunction of the following first-order conditions defines $\mathcal{L}_{st}$.

$\forall x\neg(x < x)$ (irreflexive)

$\forall x\forall y\forall z((x < y \wedge y < z) \rightarrow x < z)$ (transitive)

$\forall x\forall y(x < y \vee y < x \vee x = y)$ (total)

$\forall x(\neg(x < s) \wedge \neg (t < x))$ (endpoints)

Let $\mathcal{M} \subseteq \mathcal{L}_{st}$ be the set of odd linear orders, that is, linear orders with universe $[2n+1]$, for some n. Is $\mathcal{M}$ first-order definable relative to $\mathcal{L}_{st}$?

Here is one strategy for attempting to show that $\mathcal{M}$ is not first-order definable. For each first-order sentence φ, show that there are $A, B \in \mathcal{L}_{st}$ such that A and B agree about φ (either they both satisfy φ or they both fail to do so), $A \in \mathcal{M}$, and $B \notin \mathcal{M}$. It is clear that if we succeed in doing this, we have shown that $\mathcal{M}$ is not first-order definable. (Indeed, the converse holds as well – the strategy is nothing more than a restatement of what's required.) What makes the strategy worth pursuing is that there is a powerful, and entertaining, technique, the Ehrenfeucht game, for showing that pairs of structures agree about first-order sentences. This technique applies to both finite and infinite structures and, to some extent, fills the void left by the failure of compactness in finite model theory.

The Ehrenfeucht game is played between two players, conventionally called the Spoiler and the Duplicator. The equipment for the game consists of two boards, one representing the graph A and the other representing the graph B, and an unlimited supply of pairs of pebbles $\langle\alpha_1, \beta_1\rangle, \langle\alpha_2, \beta_2\rangle, \ldots$. The game is played through a sequence of rounds as follows. At the ith round of the game, the Spoiler chooses one of the pebbles from the pair $\langle\alpha_i, \beta_i\rangle$ and places it on a node of the corresponding board A or B, the α pebbles are played onto A and the β pebbles onto B. The Duplicator then places the remaining pebble on the other board, completing the round of play. Suppose the game has proceeded through n-rounds of play. Let a_i be the node in A covered by α_i and let b_i be the node in B covered by β_i. Let f be the mapping which sends a_i to b_i for all $1 \leq i \leq n$ and sends s^A to s^B and t^A to t^B. If f is a partial isomorphism from A to B (that is, a one to one, edge preserving map) we say the Duplicator wins the game through n-rounds of play. Thus, the Spoiler's goal is to reveal structural distinctions between A and B, the Duplicator's goal is to hide them. We say that A is n-similar to B, if and only if, the Duplicator has a strategy to win every play of the Ehrenfeucht game on A and B through n-rounds. We say structures A and B are n-equivalent, if and only if, A and B satisfy exactly the same first-order sentences of quantifier rank $\leq n$ (recall that the quantifier rank of a formula is the maximum depth of nesting of quantifiers in the formula). The Ehrenfeucht–Fraïssé Theorem tells us that n-similarity and n-equivalence coincide, that is, for all structures A and B and for every n, A is n-similar to B, if and only if, A is n-equivalent to B (see Ehrenfeucht 1961; and Fraïssé 1954).

Armed with the Ehrenfeucht–Fraïssé Theorem, we can now implement our strategy for showing that $\mathcal{M}$ is not first order definable. For each n, it suffices to construct a pair of finite linear orders A and B such that $A \in \mathcal{M}$, $B \notin \mathcal{M}$, and A is n-similar to B. We accomplish this by overkill – for each n, if A and B are finite linear orders of length $>2^n$, then A is n-similar to B. To see this, consider the following strategy for the Duplicator in the n-round game played on two such linear orders. At round m, the Duplicator plays as follows. Suppose, without loss of generality, that the Spoiler has played into A. This play falls into one of m intervals into which A has been divided by the play of pebbles at earlier rounds of the game and it determines distances d_1 and d_2 between the newly pebbled point and the left and right endpoints of that interval, respectively. The Duplicator plays into the corresponding interval in B so as to achieve the following approximation between these distances and the corresponding distances d_1' and d_2' between the point he/she pebbles and the endpoints of his/her interval. Namely, for $i = 1, 2$ if $d_i \leq 2^{(n-m)}$, then $d_i = d_i'$, and if $d_i > 2^{n-m}$, then $d_i' > 2^{n-m}$. The initial condition on the lengths of A and B insures that the Duplicator can maintain these approximations through n-rounds of play. Thus, $\mathcal{M}$ is not first-order definable. Indeed, any first-order definable collection of finite linear orders is a finite or cofinite subset of $\mathcal{L}_{st}$.

Now, we reduce the problem of defining odd length linear orders ($\mathcal{M}$) to the problem of defining reachability (S). Let $\rho(x, y)$ be a first-order formula which is true of a pair of elements of a linear order, if and only if, the second is the successor of the successor of the first, and let $\chi(x, y)$ be the formula $\rho(x, y) \vee \rho(y, x)$. Suppose $A \in \mathcal{L}_{st}$. We may use the formula χ to define a simple source–target graph B from A. We let $|B| = |A|$, $s^B = s^A$, $t^B = t^A$, and $E^B = \{\langle u, v\rangle \mid A \models \chi[u, v]\}$. Now, observe that the graph B thus defined is reachable, if and only if, $A \in \mathcal{M}$. Suppose that there is a first-order sentence θ which defines S. Let θ' be the result of replacing each subformula of the form Exy in

θ with $\chi(x, y)$. Then, θ' defines $\mathcal{M}$. We have exhibited a 'first-order reduction' of $\mathcal{M}$ to S; it follows at once that S is not first-order definable, since $\mathcal{M}$ is not. Such first-order reductions are an important descriptive analog of the resource bounded reductions of computational complexity theory.

The foregoing examples show that some simple properties of finite graphs are not first-order definable. These examples can be easily multiplied – acyclicity, regularity, 2-colorability, etc. all fail to be first-order definable. Lest the reader be left with the impression that no interesting classes of finite graphs are first-order definable, note that the collection $\mathcal{FR}$ of finite nonempty ranks of the cumulative hierarchy of sets equipped with the membership relation as their edge relation is first-order definable (see Dawar et al. 1998). In Section 6, we will see that questions concerning the expressive power of first-order logic relative to $\mathcal{FR}$ are directly related to open problems in the theory of computational complexity.

5 Second-Order Definability

What logical resources are required to define reachability over finite graphs? As we've just seen, first-order logic doesn't suffice. There are several routes to the definability of reachability. Let's begin with Frege's (1884). The transitive closure (sometimes called the ancestral) of a binary relation R is the smallest relation (in the sense of inclusion) which is transitive and includes R. For example, the relation 'ancestor of' is the transitive closure of the relation 'parent of.' If R is a binary relation, we write $\mathrm{tc}(R)$ for the transitive closure of R.

Frege observed that the relational operator tc is uniformly definable by a formula $\tau(x, y)$ of second-order logic; that is, for every structure $A \in \mathcal{D}^*$, $\mathrm{tc}(E^A) = \{\langle u, v\rangle \mid A \models \tau[u, v]\}$. The formula $\tau(x, y)$ may be chosen to be:

$$\forall P((\forall z(Exz \to Pz) \wedge \forall v \forall w((Pv \wedge Evw) \to Pw)) \to Py).$$

This formula has a couple of noteworthy features. First, it is a universal second-order formula, that is, it is of the form

$$\forall P_1 \ldots \forall P_n \theta$$

with θ first order. Second, it is monadic universal, that is, each of the universal quantifiers binds a monadic second-order variable. We call the fragment of second-order logic consisting of all such formulas mon-Π^1_1. Now, let $\mathcal{R}^* \subseteq \mathcal{D}^*_s$ be the collection of reachable directed source–target graphs. It is clear that $\tau(s, t)$ defines $\mathcal{R}^*$ relative to $\mathcal{D}^*_s$; directed reachability is mon-Π^1_1 definable.

Is $\mathcal{R}^*$ also definable by a monadic existential second-order sentence? Since the full existential fragment of second-order logic is compact, the argument we gave at the beginning of Section 3 to show that S^* is not first-order definable, also shows that S^* (and hence $\mathcal{R}^*$ as well) is not definable by an existential second-order sentence, monadic or otherwise. In the finite case, the situation is subtler. Paris Kanellakis observed (see

Immerman 1999) that S is definable by a monadic existential second-order sentence $\exists P\theta$, where θ is the conjunction of the following first order conditions.

$$Ps \wedge \exists! x(Px \wedge Esx) \quad (s \text{ has degree 1 in } P)$$

$$Pt \wedge \exists! x(Px \wedge Etx) \quad (t \text{ has degree 1 in } P)$$

$$\forall x((Px \wedge x \neq s \wedge x \neq t) \rightarrow \exists y \exists z(Py \wedge Pz \wedge y \neq z \wedge \forall w(Pw \rightarrow (Exw \leftrightarrow (w = y \vee w = z))))) \quad (\text{all other nodes have degree 2 in } P)$$

If a finite simple graph A satisfies θ with respect to an assignment of a set of nodes X to P, then the nodes in X form a simple chain with end nodes s^A and t^A. (The reader should construct an infinite simple graph which is not reachable, but satisfies $\exists P\theta$.)

Let $\mathcal{R} \subset \mathcal{R}^*$ be the collection of finite reachable source–target graphs; this class differs from S in omitting the requirement of simplicity. Ajtai and Fagin (1990) established that $\mathcal{R}$ is not definable by a monadic existential second-order sentence. Their argument blends an extension of the Ehrenfeucht game to monadic existential second-order logic with probabilistic techniques (see Section 8 for a discussion of such techniques). This result establishes a difference in the descriptive complexity of S and $\mathcal{R}$, the former is definable in both mon-Π^1_1 and mon-Σ^1_1 (the monadic existential fragment of second-order logic), the latter only in mon-Π^1_1. From an intuitive point of view, the problem of determining whether a finite directed graph is reachable is more complex than the same problem restricted to simple graphs. It appears that descriptive complexity provides a more convincing account of this intuitive distinction than analysis of the computational complexity of these problems has yet been able to offer (see Ajtai and Fagin (1990) for further discussion).

The foregoing considerations leave open the question whether $\mathcal{R}$ is definable by an existential second-order sentence not subject to the monadic restriction. Rather than exhibiting such a sentence directly, which is straightforward, we will see that a positive answer to this question is a corollary of a celebrated result of Fagin (1974), namely: for all σ, $\mathsf{NP}(\mathcal{F}_\sigma) = \Sigma^1_1(\mathcal{F}_\sigma)$ (Σ^1_1 is the set of existential second-order sentences). Fagin's Theorem has been dubbed the first theorem of descriptive complexity theory. It equates the important computational complexity class of queries whose decision problems are solvable by nondeterministic Turing machines in polynomial time with the descriptive complexity class of queries which are definable by existential second-order sentences. Fagin's Theorem provides a machine independent characterization of NP – in order to verify that a query is in NP, one needn't tinker with machines and time bounds, just produce a Σ^1_1 sentence which defines it. In a sense, Fagin's Theorem shows that existential second-order logic is an alternative, what might be called, 'higher-level,' programming language for specifying exactly the NP queries: the proof of the theorem yields an effective procedure F for 'compiling' an arbitrary existential second-order sentence φ into a polynomially clocked nondeterministic Turing machine $F(\varphi)$ which accepts the query defined by φ and establishes that every query in NP is accepted by one of the machines $F(\varphi)$. Thus, existential second-order logic yields an effective enumeration of the NP queries, with the relation of satisfaction as the enumerating relation.

To return to our story of reachability, $\mathcal{R}$ is in NP – indeed it is in NL, the class of problems solvable by nondeterministic Turing machines using only logarithmic work space, and this class is included in P the class of problems solvable by deterministic Turing machines in polynomial time. It is generally believed that both the inclusions $\mathsf{NL} \subseteq \mathsf{P}$ and $\mathsf{P} \subseteq \mathsf{NP}$ are strict, but three decades of intense investigation have failed to produce a proof for the strictness of either. Fagin's Theorem opened up the possibility of attacking such outstanding problems in the theory of computational complexity by means of logical techniques. For example, in order to show that $\mathsf{P} \neq \mathsf{NP}$, it would suffice to show that there is a query $\mathcal{Q}$ such that $\mathcal{Q} \notin \Sigma^1_1(\mathcal{D})$ and $\mathcal{Q} \in \Pi^1_1(\mathcal{D})$, for, by Fagin's Theorem, this would establish that NP is not closed under complementation. The results mentioned earlier on the monadic fragments of Π^1_1 and Σ^1_1 are of some interest in this connection. We saw that $\mathcal{R} \in \text{mon-}\Pi^1_1(\mathcal{D})$ whereas $\mathcal{R} \notin \text{mon-}\Sigma^1_1(\mathcal{D})$. This does not resolve any outstanding problem concerning computational complexity since mon-Σ^1_1 does not correspond to any natural level of computational complexity. On the one hand, as we've just noted, $\mathcal{R}$ is in NL but not in mon-Σ^1_1. On the other hand, mon-Σ^1_1 contains NP-complete problems, that is, problems which are of maximal complexity among problems in NP with respect to polynomial time reduction. For example, the NP-complete query graph 3-colorability is easily seen to be in mon-Σ^1_1. Thus, though the result of Ajtai and Fagin (1990) does not lead to a separation of computational complexity classes, it does indicate how logic can contribute to a richer understanding of complexity by focusing attention on complexity classes which are orthogonal to the standard computational complexity measures, yet natural from a descriptive point of view.

6 Inductive Definability

In this section, we will pursue a more constructive approach to the definability of the set of reachable graphs. We will see that there are interesting connections between constructivity and complexity in this context.

One of the outstanding open problems of descriptive complexity theory concerns the existence of logics which characterize computational complexity classes below NP. An important result, due independently to Immerman (1986) and Vardi (1982), is that P is characterized by FO + LFP relative to ordered finite structures. FO + LFP is the extension of first-order logic by a least fixed point operator for defining relations by induction. Least fixed point operators have played a major role in studies of definability on fixed infinite structures (see Moshovakis 1974). Let $\varphi(R, x_1, \ldots, x_k)$ be a first-order formula with a distinguished k-ary relation symbol R. On a structure, A, we can use φ to define the relational operator, $\Phi_A(X) = \{\langle a_1, \ldots a_k\rangle \mid A \models \varphi[X, a_1, \ldots, a_k]\}$ (here, X is a k-ary relation on A and the notation stands for the assignment of X to R). If φ is an R-positive formula, Φ_A is monotone in the sense that for all $X \subseteq Y \subseteq |A|^k$, $\Phi_A(X) \subseteq \Phi_A(Y)$. We may view φ as determining an induction on A the stages of which are defined as follows: $\varphi^0_A = \emptyset$; $\varphi^{m+1}_A = \Phi_A(\varphi^m_A)$. Since Φ_A is monotone and A is finite, it follows immediately that for some m, $\varphi^m_A = \varphi^{m+1}_A$. The least such m is called the *closure ordinal* of φ on A and is denoted $\|\varphi\|_A$. It is easy to see that $\|\varphi\|_A \leq l^k$, for a finite structure A of size l (in the case of an infinite structure A, the closure ordinal of an induction may be a transfinite ordinal α whose cardinality is equal to the cardinality of $|A|$). Moreover, one can

readily verify that for $m = \|\varphi\|_A$, φ_A^m is the *least fixed point* (lfp) of the relational operator Φ_A, that is, $\Phi_A(\varphi_A^m) = \varphi_A^m$ and for all $X \subseteq |A|^k$, if $\Phi_A(X) = X$, then $\varphi_A^m \subseteq X$. We use φ_A^∞, to denote the least fixed point of the operator Φ_A. For example, if $\chi(R, x, y)$ is the formula

$$Exy \vee \exists z(Exz \wedge Rzy)$$

then for every structure $A \in D$, χ_A^∞ is the transitive closure of E^A. We write FO + LFP for the extension of first-order logic with the lfp operation which uniformly determines the least fixed point of an R-positive formula. That is, for any R-positive formula φ, $\mathrm{lfp}(R, x_1, \ldots, x_k)\varphi$ is a formula of FO + LFP and $A \models \mathrm{lfp}(R, x_1, \ldots, x_k)\varphi[\bar{a}]$ if and only if, $\bar{a} \in \varphi_A^\infty$.

Let us attend once again to reachability. For $\chi(R, x, y)$ as above, the sentence $\mathrm{lfp}(R, x, y)\chi(s, t)$ defines $\mathcal{R}$ relative to $\mathcal{D}$. This approach to the definability of $\mathcal{R}$ has been regarded as more constructive than the Fregean approach described in the preceding section: many find the general notion of iteration to be more transparent than universal second-order quantification. Since, as we will see in the next section, FO + LFP ($\mathcal{D}$) is *properly* included in $\mathsf{P}(\mathcal{D})$, the 'more constructive' approach actually yields a stronger bound on the descriptive complexity of $\mathcal{R}$. It is interesting to observe, as a corollary of Fagin's Theorem and the Immerman–Vardi Theorem, that in the case of finite ordered structures, the relative power of first-order positive induction versus universal second-order quantification amounts exactly to the question whether $\mathsf{P} = \mathsf{NP}$.

Let us look a bit more carefully at the case of ordered structures. For simplicity, let's focus on the set $\mathcal{O} \subseteq \mathcal{F}_{\langle E,<\rangle}$ of ordered graphs – a structure A is a member of $\mathcal{O}$, if and only if, the reduct of A to $\{E\}$ is in $\mathcal{D}$ and the reduct of A to $\{<\}$ is in $\mathcal{L}$, the set of finite linear orders. The Immerman–Vardi Theorem tells us that FO + LFP($\mathcal{O}$) = P($\mathcal{O}$). It follows from the results of Section 4 that the set of ordered graphs of odd size, a query in P($\mathcal{O}$), is not first-order definable relative to $\mathcal{O}$. We may conclude that that FO($\mathcal{O}$) is properly included in FO + LFP($\mathcal{O}$). In fact, there is no known example of an infinite query $\mathcal{Q} \subseteq \mathcal{O}$ such that FO($\mathcal{Q}$) = FO + LFP($\mathcal{Q}$). Kolaitis and Vardi (1992a) conjectured that for every infinite query $\mathcal{Q} \subseteq \mathcal{O}$, FO($\mathcal{Q}$) is properly included in FO + LFP($\mathcal{Q}$). This Ordered Conjecture is an important open problem in finite model theory which turns out to have connections to a number of open problems in the theory of computational complexity. Even the special case of this conjecture concerning the power of first-order versus fixed point definability relative to the set $\mathcal{FR}$ of finite ranks of the cumulative hierarchy of sets is open, and its resolution would have significant complexity theoretic consequences (see Dawar et al. 1996; Gurevich et al. 1994). (This counts as a special case, since a linear order is uniformly first-order definable on the structures in $\mathcal{FR}$, see Dawar et al. (1998).)

The Ordered Conjecture asks whether there is an infinite set of finite ordered structures relative to which first-order logic characterizes polynomial time computability. If we turn our attention away from ordered structures, we can formulate what has been regarded as the central open problem of descriptive complexity theory, namely: Is there a logical characterization of polynomial time computability over structures without a built-in order? Gurevich (1988) has given a rigorous formulation of this question. In connection with Fagin's Theorem, we noted that existential second-order logic characterizes NP in a strong sense – not only is $\mathrm{NP}(\mathcal{F}_\sigma) = \Sigma_1^1(\mathcal{F}_\sigma)$, for all σ; there is an effective

procedure for transforming sentences of existential second-order logic into polynomially clocked nondeterministic Turing machines that witness the membership of the queries they define in NP. Likewise, in the case of P, we can ask if there is a logic $L = \langle S_L, \models_L \rangle$ such that both S_L and $\models_L$ are recursive and

1. $L(\mathcal{T}_\sigma) = \mathrm{P}(\mathcal{T}_\sigma)$;
2. there is an effective procedure F such that for every $\varphi \in S_L$, $F(\varphi)$ is a polynomially clocked deterministic Turing machine which accepts $c(\varphi(\mathcal{T}_\sigma))$.

We call a logic meeting these requirements a logic for P. A logic for P amounts to an effective list of polynomially clocked deterministic Turing machines, each of which decides a query, and which lists at least one machine deciding each query in P. The difficulty in constructing such an effective list lies in the requirement that the machines must decide queries, that is, isomorphism invariant sets of structures. The set of machines meeting this requirement is not recursively enumerable. This is not fatal to the enterprise of constructing a logic for P, since we do not need to enumerate all the polynomially clocked, isomorphism invariant machines, just a rich enough subset of them. An obvious way to proceed would be as follows. A function C: $\mathcal{D} \mapsto \mathcal{D}$ is called a graph canon, if and only if,

1. $\forall G \in \mathcal{D}(G \cong C(G))$, and
2. $\forall G, H \in \mathcal{D}(G \cong H \rightarrow C(G) = C(H))$.

A graph canon extracts a unique representative from each equivalence class of $\mathcal{D}$ under the equivalence relation of isomorphism. If there is a graph canon C that is computable in polynomial time, then there is a logic for P. This is easily seen by composing C with an effective list of polynomially clocked deterministic Turing machines which, for each set of strings $X \in \mathrm{P}$, includes a machine which decides X – such an effective list can be constructed absent the requirement that the machines decide queries. It is well-known that if $\mathrm{P} = \mathrm{NP}$, then there is a polynomial time computable graph canon, which yields the conclusion that if there is no logic for P, then $\mathrm{P} \neq \mathrm{NP}$. There is no evidence that the converse holds, and the quest for a logic for P remains an active area of research in descriptive complexity theory.

7 Infinitary Logics

In this section, we investigate a measure of logical complexity that has played a prominent role in recent research in finite model theory. The measure is the total number of variables, both free and bound, which occur in a formula of first-order logic, or its infinitary extension, $L_{\infty\omega}$. First-order sentences which involve the reuse of bound variables within the scopes of quantifiers already binding those same variables are generally frowned on from a pedagogical and stylistic point of view. Thus, the study of finite variable fragments of first-order logic and infinitary logic, whose point is to exploit the possibility of such reuse, typically seems a bit unusual, if not perverse, to most logicians.

Consider the following sequence of first-order sentences, each of which contains occurrences of only the two variables x_1 and x_2:

φ_0 Est

φ_1 $\exists x_1(Esx_1 \wedge Ex_1t)$

φ_2 $\exists x_1\exists x_2(Esx_1 \wedge Ex_1x_2 \wedge Ex_2t)$

φ_3 $\exists x_1\exists x_2(Esx_1 \wedge Ex_1x_2 \wedge \exists x_1(Ex_2x_1 \wedge Ex_1t))$

φ_4 $\exists x_1\exists x_2(Esx_1 \wedge Ex_1x_2 \wedge \exists x_1(Ex_2x_1 \wedge \exists x_2(Ex_1x_2 \wedge Ex_2t)))$

$\vdots$ $\vdots$

Clearly, the sentences φ_i are pairwise inequivalent (consider the structures A_n for $n > 1$ which interpret E as the successor relation on $[n]$ and assign 1 to s and and n to t; $A_n \models \varphi_i$, if and only if, $i + 2 = n$). Note that although the sentences involve only two variables, their quantifier rank is unbounded. Needless to say, these sentences cannot be brought to prenex normal form without increasing the number of variables.

The logic $L_{\infty\omega}$ is the infinitary extension of first-order logic which is closed under the formation of arbitrary conjunctions and disjunctions of sets of formulas. In Section 2, we observed that every finite structure is characterized up to isomorphism by a single first-order sentence, from which it follows that for every signature σ, every query $\mathcal{Q} \subseteq \mathcal{F}_\sigma$ is $L_{\infty\omega}$ definable. Thus, $L_{\infty\omega}$ is too strong to be of interest from the point of view of finite model theory. Let us consider the weaker finite variable fragments of $L_{\infty\omega}$. We define $L^k_{\infty\omega}$ to be the k-variable fragment of $L_{\infty\omega}$, that is, $L^k_{\infty\omega}$ consists of all formulas of $L_{\infty\omega}$ all of whose individual variables, either free or bound, are among $x_1, \ldots, x_k$. We let $L^\omega_{\infty\omega} = \cup_{k<\omega} L^k_{\infty\omega}$. For example, let θ, a sentence of $L^2_{\infty\omega}$, be the infinite disjunction of the sentences φ_0, $\varphi_1, \ldots,$ exhibited above. Observe that θ defines $\mathcal{R}$ (directed reachability) relative to $\mathcal{D}$ (the set of finite directed graphs). This is no accident: Kolaitis and Vardi (1992b) established that for every σ, FO + LFP($\mathcal{F}_\sigma$) $\subseteq L^\omega_{\infty\omega}(\mathcal{F}_\sigma)$. Thus, the finite variable fragment of infinitary logic provides a tool for analyzing inductive definability over finite structures.

One of the main techniques for studying $L^\omega_{\infty\omega}$ definability is the k-pebble game, a variant of the Ehrenfeucht game, essentially due to Barwise (1977). In the k-pebble game, instead of an unlimited supply of pebble pairs, the equipment contains only the pebble pairs $\langle\alpha_1, \beta_1\rangle, \ldots \langle\alpha_k, \beta_k\rangle$. At each round of play, the Spoiler may now either play a pebble from a pair that has not yet been played and place it on the associated board, or move a pebble that has already been played to a new position. As before, the Duplicator must follow by moving the matched pebble on the other board. The winning condition for the n-round game remains the same as before. There is also an infinite version of the k-pebble game which we call the eternal k-pebble game. In this version, play continues through a sequence of rounds of order type ω. The Spoiler wins a play of the eternal game, if and only if, he wins at some finite round; otherwise, the Duplicator wins. We say that structures A and B are indistinguishable by sentences of $L^k_{\infty\omega}$ ($A \equiv^k_{\infty\omega} B$), if and only if, for every sentence $\varphi \in L^k_{\infty\omega}$,

$$A \models \varphi \Leftrightarrow B \models \varphi.$$

Barwise proved that the Duplicator has a winning strategy for the eternal k-pebble game played on A and B, if and only if, $A \equiv^k_{\infty\omega} B$. Thus, we can show that a query $\mathcal{Q} \subseteq \mathcal{D}$ is not $L^k_{\infty\omega}$ definable by exhibiting structures A, $B \in \mathcal{D}$, such that $A \in \mathcal{Q}$, $B \notin \mathcal{Q}$, and the Duplicator has a winning strategy for the eternal k-pebble game played on A and B.

As an illustration of this technique, we show that $\mathrm{P}(\mathcal{D}) \not\subseteq L^\omega_{\infty\omega}(\mathcal{D})$. We say that $A \in \mathcal{D}$ is an *empty graph*, if and only if, $E^A = \emptyset$. It is easy to see, by playing the k-pebble game, that for all empty graphs A and B, if A and B both have at least k nodes, then $A \equiv^k_{\infty\omega} B$. It follows at once that the set of graphs which have an odd number of nodes, a query in P, is not definable in $L^\omega_{\infty\omega}$. It also follows that the languages $L^k_{\infty\omega}$ form a strict hierarchy in terms of expressive power relative to $\mathcal{D}$. We will meet $L^\omega_{\infty\omega}$ again in the next section.

8 Random Graphs and 0–1 Laws

In this section, we will take up some connections between finite model theory and combinatorics. We focus attention on the study of random graphs, an active area of research in contemporary combinatorics.

Random graphs

Consider the following procedure for determining a directed graph with node set $[n]$. For each of the n^2 ordered pairs of nodes flip a fair coin to determine whether or not there is a directed edge from the first to the second; we assume the outcomes of the tosses are mutually independent. For each n, this procedure gives rise to the uniform probability distribution over $\mathcal{D}_n$, the collection of directed graphs with node set $[n]$. We may use this probability distribution to answer questions about how many graphs there are with certain properties. We write $\Pr_n(\theta)$ for the probability (with respect to this distribution) that a graph with node set $[n]$ satisfies θ. Note that,

$$\Pr_n(\theta) = \frac{\mathrm{card}\{G \in \mathcal{D}_n \mid G \models \theta\}}{\mathrm{card}\mathcal{D}_n}$$

We will be interested in the behavior of $\Pr_n(\theta)$ as a function of n for various choices of θ. We write $\Pr(\theta) = \lim_{n\to\infty} \Pr_n(\theta)$. In general, $\Pr(\theta)$ may not be defined. For example, when $\theta \in \Sigma^1_1$ expresses the condition that there are an even number of nodes, $\Pr_n(\theta)$ endlessly oscillates between the values 0 and 1 and thus has no well defined limit. On the other hand, many interesting graph theoretic properties do possess a 'limit probability' with respect to the uniform distribution. We will see how logic provides some explanation of this fact.

Let us begin with the example of connectivity: a directed graph A is connected, if and only if, for each pair i, j of distinct nodes of A, there is a path from i to j. Let θ be the sentence of FO + LFP that defines the set of connected graphs relative to $\mathcal{D}$. We wish to discover whether $\Pr(\theta)$ is well defined, and if it is, whether we can determine its

value. In order to do so, we will attempt to approximate the value of $\Pr_n(\theta)$ for large values of n.

Rather than dealing directly with θ, let us consider the following first order condition which implies θ. Let φ be the following sentence:

$$(\forall x)(\forall y)(x \neq y \rightarrow (\exists z)(x \neq z \wedge y \neq z\ (Exz \wedge Ezy).$$

The sentence φ expresses the 'two degrees of separation' property – we can proceed from any node to any other by a path of length two. Clearly, φ implies θ. Hence, for all n,

$$\Pr_n(\varphi) \leq \Pr_n(\theta).$$

Therefore, if we can show that $\Pr_n(\varphi)$ becomes large, as a function of n, the same will be true of $\Pr_n(\theta)$.

Let's perform the calculation. Fix a pair of distinct nodes $i, j \in [n]$. We say that a node k links i to j, if and only if, there is an edge from i to k and an edge from k to j. Clearly, for any fixed node k, distinct from i and j, the probability that k does not link i to j is $.75$. So the probability that no node distinct from i and j links i to j is $(.75)^{n-2}$. Now, there are $n(n-1)$ ordered pairs of distinct nodes in $[n]$. Therefore, the probability that some pair of distinct nodes in $[n]$ fail to be linked is bounded from above by $n(n-1)\cdot(.75)^{n-2}$. That is,

$$\Pr_n(\neg\varphi) \leq n(n-1)\cdot(.75)^{n-2}.$$

It is easy to show that

$$\lim_{n\to\infty} n(n-1)\cdot(.75)^{n-2} = 0.$$

It follows at once that

$$\Pr(\theta) = \Pr(\varphi) = 1.$$

So we have succeeded in analyzing the limiting behavior of graph connectivity by reducing the problem to a simple calculation of the limiting behavior of a first-order condition; and the limit probability of that condition is 1. To what extent can we generalize this example?

0–1 Laws

In this section we will consider a sweeping generalization of the preceding example of connectivity. We say that a logical language L satisfies the 0–1 law with respect to the uniform distribution over directed graphs, if and only if, for every sentence φ of L,

$$\Pr(\varphi) = 0 \quad \text{or} \quad \Pr(\varphi) = 1.$$

A bold generalization of the example of connectivity would be the following: FO + LEP satisfies the 0–1 law for the uniform distribution over directed graphs. Indeed, this generalization is true, as was established by Blass et al. (1985). This result itself generalized the 0–1 law for first-order logic due to Fagin (1976) and Glebskij et al. (1969). A striking generalization of these (and additional) results, which provides a beautiful explanation for the limiting behavior of a variety of graph theoretic properties, is the following 0–1 law for $L^{\omega}_{\infty\omega}$ due to Kolaitis and Vardi (1992b): $L^{\omega}_{\infty\omega}$ satisfies the 0–1 law for the uniform distribution over directed graphs. Not only does this result generalize the example of connectivity given above; its proof also follows the lines of the argument given for the example. In particular, the theorem is a corollary of the following fascinating result, also due to Kolaitis and Vardi (1992b): For every $k \geq 2$, there is a k-variable first order sentence γ_k such that

1. $\Pr(\gamma_k) = 1$, and
2. for every sentence $\theta \in L^{k}_{\infty\omega}$, either $\gamma_k \models \theta$ or $\gamma_k \models \neg\theta$.

In other words, for each k, there is a single first-order sentence which has limit probability 1 with respect to the uniform distribution on directed graphs and axiomatizes a complete $L^{k}_{\infty\omega}$ theory.

The sentence γ_k may be constructed as follows. A k-literal is a formula of the form Ex_ix_j or its negation with $1 \leq i, j \leq k$. A basic k-type is a maximal consistent conjunction of k-literals. A k-extension condition is a sentence of the form:

$$\forall x_1 \ldots \forall x_{k-1}\Big(\Big(\bigwedge_{i \neq j} x_i \neq x_j \wedge \varphi\Big) \rightarrow \exists x_k\Big(\bigwedge_{i<k} x_i \neq x_k \wedge \psi\Big)\Big),$$

where φ is a $(k+1)$-type, ψ is a k-type, and ψ extends φ. A graph satisfies such a k-extension condition, if and only if, each of its size $k-1$ subgraphs of type φ can be extended to a size k subgraph of type ψ. We let γ_k be the conjunction of all the l-extension conditions for $2 \leq l \leq k$. The sentence γ_k expresses a 'bounded principle of plenitude:' every subgraph of size $l < k$ can be extended in every possible way to a subgraph of size $l+1$ (compare the two degrees of separation principle above). For $k \geq 3$, it is not at first sight obvious that there are finite structures with satisfy γ_k. However, an easy computation, of just the sort sketched for the two degrees of separation principle, reveals that $\Pr(\gamma_k) = 1$ for all $k \geq 2$. That is, for every $\varepsilon > 0$, for large enough n, all but an ε fraction of the directed graphs of size n satisfy γ_k.

In order to verify that γ_k axiomatizes a complete $L^{k}_{\infty\omega}$ theory, it suffices to show that for all directed graphs A, B, if $A \models \gamma_k$ and $B \models \gamma_k$, then $A \equiv^{k}_{\infty\omega} B$. But this follows directly from Barwise's characterization of $\equiv^{k}_{\infty\omega}$ given in Section 7, since it is easy to see that the Duplicator has a winning strategy for the eternal k-pebble game played on A and B, if both A and B satisfy γ_k. (Play the game! The description of γ_k as a bounded principle of plenitude is exactly what's required for the Duplicator's strategy.)

Let us call a sentence φ of first order logic *stochastically valid*, if and only if, $\Pr(\varphi) = 1$, and let Sval be the set of stochastically valid sentences of first order logic. It clear from the preceding discussion that $\Gamma = \{\gamma_k \mid k \geq 2\}$ axiomatizes a complete first-order theory, a result due to Gaifman (1964). In particular, Γ axiomatizes Sval. It follows at once that Sval is decidable. This provides an interesting contrast to the results described in Section 1.

Acknowledgements

The preparation of this work was supported in part by NSF CCR-9820899. I would like to thank the Graduate Program in Logic and Algorithms at the University of Athens for support while on leave from the University of Pennsylvania and for providing the stimulating research environment in which my work on this paper was completed. I am especially grateful to Steven Lindell for a decade of valuable discussions on the topics of this paper and to Mary-Angela Papalaskari for valuable comments on earlier drafts of this chapter.

References

Ajtai, M. and Fagin, R. (1990) Reachability is harder for directed than for undirected finite graphs. *Journal of Symbolic Logic*, 55, 113–50.

Alon, N. and Spencer, J. (1992) *The Probabilistic Method*. New York: John Wiley.

Barwise, J. (1977) On Moschovakis closure ordinals. *Journal of Symbolic Logic*, 42, 292–6.

Blass, A., Gurevich, Y. and Kozen, D. (1985) A zero-one law for logic with a fixed point operator. *Information and Control*, 67, 70–90.

Dawar, A. (1999) Finite models and finitely many variables. In D. Niwinski and R. Maron (eds.), *Logic, Algebra and Computer Science*, vol. 46 of *Banach Center Publications* (pp. 93–117). Polish Academy of Sciences.

Dawar, A., Lindell, S. and Weinstein, S. (1996) First order logic, fixed point logic, and linear order. In H. Kleine-Buening (ed.), *Computer Science Logic '95* (pp. 161–77). Berlin: Springer.

Dawar, A., Doets, D., Lindell, S. and Weinstein, S. (1998) Elementary properties of the finite ranks. *Mathematical Logic Quarterly*, 44, 349–53.

Ebbinghaus, H.-D. and Flum, J. (1999) *Finite Model Theory*. Berlin: Springer-Verlag.

Ehrenfeucht, A. (1961) An application of games to the completeness problem for formalized theories. *Fund. Math.*, 49, 129–41.

Fagin, R. (1974) Generalized first-order spectra and polynomial-time recognizable sets. In R. M. Karp (ed.), *Complexity of Computation, SIAM-AMS Proceedings*, vol. 7 (pp. 43–73).

Fagin, R. (1976) Probabilities on finite models. *Journal of Symbolic Logic*, 41(1), 50–8.

Fraïssé, R. (1954) Sur quelques classifications des systèmes de relations. *Publications Scientifiques de l'Université d'Algerie, Séries A*, 1, 35–182.

Frege, G. (1884) *Die Grundlagen der Arithmetik*. Breslau: Wilhelm Koebner.

Gaifman, H. (1964) Concerning measures in first-order calculi. *Israel Journal of Mathematics*, 2, 1–18.

Gaifman, H. and Vardi, M. (1985) A simple proof that connectivity of finite graphs is not first order. *Bulletin of the EATCS*, 43–5.

Glebskij, Y., Kogan, D., Liogon'kij, M. and Talanov, V. (1969) Range and degree of realizability of formulas in the restricted predicate calculus. *Cybernetics*, 5, 142–54.

Grohe, M. (1998) Finite variable logics in descriptive complexity theory. *Bulletin of Symbolic Logic*, 4, 345–98.

Gurevich, Y. (1984) Toward logic tailored for computational complexity. In M. Richter et al. (eds.), *Computation and Proof Theory* (pp. 175–216). Heidelberg: Springer-Verlag.

Gurevich, Y. (1988) Logic and the challenge of computer science. In E. Börger (ed.), *Current Trends in Theoretical computer Science* (pp. 1–57). Computer Science Press.

Gurevich, Y., Immerman, N. and Shelah, S. (1994) McColm's conjecture. In *Proceedings of the 9th IEEE Symposium on Logic in Computer Science*, pp. 10–19.

Immerman, N. (1986) Relational queries computable in polynomial time. *Information and Control*, 68, 86–104.

Immerman, N. (1999) *Descriptive Complexity*. New York: Springer-Verlag.

Kolaitis, P. G. and Vardi, M. Y. (1992a) Fixpoint logic vs. infinitary logic in finite-model theory. In *Proceedings of the 7th IEEE Symposium on Logic in Computer Science*, pp. 46–57.

Kolaitis, P. G. and Vardi, M. Y. (1992b) Infinitary logics and 0–1 laws. *Information and Computation*, 98(2), 258–94.

Moschovakis, Y. N. (1974) *Elementary Induction on Abstract Structures*. Amsterdam: North Holland.

Otto, M. (1997) *Bounded Variable Logics and Counting*. Berlin: Springer-Verlag.

Papadimitriou, C. (1994) *Computational Complexity*. Reading: Addison-Wesley.

Rosen, E. (1997) Modal logic over finite structures. *Journal of Logic, Language, and Information*, 6, 427–39.

Trakhtenbrot, B. A. (1950) Impossibility of an algorithm for the decision problem in finite classes. *Dokdaly Akademii Nauk SSSR*, 70, 569–72.

Vardi, M. Y. (1982) The complexity of relational query languages. In *Proceedings of the 14th ACM Symposium on the Theory of Computing*, pp. 137–146.

Further Reading

Two excellent texts are available which cover the topics presented here in depth. They are Ebbinghaus and Flum (1999) and Immerman (1999). An invaluable introduction to the theory of computational complexity is Papadimitriou (1994). For readers wishing further background on finite variable logics there are valuable survey articles by Dawar (1999) and Grohe (1998) and an excellent monograph by Otto (1997). An excellent introduction to the theory of random graphs is Alon and Spencer (1992).

Part VIII

LOGICAL FOUNDATIONS OF SET THEORY AND MATHEMATICS

This page intentionally left blank

23

Logic and Ontology: Numbers and Sets

JOSÉ A. BENARDETE

From the standpoint of philosophical logic, a great gulf separates elementary arithmetic, understood here as involving only the so-called adjectival use of numerals, from advanced arithmetic which features their substantival use, where the distinction turns on a point of grammar. Thus, '5 is odd' will count as a truth of advanced arithmetic, with the substantival expression '5' serving as a proper name that denotes a Platonic entity (never to be seen on land or sea). If advanced, Platonic arithmetic takes numerical sentences grammatically at face value, it is nominalistic, adjectival arithmetic that proves more grammatically devious, in subjecting numerical sentences to a reductive paraphrase via a detour through first-order predicate logic. Thus 'there are at least 2 (i.e. two) Fs' will be paraphrased as (1).

(1) $(\exists x)(\exists y)\,(Fx\ Fy\ \&\ \sim (x = y))$

For in (1) we are quantifying in Quine's jargon (via the existential quantifiers '$\exists x$' and '$\exists y$') only over Fs, for example dogs, and any putative reference to 2 as a Platonic object in our ontology is deftly conjured away, in accordance with his slogan 'Explication is elimination'. Because philosophy of mathematics today can only be described as being positively spooked by the neo-nominalist challenge – first implicitly posed in Benacerraf's 1973 "Mathematical Truth" (Benacerraf and Putnam 1983) and soon after, much more aggressively, implemented in Field (1980) – my own agenda can be expected never to stray very far from the specter of that challenge.

In the nominalistic vein of (1) we can even take the proto-equation 'Two and two make four' to say that if there are at least two Fs and at least two Gs (no F being a G), then there are at least four Hs (every F and G being an H). Typographically, distinguish now adjectival 'two' from substantival '2,' thereby being afforded the opportunity of registering '2 and 2 make 4' at face value as a truth of advanced arithmetic. One and the same unregimented English sentence, where '2' and 'two' are taken to be synonymous, is seen here to be ambiguous, needing to be disambiguated as between a nominalistic 'two' and a Platonistic '2.' Because the nominalistic version of the sentence can be displayed as a valid statement-form of first-order logic, Frege's program of reducing arithmetic to logic is thereby vindicated. But only in respect to elementary arithmetic (as herein defined). For the vernacular '5 (or five) is odd' remains irreducible to first-order logic.[1]

No accident surely that predicate logic is also styled as quantification theory, reminding us that in addition to the general quantifiers '∃' and '∀' we are free to recognize the following numerical quantifiers: 'there are at least (exactly) two (three, four . . .) Fs' where the identity predicate found in (1) is smuggled in as a constituent of these complex quantifiers. No more than an advertising trick of relabeling, these ostensibly new quantifiers are already available for free in standard first-order predicate logic with identity. Recalling one traditional definition of mathematics as the science of quantity, we trust that more than a mere verbal trick is involved in now undertaking to recycle Frege's logistic thesis by assimilating mathematics as so defined to predicate logic characterized as quantification theory.

1 Sher's Weak Logicism

Pursuing that suggestion, one may even dare to emulate Gila Sher (1991) by enriching standard first-order logic with such adjectival yet Cantorian quantifiers as 'there are uncountably many x,' thereby inviting all of Cantor's alephs into 'logic.' Not that any pretense of reducing those alephs to logic along properly Fregean lines can be expected here, seeing that they will be supplied outright by Zermelo-Fraenkel set theory, taken to be coeval with logic itself. Even so, a convincing, recognizably Fregean case can be made for allowing the Cantorian quantifiers (as embedded in an extended first-order logic) to be certified as logical constants, thereby recasting set theory itself in terms of a weak logicism. Thus Sher writes, "Frege construed the existential and universal quantifiers as second-level quantitative properties that hold (or do not hold) for a first-level property in their range due to the size of its extension" (Sher 1991: 10). We may then suppose that the second-level property expressed by the Cantorian quantifier 'there are aleph-50 x' will fail to hold of the property expressed by the first-level predicate 'x is a dog' if only because the suggestion of there being aleph-50 dogs can only strike standard set theory as being impossibly droll, smacking of a category-mistake. Not at all the sort of scenario one routinely envisages.

Moreover, let all of space–time be packed solid with dogs cheek by jowl; that will yield no more than aleph-zero of them. A disappointing result really, for our program. In the general case one wants to say that for any F (absent information to the contrary) it is an open question how many Fs there are; and with dogs being the very sort of thing paradigmatic of what the question 'how many?' addresses, all of the cardinal numbers, transfinite as well as finite, ought to be available to draw upon. That at any rate is a new slant on set theory, viewed precisely as the general theory of cardinality, that is activated by Benacerraf's challenge, under the aegis of which the nominalist is free to argue that ruling out aleph-50 dogs is tantamount to ruling out aleph-50 itself as a genuine cardinal. No genuine cardinal, then no chance of figuring as a logical constant in its capacity of being a genuine quantifier.

If Sher's program draws on A. Mostowski's seminal (1957) "On a Generalization of Quantifiers," the latter in its turn draws on Cantor's generalization of the concept of (finite) cardinality, both of which can only prove the more nominalistically attractive if aleph-50 dogs were to be admitted as a serious option. The immediate obstacle lay in the exiguous accommodations that all of (our) space–time affords to dogs, combined

with the tacit insistence (reminiscent of Kant) that there can only be one Space. Long out of favor, Kant's synthetic *a priori* has recently been making a modest comeback (notably in van Cleve 1999 but see also Tennant 1996), and it may now be invoked in support of the One Space thesis. Nominalists, however, will be least inclined to defer to it, preferring to canvass the copious plurality of worlds of Lewis (1986) as well as Everett's 'many worlds' hypothesis in quantum physics. In rejecting any aprioristic constraint on the cardinality of dogs, one must recognize (according to the official semantics anyway) that simply to say 'There are aleph-50 dogs' is to say that the *set* of dogs is equinumerous (via one-to-one correspondence) with . . . Assume here the simplest case where the (generalized) Continuum Hypothesis is true. Then the net result of filling the gap – with the words 'the power-set of the power-set of . . . the natural numbers' – will be that even nominalists can endorse the following argument A as valid: 'there are aleph-50 dogs, therefore there are at least aleph-50 Platonic objects,' sticking with our standard semantics. Assume, however, with Field (1993) that the (non)existence of sets is a contingent matter, meaning that given any set of dogs those dogs can jointly exist in the absence of all sets. Pretend now that the premise of A is true, and focus on the very dogs themselves, ignoring their cardinality. Following Field, those dogs will be found in a possible world where there are no sets, and there will even be aleph-fifty dogs there, seeing that for each dog here there will be exactly one dog there (indeed the same dog).

Not substantival 'aleph-50' then but rather an adjectival (and nominalistic) 'aleph-fifty' emerges as a logical constant (and transfinite quantifier) in a Sherian extension of first-order logic that a Fieldian nominalist can accept. Further warrant for speculations in that vein will be found in the megethology of Lewis (1998: 203–29) where a nominalistic universe is envisaged with quite as many entities as ZF supplies, doubtless to be sought in his plurality of worlds.

2 Finiteness, an Infinite Sentence and Skolem

Recoiling from these excesses, one may well wish to stick with first-order logic plain and simple, though (in the absence of both set theory and substantival arithmetic) Shapiro (1991: 9) indicates how one will then lack even the mere means to say that there are only finitely many dogs, or worse still (since it threatens first-order logic itself) that every first-order sentence consists of only finitely many expressions, thereby in effect joining Field in his recent "doubts about the determinacy of the notion of finiteness" (Schirn 1998: 99). It is here above all, with the finite itself, that nominalistic doubts about numbers and sets trickle all the way down to logic. Encouraged, however, by at least two linguists (Langendoen and Postal 1984), who insist that the grammar and syntax of ordinary language allow for sentences of (any arbitrary finite or) transfinite length, one can always liberalize our standard first-order logic and (try to) say in an infinitary notation, "There are exactly one or two or three or . . . dogs," in the adjectival mode of (1).[2]

How our failure to complete the sentence (whose length is presumed to be $\omega + 5$) might bear on (the constraints of) logic, may prove (a little) less obscure in the light of a marvelous exchange in Shapiro (1991: 206) where in reply to "What I mean by

'natural number' is 'member of the sequence 0, 1, 2, 3 . . .'," a Skolemite skeptic queries the meaning (and use) here of the dots . . . , along the lines of the following, simpler scenario of mine. Posit an infinite sequence of men, S, with a first, second, . . . where the dots can be shown to harbor a genuine indeterminacy which – to our astonishment – Skolem reads back into the natural numbers themselves. Naively, the two sequences, that is N and S, are on a par but the second allows for a man in the (the?) infinite sequence who is separated from the first man by infinitely many intermediate men. Start with a shortest man 6 ft. tall launching an infinite progression with each man taller than his predecessor yet with none reaching 7 ft. Here then is one infinite sequence, S_1, whose ordinality is ω. Add to it a 7-ft man, yielding a second sequence, S_2, of order-type $\omega + 1$. Providing for S_1, do the dots in my scenario of S also extend to our last man in S_2? Although the answer is doubtless no when it comes to the speech-act pragmatics of most occasions where S_1 will supply our standard or intended model of S, I take Skolem to be saying that as to mere semantics '. . .' as it figures in my scenario is infected with indeterminacy as between S_1 and S_2.

Because the implicit exposure of most people to Peano's (five) postulates for arithmetic extends only to the first four, which allow for just the sort of nonstandard model as my scenario (hence the need for Peano's fifth postulate), it can be a great mystery how the vulgar ever do acquire the concept of a natural number. Another, more philosophically urgent puzzle turns on how *we* succeed in doing so, even as favored with the fifth postulate of mathematical induction that – according to Skolem's devious argument, as richly discussed both in Shapiro (1991) and in Lavine (1994) – also fails to secure determinacy for it. Reminiscent of Kripgenstein's 'quus' paradox with its skepticism (Kripke 1982) as to how '+' can signify addition rather than quaddition, Skolem's regarding '. . .' will probably be fully resolved only after Kripgenstein's much more general worries about meaning and reference have been appeased, not to mention Putnam's (1977) "Skolemization of absolutely everything" (Putnam 1983: 15). If the most intriguing response to Kripgenstein lies in the 'saving constraint' of objective similarity out in the world that is "not of our own making," invoked in Lewis (1999: 45–55, 63–7) in order to fix content, one can at least see how it bears on Skolem, for '. . .' is doubtless synonymous with 'etc.' which in its turn, just means 'and (all) the others (of the same sort)' where what is to count as the same sort of like items needs to be pinned down. No surprise surely if Tennant's Schema C in the next section, which undertakes to ground the natural numbers in logic itself, should come to supply a piece in the puzzle.

3 Back to Strong Logicism?

Highly controversial, Crispin Wright's (1983) reactivation of Frege's logistic program, which for decades just about everyone assumed to be a lost cause, has forced researchers to rethink some of the more fundamental issues in logic. Benacerraf himself in a retrospective look at responses to his 1973 challenge regards Wright's 1983 as "the only line of inquiry that seems at all sensitive to arithmetical practice" (in Schirn 1998: 57). Neatly sidestepping Russell's Paradox of 1902 which ditched Frege's mature program, Wright reverts to an earlier version that features HP (Hume's

Principle) where the functional expression 'the number of x such that x is F' is abbreviated by 'Nx:Fx'.

(HP) Nx:Fx = Nx:Gx iff the Fs can be placed in one-to-one correspondence with the Gs.

Better still, Tennant (1997a: 310) features the simplified Schema C which directly effects an a priori synthesis of adjectival with substantival (finite) arithmetic.

(C) There are n Fs iff Nx:Fx = *n*

where '*n*' on the right indicates a numeral and 'n' on the left indicates its adjectival correlate which is to be unpacked as in (1). Taking HP and C to be analytic propositions, it will be easy now with either to produce an a priori proof of (the existence of) the natural numbers. Thus using C to derive '~(∃x) ~ (x = x) iff Nx:~(x = x) = 0,' we find that 0 emerges as the number of things which fail to be identical with themselves, while 1 emerges as Nx:(x = 0) and 2 as Nx:(x = 0 ∨ x = 1), etc.

Being analytic, won't all of these propositions, for example '(∃x) (x = 9),' be on a par with 'All bachelors are unmarried' and hence merely verbal truths that can tell us nothing about the world, being true solely by convention? Prompted by early Wittgenstein and dominant in the 1930s, the Conventionalist doctrine of analyticity, which was widely used to deflate Frege's program, has long been absent from contemporary discussions. Nor have I seen any sign of its being revived in response to Wright, largely (I suppose) thanks to Tarskian semantics, with an assist from Davidson (1967a), for whom to grasp the meaning of a sentence is to grasp its truth condition in terms of a Tarskian biconditional. Thus to grasp the meaning of the sentence 'All unmarried men are unmarried' is just to recognize that the sentence is true iff all unmarried men are unmarried, where our attention, initially fixed on a sentence, is guided away from it to the world, maybe even to the state of affairs of each unmarried man's being unmarried, though the genius of Tarski lay in declining to reify states of affairs. The mere availability of that option, however, has been widely felt to dispel Conventionalism. More characteristic of current controversy over neologicism – ranged on one side are Wright, Hale (1988) and Tennant, on the other Dummett (1991), Boolos (1998) and Field – is Field (writing in 1984), "I don't see how the existence of objects of any sort [e.g. numbers] can follow logically from the existence of objects of an entirely different sort [e.g. planets]" (Field 1989: 166) as in 'There are nine planets. ∴ The number of planets is 9' where our typographical innovation alerts us to a difficulty that goes unnoticed in the vernacular when premise and conclusion are seen as broadly synonymous.

Despite being widely though perhaps only subliminally shared, Field's worry detracted very little from the acclaim Davidson enjoyed when in "The Logical Form of Action Sentences" (1967b), in a comparable case, he urged that the following should count as a valid argument, to be certified as such by first-order logic: 'Tom is walking slowly. ∴ (∃x) x is a walking & x is slow' (Davidson 1982: 105–48) even though the premise was standardly supposed, by Quine and Co., to involve 'ontological commitment' only to Tom, while the conclusion was widely held at the time to feature a very

dubious sort of entity, namely an event. Assume, however, that real logic is just first-order logic (a view still very much in fashion), and try formalizing the valid argument 'Tom is walking slowly. ∴ Tom is walking' (where the adverb proves recalcitrant) without positing in the premise (the event of) Tom's walking.

Still another example of Frege's legerdemain of a priori synthesis, which cannot then fail to smack of Kant's synthetic a priori, is found in Armstrong's aprioristic appeal to the principal Truthmaker – every true statement is made true by some object(s) – whereby the argument 'Socrates is wise. ∴ There is at least one state of affairs (of Socrates being wise)' emerges as valid (Armstrong 1997: 115). Frege, Davidson, Armstrong: these eminent instances of (what one might pejoratively call) philosophical logic invite Field to reply, "Plain, flat-footed logic is good enough for me." If in connection with the former sort of logic my harking back to the synthetic a priori will strike some readers as quaintly anachronistic, two considerations may appease them. First, the more general. Frege invents modern logic for one purpose only, namely to prove in detail the analyticity of arithmetic, as against Kant's synthetic a priori construal of it that may now be feared to infect HP and Schema C in a return of the repressed. Even assuming their analyticity, however, Tennant (1997b: 293–4) rightly senses a difficulty in Gödel's true but unprovable sentence of arithmetic that might then be urged to have a synthetic a priori status by way of contrast.

My second consideration bears directly on HP which Hale (1994: 124) urges us to view in terms of "a broader conception of analyticity that covers such cases" as 'Nothing is both red and green all over,' a proposition much contested over the years that van Cleve (1999: 226–9) joins a distinguished tradition in defending as paradigmatic of the synthetic *a priori*. Resolving this dispute between Hale and van Cleve offers the best researchable prospects today for assessing neologicism.

4 Benacerraf's Challenge

By invoking Davidson's events and Armstrong's states of affairs – both taken to be concrete entities – as foils for understanding Frege's abstract objects, we are given a window of opportunity for meeting Benacerraf's challenge. Recall the truffle in Hermione's hand that causes her to believe in its existence, catering thereby to the recent shift in epistemology, away from the traditional, internalist emphasis on evidentialism (the weighing of evidence) and toward the new externalist focus on reliabilism, with underlying belief-forming mechanisms that track the truth. Platonic objects in their causal impotence prove thus to be at a distinct naturalistic disadvantage when compared with the true-belief-inducing efficacy of perceived truffles. An unfair contrast when it comes to playing the ontology game! Contrast rather Frege in his ontology quantifying over 0 with Armstrong quantifying in his over the state of affairs of a truffle in his hand being seen by him, altogether waiving the success or failure of either's philosophical line of argument. Although Armstrong's everyday true belief in the truffle is caused by it, no such reliable mechanism is naturalistically credible when he undertakes to posit in his armchair ontology such *recherché* entities as his universals and states of affairs. Think here of Hume: the belief in truffles is caused by force of nature or habit. Not so with more rarefied speculations where ontologists go different ways. Let Armstrong's

universals be allowed to exist. No matter. They play no role, externalist or internalist, in explaining why Armstrong does and Quine does not believe in them.

Actually, truffles themselves lose their innocence on being co-opted into (or banished from) the ontology game. After eliminating all Platonic objects, our nominalist may become emboldened to wield Ockham's razor afresh, now eliminating even truffles in the course of quantifying only over the elementary particles of physics. So our belief in truffles may not really be caused by them, but by elementary particles suitably arranged (to mimic truffles)? If one recoils from this suggestion mereology may be invited to kick in, on the ground that "mereological wholes" like truffles, "are not ontologically additional to all their [ultimate] parts," namely the particles, being rather "*identical* with all their parts taken together" (emphasis in original). Given the particles then, the truffles are supplied by way of an "ontological free lunch" which "like other such lunches . . . gives and takes away at the same time. You get the supervenient for free, but you do not really get an extra entity" (Armstrong 1997: 11–12). No extra entity? So $(\forall x)$ (x is an elementary particle)? Even though $(\exists y)$ (y is a truffle) & ~ $(\exists z)$ (z is a truffle & z is a particle)? That Armstrong's legerdemain here may be all of apiece with Frege's, one is encouraged to believe when Tennant, affecting equal facetiousness, appeals to his Schema C by way of "getting something for nothing" (Tennant 1997b: 322), and the comparison between the two cases may even dispel the invidious bugbear of Platonism that attaches to one of them. Not that an alternative model for understanding neologicism cannot be found in Armstrong's case for states of affairs, seeing that his independent appeal to Truthmaker – arguably playing the role of Schema C – is not taken by him to involve any ontological free lunch.

5 An Anti-realist Frege?

If sets have been seen as being constituted by their members (Parsons 1983: 217, 275 and 286), truffles have been taken to be constituted by elementary particles, and in both cases the 'constitution' relation may be viewed either in a realist mode as being metaphysically deep or in an anti-realist one as smacking of eliminativism. In this scheme the causal impotence of the one sort of item and the (putative) causal efficacy of the other may come to play very little role in any final reckoning.

In a somewhat different vein Dummett adjudges Wright's neologicism to be a success but only if it is viewed in terms of an anti-realist Frege of 1884 in *Grundlagen*. Not to be confused with the realist Frege of 1893 in *Grundgesetze* who profits by his interim discovery of the sense/reference distinction. The shift turns on the difference between a thin (anti-realist) and a thick (realist) notion of reference where the former is content to view "any legitimate question about the meaning of a term, that is, about what we should call its reference [as being] reducible to a question about the truth or otherwise of some sentence in the language" (Dummett 1991: 192). Thus 'the number of planets' will (trivially) succeed in denoting an object if 'the number of planets = the number of Jones's fingers' is true, while (the really important point) the non-trivial truth-condition of the sentence will be reductively satisfied just in case there are (exactly) nine planets and fingers. As of 1894, however, compositionality will decree that a whole sentence can have a sense (and reference) only if each of its (unitary) parts

antecedently has a sense (and reference), as when 'that truffle' embedded in 'that truffle is white' comes to enjoy thick reference thanks to Benacerref's causal link.

Faithful then to the earlier anti-realist Frege, Wright's neologicism is deemed by Dummett to fail in its more ambitious goal of achieving thick reference for numerals. What Dummett forgets to mention, however, is his amazing discussion (Dummett 1973: 503) where a Kantian Frege is envisaged even as to 1893, for whom arguably all reference proves in the end to be thin.

> Our ability to discriminate, within reality, objects of any particular kind results from our having learned to use expressions, names or general terms, with which are associated a criterion of identity. . . . [W]e can in principle, conceive of a language containing names and general terms with which significantly different criteria of identity were associated, and the speakers of such a language would view the world as falling apart into discrete objects in a different way from ourselves. . . . [F]or Frege . . . it is we who . . . impose a structure on [the world]. (Dummett 1973: 503)

How to square this Fregean anti-realism with our radical Ockhamist who, in refusing to quantify over macro objects, allows only thin reference to 'that truffle' as it figures in the true sentence 'That truffle is white,' should not be very difficult. One has only to view the radical Ockhamist as clearing the ground for the still more radical position of Putnam's conceptual relativism (Putnam 1988: 113–16).

6 Second-order Logic and Sets

If an anti-realist neologicism affords one option, a realist version emerged in my proposal to assimilate Tennant's Schema C to Armstrong's Truthmaker as, in the one case, numbers and in the other case states of affairs are admitted to a realist ontology. As to sets in particular, it is so-called second-order logic, which Quine famously distinguishes from logic proper, that ostensibly treats 'x ε y' as a(nother) logical predicate along with 'x = y'. Take the second-order sentence '(∃F) (Socrates is F)' which is routinely read either as 'there is a property Fness that Socrates has' or as 'there is a set of which Socrates is a member.' Frege, however, gives it a special twist: 'There is something, for example wise, that Socrates is' as in 'Wise is something that Socrates is which I am not,' indicating how the first-order '(∃x) (∃y) (x = y)' supplies only one way of disambiguating 'Something is something,' for there is also the second-order reading, 'Something (e.g. Socrates) is something (e.g. wise)' where a first-order 'something' is followed by a second-order 'something.' 'Wise is something' is thus complete on one reading but incomplete on the other when 'Wise is something (that Socrates is)' fails to yield a complete sentence (as truncated). Hence Frege's *outré* incomplete entities or quasi-entities (his concepts and functions) none of which can be given a proper name.[3]

Psychologically at least, one can understand how the conceptual jump in Schema C from non-objectual (adjectival) arithmetic to the objectual (substantival) variety might cease to strike Frege as being objectionably abrupt only after he came to finesse the passage by positing the *tertium quid* of his (arguably) Platonistic quasi-entities; and I can well imagine efforts today along this line to narrow if not perhaps quite close the

gap, for example Hodes (1990: 255). I leave it as an open question whether accepting second-order logic in terms of Frege's second-order 'something' should persuade one to regard Quine's criterion of ontological commitment, enshrined in our first-order 'something,' as being unduly restrictive. In effect, then, I am asking the nominalist if he can live with Fregean second-order logic, being quite prepared to find that nominalists may divide on this issue owing to an inherent indeterminacy that infects our notion of ontological commitment when it comes to Fregean concepts and functions.

More encouraging may then be felt to be mereology whereby '(∃F) Socrates is F' receives this fourth, expressly nominalistic reading, 'There is something (a mereological whole) of which Socrates is a part.' Virtually patented in Boolos (1984, 1985), the new device of plural quantification supplies a fifth, nominalistic reading, 'There are some things such that Socrates is one of them' where these last two readings (I verily believe) echo in a deep way Frege's second-order 'something.' How thanks to plural quantification (2) can clarify in particular P(N), that is the problematic power-set of the natural numbers, will emerge in the sequel.

(2) There are some natural numbers that no set has as its only members.

Mystifying? Not (a trivial case) if one's set theory allows only finite sets. The trick will be to see how (some) constructivists might entertain the truth of (2) even as regards the infinite subsets of N where at least nominally P(N) can still exist with all of the subsets of N, that is all which are left over after (2) has kicked in. Because not only Sher (1991) but also Shapiro (1991) is engaged in investing set theory with a logical or quasi-logical status, the one in a first-order framework, the other in a second-order framework, the familiar reservations of constructivists pose a threat to both programs that (2) can hardly fail to illuminate.

Not to be confused with Quine, however, who convicts second-order logic of just being "set theory in disguise," Shapiro sharply distinguishes his own, logical conception of set from the standard, non-logical, iterative conception of ZF where only the former is to be equated with second-order logic. Always relative to some restricted universe of discourse, for example N, the first-order variables of Shapiro will range over the objects in the domain, while the second-order variables will range over their various subsets, reminding us of Tarski's model theory. If set theory proper is dizzyingly vertical, the logical conception of set is seen to be manageably horizontal, with a set of all non-Fs, for example non-dogs, being allowed only in the latter, Boolean system which smacks more of Aristotle than of Frege. Why the manageable system cannot really be insulated from the high tides of ZF, becomes clear when one asks after the cardinality of P(N) now that the Continuum Hypothesis has been shown, by Paul J. Cohen, to be neither provable nor disprovable in ZFC. Arguably more subversive than even the constructivist challenge to classical mathematics, there is a widespread failure of nerve today as to whether C.H. can be said to have a determinate truth value at all, in defiance of Excluded Middle that insists on it.

As to how this failure of nerve might threaten Shapiro's logical conception of set, a cheeky proposal by Field I find to be especially suggestive. Cheeky, I say, because as a nominalist he might be expected to relish the spectacle of set theory imploding from

within; and one can only query his good faith here in offering guidance to Platonists trying to cope with (just about) any Cantorian aleph being eligible, as a point of mere logic, to fix the cardinality of P(N). Elaborating on the distinction between constructible (rule-governed) and non-constructible (random) sets, Field would have us recognize the term 'set' as systematically ambiguous, depending on whether we are referring to $sets_0$, $sets_1$, $sets_2$ . . . all the way up the hierarchy of ZF ordinals (Field 1994). When it comes to $sets_0$ then C.H. will be true and the cardinality of P(N) will be aleph-one. In all other cases C.H. will be false, and the cardinality of P(N) will be, say, aleph-46 when it comes to $sets_{45}$.

If one's first reaction to Field's proposal will probably view it as being bent on trashing set theory, an instructive byproduct of his suggestion is that (2) taken now as a schema emerges as true whenever the vague term 'set' is replaced by any one of his subscripted precisifications of it. And Field aside, an irenic constructivist who recognizes that (thanks to Russell's Paradox) there are already some things that no set has its only members might allow in the same vein for there being some numbers, for example 8, 86, 862, 8625, . . . that by corresponding to the decimal expansion of a putative irrational number which no (finite) rule can generate, that is, 8625 . . . , are thereby disqualified from being the sole members of any set. No less accommodating, a moderate classicist might now be prepared to retreat from his espousal of non-constructible sets to a backup position supplied by (2) if only to stake out common ground in what has otherwise been a longstanding deadlock. The key point, of course, is that what really matters here is that some numbers do exist (or might, for all we know, exist) thanks to which (2) is or might be true, and that the putative existence of a set that gathers them up (or does whatever sets are supposed to do in respect of their members) can only play a secondary role. Because this suggestion as to 'what really matters' relies on Boolos's device of plural quantification, label it P.Q. for future reference.

Field's proposal takes on new importance when it is viewed in the light of Hallett's remarkable work (1984: 208) where, following Paul J. Cohen himself, he writes, "[W]e have no positive reason to assume that even only one application of the power-set axiom to an infinite set will not exhaust the whole universe," which is as much as to say that P(N) may not be a ZF set at all but rather a proper class of von Neumann. What *may* be the case for Cohen, Field in effect takes really to be the case, and thereby daringly proposes by means of philosophy alone to answer what is still standardly taken to be an open question in mathematics proper, namely 'what is the cardinality of P(N)?' For all of Field's subscripted subsets of N will be seen by any true classicist to be proper *subsets* of N, and hence no aleph will suffice to fix how "incredibly rich" (Cohen) is P(N), precisely the "point of view" here which Cohen as of 1966 "feels may eventually come to be accepted."

7 Skolem (Again) and Megethology

"The proponents of second-order logic *as logic*", Shapiro writes by way of exorcising Skolem (1991: 204), "hold that second-order terminology . . . is sufficiently clear, intuitive, or unproblematic. . . . The claim is that once a domain (for the first-order

variables) is fixed, there is a reasonably clear and unambiguous understanding of such locutions as . . . 'all subsets' thereof." Unambiguous? Field we find, in coping with P(N) and C.H., denies that, with his radical disambiguation of the term 'set.' But let that pass and focus on our "reasonably clear understanding" of the locution 'all subsets of N.' Suppose that Jones (standing in for the constructivist) insists on each set's having only finitely many members. Does he mean something different from the rest of us by the expression 'all the subsets of the natural numbers'? Although many will be strongly inclined to say yes, I doubt if they are fully registering my query as it is intended, with heavy emphasis on the whole notion of meaning in analytical philosophy. For we can hardly expect to bring Jones into line by saying, "So you're thinking of (river) banks while we're talking about (money) banks," seeing that – to pursue the analogy – he denies that the term 'bank' as we use it, that is 'set' or 'non-constructible set,' can have a nonempty extension.

Why not then suspend judgment as to consructivist demands, allowing that (2) might well be true owing to P.Q.? For we can continue to believe in the existence of (whatever should turn out to be) *all* the subsets of N. No longer indeed remaining faithful to Shapiro's "working realism" with its injunction to take classical mathematics at face value, and there may now even be reason to fear being led down a slippery slope to Skolemism where what counts as 'all the subsets of N' is always relative to this or that frame of reference. Recall that Peano's fifth postulate of mathematical induction was presciently designed (really by Dedekind) to rule out Skolem's nonstandard models of N precisely by quantifying over *all the subsets of N*. So Skolem may then be vindicated in either one of two ways: (a) Field's insistence on disambiguating 'set' prompted by C.H. and (b) worries about (2) having to do with constructivist scruples.[4]

Not for the first time do I now wish to show how, here in connection with our uneasiness over (2), nominalistic considerations can be pressed into the service of classical set theory itself. Returning then to mereology but not the free-lunch version of Armstrong that undertakes to co-opt 'x is a part of y' as a topic-neutral logical constant, I rely rather on a very topic-specific version of it that, in drawing on the familiar distinction between things and the stuff of which they are composed, restores mereology to its proper home, namely Helen Cartwright's quantities. Pour some water into the Atlantic ocean. Fifty years later someone might scoop it up from the Pacific. In the interim that very (quantity of) water will be sloshing about in a very scattered form. Pretend that this 'water' is composed of Democritean atoms made of adamant where, in a world with a simple infinity of them, there will be a distinct quantity of adamant corresponding to each classical subset (which in turn corresponds to each mereological sum) of those atoms. Mereology and (quantities of) stuff thus appear to be tailor-made for each other, which explains why the negation of (3), which has been designed as a mereological counterpart of (2), can strike us as an analytic proposition, featuring both nominalistic devices, namely plural quantifications as well as mereology, that are systematically combined to yield the megethology of Lewis's thesis "Mathematics is megethology" (Lewis 1993: 3–23).

(3) There are some Democritean atoms the adamant of which fails to comprise some adamant.

Compare: there are some wooden houses the wood of which (they are composed) fails to comprise some (quantity of) wood. Here if anywhere mereology supplies an ontological free lunch, and P(N) can thus be envisaged in an entirely nominalistic realization yet free of all constructivist restrictions, thereby running afoul of Feferman (1998) that takes his constructivist program to be mandated by his rejection of set-theoretical Platonism. To the contrary, nominalists are free to divide, some opting for a classical, non-constructivist mereology, while others insist on a restrictive, constructivist abridgment of it. Platonism proves then to be largely a red herring when it comes to the key issue. How the debate (over constructivist vs. non-constructivist mereology) will play out in the coming decade, one can only await in suspense.

Because my Democritean scenario takes us beyond the actual world (recalling in this respect my conceit as to aleph-50 dogs), it does look as if my gropings in this chapter toward a replay of (something like) Frege's logicism will have to look to Hellman (1989) and Chihara (1990) for expressly modal foundations of mathematics. To the question whether the modal operators 'it is possible (necessary) that' should count as logical constants, may then be added queries as to 'x is a part of y,' 'x ∈ y,' 'x is one of the ys,' and even 'x = y.' How logic itself is finally to be characterized, can be expected to accommodate various considerations raised in this chapter, having to do with numbers and sets.

Notes

1 Because 5 is defined in the Frege–Russell program as the set of all sets with exactly five members, a reduction of '5' to 'five' is effected but only via the Platonic objects of set theory. After Russell's Paradox ZF set theory rules out any such set as being 'too large.' Acceptable as a proper class, 5 in that role is debarred from being a member of any set or class. No set or class then of natural numbers, spelling defeat of the program.

2 Responding in 1931 to Gödel's true but (finitely) unprovable sentence, "Zermelo went on to propose a massive infinitary logic" along this line, writes Shapiro (1991: 191), where "Zermelo argued that Gödel's reasoning shows that any finitary notion of 'proof' is inadequate." Zermelo's option remains problematically open to this day.

3 Although Dummett (1973: 243) concludes his ch. 7 on incomplete expressions by finding Frege's position to be "in the end unjustified," in his later book (1981: 164), he reopens the issue where "this conclusion now seems to be too strong."

4 In an especially perspicuous defense of Skolemism (as against Benacerraf), Wright remarks that "there are, I suppose, two routes into the informal notion of a subset of a given set. . . . Neither of these routes, it seems to me, holds any very plausible promise of meeting the Cantorian's needs" (Benacerraf and Wright 1985: 135). If Wright now has grounds for resisting (2), I have yet to learn of them, though I would not be surprised to meet a Skolemite who felt threatened by (2).

References

Armstrong, David (1997) *A World of States of Affairs*. Cambridge: Cambridge University Press.

Benacerraf, Paul and Putnam, Hilary (eds.) (1983) *Philosophy of Mathematics: Selected Readings*, 2nd edn. Cambridge: Cambridge University Press.

Benacerraf, Paul and Wright, Crispin (1985) Skolem and the skeptic. *Proceedings of the Aristotelian Society*, supplementary vol., 59, 85–137.

Boolos, George (1984) To be is to be a value of a variable (or to be some values of some variables). *The Journal of Philosophy*, 430–49, and reprinted in Boolos (1998).

Boolos, George (1985) Nominalist platonism. *The Philosophical Review*, 94, 327–44, and reprinted in Boolos (1998).

Boolos, George (1998) *Logic, Logic, and Logic*. Cambridge, MA: Harvard University Press.

Chihara, Charles S. (1990) *Constructibility and Mathematical Existence*. Oxford: Clarendon Press.

Cohen, Paul J. (1966) *Set Theory and the Continuum Hypothesis*. New York: Benjamin.

Davidson, Donald (1967a) Truth and meaning. *Synthèse*, 17, 304–23, and reprinted in Davidson, Donald (1984) *Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, Donald (1967b) The logical form of action sentences. In N. Rescher (ed.), *The Logic of Decision and Action*. Pittsburgh: University of Pittsburgh Press and reprinted in Davidson (1982).

Davidson, Donald (1982) *Essays on Actions and Events*. Oxford: Clarendon Press.

Dummett, Michael (1973) *Frege: Philosophy of Language*. Cambridge, MA: Harvard University Press.

Dummett, Michael (1981) *The Interpretation of Frege's Philosophy*. Cambridge, MA: Harvard University Press.

Dummett, Michael (1991) *Frege: Philosophy of Mathematics*. Cambridge, MA: Harvard University Press.

Feferman, Solomon (1998) *In the Light of Logic*. Oxford: Oxford University Press.

Field, Hartry (1980) *Science without Numbers: A Defense of Nominalism*. Princeton, NJ: Princeton University Press.

Field, Hartry (1984) Critical notice of Crispin Wright: Frege's conception of numbers as objects. *Canadian Journal of Philosophy*, 14, 637–62, and reprinted in Field (1989) as Ch. 5, retitled as "Platonism for Cheap? Crispin Wright on Frege's Context Principle."

Field, Hartry (1989) *Realism, Mathematics and Modality*. Oxford: Basil Blackwell.

Field, Hartry (1993) The conceptual contingency of mathematical objects. *Mind* 102, 285–99.

Field, Hartry (1994) Are our logical and mathematical concepts highly determinate? *Midwest Studies in Philosophy*, 19, French, Peter A., Uehling, Theodore E. and Wettstein, Howard K. (eds.), 391–429 and reprinted (slightly shortened) in Schirn (1998), retitled as "Do we have a determinate conception of finiteness and natural number?"

Hale, Bob (1988) *Abstract Objects*. Oxford: Basil Blacwell.

Hale, Bob (1994) Dummett's critique of Wright's attempt to resuscitate Frege. *Philosophia Mathematica*, vol. 2, 122–47.

Hallett, Michael (1984) *Cantorian Set Theory and Limitation of Size*. Oxford: Clarendon Press.

Hellman, Geoffrey (1989) *Mathematics without Numbers: Toward a Modal-Structural Interpretation*. Oxford: Clarendon Press.

Hodes, Harold T. (1990) Ontological commitment thick and thin. In G. Boolos (ed.) *Meaning and Method*. (pp. 236–59). Cambridge: Cambridge University Press.

Kripke, Saul A. (1982) *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Langendoen, D. Terence and Postal, Paul M. (1984) *The Vastness of Natural Languages*. Oxford: Basil Blackwell.

Lavine, Shaughan (1994) *Understanding the Infinite*. Cambridge, MA: Harvard University Press.

Lewis, David (1986) *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, David (1993) Mathematics is megathology. *Philosophia Mathematica*, 3, 3–23; rpt. in Lewis (1998).

Lewis, David (1998) *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.
Mostowski, A. (1957) On a generalization of quantifiers. *Fundamenta Mathematicae*, 44, 12–36.
Parsons, Charles (1983) *Mathematics in Philosophy: Selected Essays*. Ithaca, NY: Cornell University Press.
Putnam, Hilary (1977) Models and reality. In Putnam (1983) 1–25.
Putnam, Hilary (1983) *Philosophical Papers*, vol. 3. Cambridge: Cambridge University Press.
Putnam, Hilary (1988) *Representation and Reality*. Cambridge, MA: MIT Press.
Schirn, Matthias (ed.) (1998) *The Philosophy of Mathematics Today*. Oxford: Clarendon Press.
Shapiro, Stewart (1991) *Foundations without Foundationalism: A Case for Second-Order Logic*. Oxford: Clarendon Press.
Sher, Gila (1991) *The Bounds of Sense*. Cambridge, MA: MIT Press.
Tennant, Neil (1996) The law of excluded middle is synthetic *a priori*, if valid. *Philosophical Topics*, 24, 205–29.
Tennant, Neil (1997a) On the necessary existence of numbers. *Noûs*, 31, 307–36.
Tennant, Neil (1997b) *The Taming of the True*. Oxford: Clarendon Press.
Van Cleve, James (1999) *Problems in Kant*. New York: Oxford University Press.
Wright, Crispin (1983) *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.

Further Reading

Azzouni, Jody (1994) *Metaphysical Myths, Mathematical Practice: The Ontology and Epistemology of the Exact Sciences*. Cambridge: Cambridge University Press.
Benacerraf, Paul (1973) Mathematical truth. *Journal of Philosophy*, 70, 661–80.
Benacerraf, Paul (1998) What mathematical truth could not be. In Schirn, 33–75.
Boolos, George (1997) Is Hume's principle analytic? In Heck (1997) and in Boolos (1998).
Davies, Martin (1981) *Meaning, Quantification and Necessity*. London: Routledge & Kegan Paul.
Dummett, Michael (1998) Neo-Fregeans: in bad company? In Schirn (1998) 369–87.
Field, Hartry (1998) Do we have a determinate conception of finiteness and natural number? In Schirn (1998) 99–129.
Hazen, A. P. (1993) Against pluralism. *Australasian Journal of Philosophy*, 71, 132–44.
Heck, Richard G. (ed.) (1997) *Logic, Language, and Thought*. Oxford: Oxford University Press.
Hughes, R. I. G. (ed.) (1993) *A Philosophical Companion to First-Order Logic*. Indianapolis: Hackett.
Lewis, David (1988) *Papers in Philosophical Logic*. Cambridge: Cambridge University Press.
Maddy, Penelope (1997) *Naturalism in Mathematics*. Oxford: Clarendon Press.
McCarty, C. and Tennant, Neil (1987) Skolem's paradox and constructivism. *Journal of Philosophical Logic*, 16, 165–202.
Mortensen, Chris (1998) On the possibility of science without numbers. *Australasian Journal of Philosophy*, 76, 182–97.
Parsons, Charles (1977) What is the iterative conception of set? In R. E. Butts and Jaako Hintikka (eds.), *Logic, Foundations of Mathematics, and Computability Theory*, (pp. 335–67). Dordrecht: Reidel and reprinted in Benacerraf and Putnam (1983) and Parsons (1983).
Quine, W. V. O. (1970) *Philosophy of Logic*. Englewood Cliffs, NJ: Prentice-Hall.
Wright, Crispin (1998) Response to Dummett. In Schirn (1998) 389–405.

24

Logical Foundations of Set Theory and Mathematics

MARY TILES

For much of the twentieth century the philosophy of mathematics centered around studies in the foundations of mathematics and these typically concentrated on set theory, arithmetic and the theory of real numbers. They also typically involved work in formal logic. Thus it is natural enough to think that we know what we should be talking about when given the above title. But do we really? To make sure that we do, we had better start by considering what it means to talk of foundations in general and logical foundations in particular, especially in the context of mathematics. We also have to find out how set theory gets into the act.

1 Foundations and Logical Foundations

> I was especially pleased with mathematics because of the certainty and self evidence of its proofs; but I did not yet see its true usefulness, and thinking that is was good only for the mechanical arts, I was astonished that nothing more noble had been built on so firm and solid a foundation. On the other hand I compared the ethical writings of the ancient pagans to very superb and magnificent palaces built only on mud and sand . . . when it came to the other branches of learning, since they took their cardinal principles from philosophy, I judged that nothing solid could have been built on so insecure a foundation. (Descartes 1960: 7–8)

The metaphor of building a dwelling on new, secure foundations pervades Descartes two most popular works, *Discourse on Method and Meditations*, and is one of the ways in which he influenced subsequent developments in philosophy. It was only in the second half of the twentieth century that this quest for foundations ceased to dominate philosophical discourse, although it remains a persistent theme, especially within the philosophy of mathematics, although even here it is increasingly being challenged.

When Descartes talked about foundations his conception of how to find them was influenced by what he conceived to have been the ancient mathematical method of analysis. Citing Pappus, he saw the method of analysis as a procedure for working back

to the first principles upon which any putative item of knowledge would be based. The passage from Pappus in which analysis is described is as follows:

> in analysis we assume that which is sought as if it were (already) done, and we inquire what it is from which this results, and again what is the antecedent cause of the latter and so on, until, by so retracing our steps we come upon something already known or belonging to the class of first principles, and such a method we call analysis as being a solution backwards.
>
> But in synthesis, reversing the process, we take as already done that which was last arrived at in the analysis and, by arranging in their natural order as consequences what were before antecedents, and successively connecting than one with another, we arrive finally at the construction which was sought, and this we call synthesis. (Editor's note in Euclid 1926: vol. I, pp. 138–9)

It is thus by analysis that we get back to first principles, but they are only shown to be adequate as first principles if the corresponding synthesis can be completed, that is if we can show that they provide a basis from which the knowledge to be grounded can be deduced.

In Descartes' hands the analysis which provides a grounding for knowledge involves revealing the complexity of what is to be known by taking it apart into its simpler components. This may involve analysis both of objects and of concepts. Indeed terms for objects will have to be redefined by reference to the way the objects are constructed. Cartesian foundationalism was inseparable from its mechanistic reductivism. An adequate foundation is then also an ontological foundation; it tells us what our knowledge is really knowledge about. It is also epistemological in the sense of showing that and how our knowledge claims are justified. If in addition one believes that it is the ability to provide a justification that constitutes possession of knowledge, then the analysis–synthesis circuit is also a route to knowledge acquisition and analysis is a method of discovery, as Descartes himself claimed.

Descartes was not, however, looking for logical foundations. A logical foundation would be a set of first principles from which one can, using definitions and logically valid deductive arguments provide proofs for the desired knowledge claims. Descartes had a low opinion of logic, which in the seventeenth century encompassed little more than the theory of syllogisms. He wanted his foundations to provide a bedrock on which to build an edifice of human knowledge. The building would be raised by a process of deductive synthesis from principles which could 'clearly and distinctly perceived to be true' by a sequence of steps each of which was accompanied by the same sense of self-evident correctness. These steps are neither confined to the rules of any formal logic nor would logical validity automatically be sufficient for the kind of support required. A deduction must reveal what it is in virtue of which the deduced statement is true. A logically valid indirect proof using *reductio ad absurdum* frequently does not do this, it merely shows why the opposite of what is to be proved cannot be true. Moreover, if analysis involves the analysis of objects, the reversing synthesis involves the construction of objects. Knowledge of objects is grounded in their method of construction, a kind of construction altogether different from the construction of concepts using definition by genus and differentia (man is a rational animal, for example) which was the form well-suited to syllogistic reasoning.

2 Foundations for Mathematics

Why should mathematics be thought to need any foundation? This after all was the discipline which inspired Descartes, providing him with a paradigm of a structure solidly erected on firm foundations. The mathematical work central to the formation of this paradigm was Euclid's *Elements of Geometry*. There the foundations, namely, axioms, postulates and definitions, are laid out at the beginning of each book, and step by step a body of knowledge is erected through the proof of theorems, later theorems building on results established earlier. But, as a glance at Descartes' own treatise on geometry (Descartes 1925) quickly confirms, the mathematics of the seventeenth century was already moving well beyond the confines of Euclidean geometry. Newton's *Principia* (Newton 1999), written in 1686, strikingly confirms both the hold of the Euclidean paradigm as a paradigm for organizing a body of knowledge, and the extent to which mathematics and its methods have moved away from classical geometry. Newton proceeds by presenting axioms and definitions. His laws of motion are presented as axioms. His text consists of propositions proved on their basis but the means of proof introduce mathematical methods unknown to Euclid. Like Descartes Newton uses algebraic methods and builds an understanding of 'complex' motions on the basis of their composition from 'simpler' motions. This is what Kant (1996) would later call reasoning from the construction of concepts, reasoning which he contrasted with logical reasoning (reasoning from concepts). Kant saw the distinctive power of mathematics as deriving from the fact that it employs this form of constructive reasoning, reasoning grounded in the way its objects are constructed in pure intuition. Reasoning from concepts according to the laws of syllogistic logic could establish analytic truths (those based on the analysis of concepts), whereas mathematical reasoning from the construction of concepts establishes synthetic *a priori* truths.

In addition to using algebraic methods Newton introduced the language of fluxions in the process of developing the techniques which were to become 'infinitesimal calculus.' The soundness of proofs constructed by these means was quickly challenged (by Berkeley (1992) and others). The methods used seemed to many to be inherently insecure because they involved trying to treat continuous magnitudes as if they could be made up of infinitely many discrete parts. This is spite of the fact that Zeno's paradoxes (discussed by Aristotle (1996: 238b23–240b8), and used by him as a basis for insisting on an absolute distinction between discrete and continuous magnitudes) and other well-known paradoxes of the infinite suggested that such moves would lead to inconsistencies and contradictions. The mathematics that, as its use proliferated, came more and more urgently to seem to be in need of 'foundations' – solid construction on a secure base – was that of analytic geometry and the methods of infinitesimal calculus.

The challenge that infinitesimals pose to a foundationalism centered on the idea of knowledge based on methods of construction is that, even supposing there are infinitesimally small limits of division (analysis) of a continuous line, the reverse synthetic process can never be humanly completed – it would be an infinite process. It would seem to require an infinite mind to understand an infinitely complex whole on the basis of its parts. Both Kant (1996: 531) and subsequently Cantor (see Dauben 1979: 130–1)

firmly declared the idea that analysis should reveal infinite complexity – structure all the way down – to be absurd. It is absurd to the extent that it violates a core principle on which the Cartesian foundational program was mounted – understanding is grounded in methods of construction. If we cannot locate simple parts at the end of a finite analysis, we humans will never reach a foundation on which to begin building. The challenge to provide a foundation for the new, infinitistic mathematics, was thus to find a way round this problem.

3 Mathematics and Set Theory

In Descartes' geometry, as also in Kant's treatment of mathematics, the problem noted above is finessed in the following way for the case of continua. They think of continuous magnitudes as constructed objects by invoking the concept of continuous action (motion). A line is constructed as the continuous motion of a point, which moves according to a law given in the form of an algebraic expression. This law expresses a complex ratio of distances from given, fixed lines (axes) whose value is constantly expressed by the moving point, and which is thus exhibited by any and every point on the constructed line. To study a curve through its algebraic characterization is then to learn about it on the basis of its method of construction, which is not a building up of discrete parts, but a continuous generation of a continuous whole.

This is a viable position as long as it is possible to think, as had been done since Aristotle's discussion of Zeno's paradoxes, that there are two irreducibly distinct kinds of whole – continuous and discrete. Europe inherited from the ancient Greeks the view that mathematics has two distinct branches – geometry, the science of continuous magnitudes, and arithmetic, the science of discrete magnitudes. Discrete magnitudes are aggregates of parts (elements); they are formed by heaping together a number of discrete items and are thus said to be 'wholes given after their parts.' A continuous whole, on the other hand, can be divided without limit and can be divided anywhere; its parts are 'created' by division which is a process of delimiting the boundaries of a part. Thus a part here is always essentially a part *of* the whole from which it is marked off and it is for this reason that continuous wholes are said to be 'wholes given before their parts.' Furthermore, because a continuous whole can be divided without limit, it potentially contains infinitely many parts. The point of distinguishing firmly between wholes given before and wholes given after their parts was to underscore the point that one cannot, on pain of contradiction, think of a continuous whole as something constructed out of the infinitely many parts it potentially contains; these parts cannot be treated as independently given discrete parts to be heaped into an aggregate.

The position taken by Kant and Descartes proved unstable for two reasons. First it was criticized for relying on the concept of motion, which, being drawn from mechanics, was unsuitable for use in thinking through the foundations of pure mathematics. Second, because it appears to place restrictions on the possible objects of mathematical study which mathematicians themselves saw no reason to recognize. It is possible to write (construct) algebraic expressions which don't correspond to any continuous or even drawable curve. What reason could be given for ruling that the complex relationships expressed in such equations should not be legitimate objects of mathemati-

cal investigation? In the eighteenth century mathematicians such as Euler and D'Alembert argued over what was to count as a function. In the end the notion of a function was liberalized in such a way that any collection of points in the plane could count as the graph of a function, and any method of calculating a real number as value for other real numbers as arguments would count as a function.

In many ways this simply reflects recognition that the move of introducing algebraic methods into geometry, of which Descartes' work was a part, and the introduction of Cartesian coordinates presupposes that each point in the Euclidean plane can be indexed by a pair of numbers, its coordinates. This in turn presupposes that a continuous line or plane can be represented by a set of numbers, or of pairs of numbers. Thus one must after all be able to view a continuum as composed of infinitely many points, in spite of the well-known contradictions arising from the supposition that one can add dimensionless points, items having length zero, together in such a way that they make up a continuous line having a positive length. The move thus involves unification of two opposed ways of thinking about part and wholes and their associated concepts of magnitude. The challenge was to find a way of doing this while avoiding the known and very real hazard of ending up with an inconsistent theory. Modern set theory proposes a solution, but without, as we shall see below, solving all the puzzles.

Mathematicians were thus firmly pushed in the direction of thinking of the Euclidean plane as an aggregate of points, if not as an aggregate constructed from points. The direction taken by Hilbert, Cantor and others was not to think about how to build up a continuum out of points, but to try to state the conditions which would have to be satisfied by a given collection of infinitely many points for them to count as constituting the points of a Euclidean plane, or a continuous line. Similarly instead of thinking about functions by starting from the lines which are their graphs, a (real valued) function of a single real variable is to be thought of as the set of ordered pairs which would be the coordinates of the points on its graph. One can then investigate what characteristics this set must possess if the function is to be continuous at a given point, differentiable at that point, and so on. Indeed Hilbert (1971) provided a new axiomatization of geometry along these lines and then proved that the real numbers could be used to 'construct' a structure (model) in which the axioms were satisfied. This appears to effect a reduction of geometry to the study of sets of points and their possible structures in conjunction with the study of real numbers.

But what is a real number? How are the real numbers defined? By making use of the concepts ordered pair, and infinite sequence mathematicians such as Cantor and Dedekind showed that one could start from the natural numbers – 0,1,2, . . . – to define the integers (negative and positive whole numbers) as ordered pairs of natural numbers where, for example, (1,2) represents 1–2, that is –1, and (2,1) represents 2–1, that is 1. Ordered pairs of integers represent the rational numbers, (1,4) is 1/4, etc. Real numbers can be defined as infinite convergent sequences (Cauchy sequences) of rational numbers. (A sequence of rational numbers is convergent if after some point the difference between successive terms gets smaller and smaller, as in 1, 1/2, 1/4, 1/8, . . .) In each case it has to be shown that the representatives have all the properties required of the numbers they are to represent. This is done by providing an axiomatic characterization of the structure required and then showing that these entities and operations defined over them can be shown to satisfy the axioms.

These moves have three possible philosophic interpretations. One (the logisist) says that the definitions show what the different kinds of numbers are and thus we have an ontological reduction of integers, rational and real numbers to natural numbers. Another (the formalist) says that these constructions prove the consistency of the axioms for integers, rational numbers and real numbers, relative to those for the natural numbers and whatever is needed for the constructions in terms of ordered pairs and infinite sequences. A third (the intuitionist) says that because the real numbers are defined as infinite, incompletable sequences, our reasoning about them has to proceed in a different way that our reasoning about the integers or rational numbers, assertion about real numbers cannot be presumed to obey the law of excluded middle. Intuitionists and constructivists resist assimilations of mathematical reasoning to logical reasoning along with any presumption that the infinite can be treated by analogy with the finite.

If the reduction could continue and the natural numbers could themselves be defined in terms of sets, then it would seem that one might be able to claim that set theory provides the ultimate foundation for mathematics. All the objects seem to be definable as sets and so in principle all theoretical results should be translatable, in principle into language which talks only about sets and operations on sets. The Bourbaki program, carried out by a group of French mathematicians, shows that this really is possible for large areas of mathematics.

The step that is made in the development of modern set theory, which allows the above constructions and allows it to accommodate aspects of the theory of both discrete and continuous wholes, wholes given before and whole given after their parts, is the distinction between set membership and set inclusion. The relationship between a set and its members, corresponds to that between a discrete whole (aggregate) and its parts and the relationship between a set and its subsets has to take over the work done by the relationship between whole given before its parts and those parts.

Sets are assumed to be identical if and only if they have the same members, so in this sense sets are defined by their members. Moreover, since the subset relationship can be defined in terms of the membership relation (A is a subset of B if and only all members of A are members of B) the barrier between these two ways of thinking about wholes and parts becomes permeable. In principle all sets are regarded as discrete wholes, even though some are infinite. However, it is also assumed that a subset of a given set A can be defined as the set of all elements of A having property P. This way of defining sets makes them subsets *of* a given set, that is parts given after the whole. It is further assumed that for any set A there is a set, the power set of A, containing as its elements all and only subsets of A. The barrier between the theory of discrete and continuous wholes, wholes given before, and those given after their parts is transformed into a double gulf (1) between finite and infinite sets and (2) between an infinite set and its power set – the set of all its subsets. The power sets of infinite sets are resistant to being treated as discrete wholes – things to which one might put a number in the same sense in which one can put a number to a finite set. This resistance is reflected in the independence of Cantor's Continuum Hypothesis from the remaining axioms of ZF set theory. (This hypothesis says that the cardinal number of the set of all subsets of the natural numbers is the next infinite cardinal number after that of the set of natural numbers. Cantor had already proved that the cardinal number of the set of real

numbers is the same as the cardinal number of the set of all subsets of the natural numbers.)

4 Sets, Classes, and Logic

So how does enquiry into the foundations of mathematics become a quest for logical foundations? By relating sets to classes and in this way making set constructions the product of corresponding logical operations for defining predicates. Then the way is cleared for losing the distinction between the synthesis which is logical deduction from first principles and the synthesis which is building up from simple component parts, and hence also the distinction between logical analysis (analysis of concepts) and analysis of objects. This will work if sets or classes are objects which can be constructed by logical operations on their corresponding concepts, but would not be possible without extending logic to cover relations and functions, as well as concepts, and the various operations used by mathematicians to define these. Accomplishing this task was Frege's major achievement.

Frege (1950) aimed to show that arithmetic is a body of analytic truths; that it really is a part of logic, in his new extended sense of logic. This includes the claims that classes are logical objects, that numbers are classes and that the application of any arithmetical truth is a matter of logical deduction. If Frege had succeeded he would thus have explained the universal applicability of arithmetic at the same time as providing it with a foundation in logic and the theory of classes.

The notion of set, or class, invoked in an informal way by Cantor and other mathematicians, already had a history in logic and attempts to introduce algebraic methods into logic, from Leibniz to Venn, De Morgan, and Boole. In traditional logic a class is the extension of a term – the collection of objects of which that term can be correctly predicated. Classes are thus wholes to which the theory of discrete, rather than continuous magnitude would apply.

The first thing that Frege needed to do was to introduce into logic a reflection of the distinction between the membership and subset relations. In Aristotelian logic this was not marked because singular statements, such as 'Aristotle was bald' were, for the purposes of syllogistic logic, treated as universal sentences, that is by analogy with 'All Greek males are bald.' Both of these would have been assigned the form 'S a P' and would then be viewed as expressing either an intensional relation (the predicate P is included in the concept of the subject S) or an extensional relation (the extension of the subject term S is included in the extension of the predicate term P). Frege on the other hand insisted on the distinction between object and concept as a logical distinction and one that should be reflected in logical notation. Objects have to be reflected at the logical level if the application of numbers is to be a logical operation, for it is objects that are counted and it is objects that are formed into sets.

The logic we have inherited from Frege, via Russell and others, thus starts from the singular sentence, P(a) which corresponds to the set theoretic form 'a $\in$ {x:Px}.' The universal then has the form '$\forall x(S(x) \in P(x))$' which in turn can be used to define the subset relation: $A \subseteq B$ if and only if $\forall x(x \in A \rightarrow x \in B)$. Frege also argued that set theory had to be based in logic if it was to hope to account for numbers and our use of

them. The idea of a set as an aggregate of objects runs into problems trying to account for the bases of the system of natural numbers – 0 and 1. How can there be a heap containing no objects? Moreover what is the difference between a heap containing a single object and just a single object. Frege's insistence that sets should be thought of as classes, the extensions of concepts, avoids these puzzles. It is easy to define a concept ('is a round square' or '$x \neq x$' for example) under which no object can possibly fall, and which hence has an empty extension. So 0 is the number of the concept '$x \neq x$.' Similarly there can be concepts under which only one object falls ('$0 = x$,' for example) whose extensions contain a single object. So 1 is the number of the concept '$0 = x$.' Frege thus asserted that a statement in which a number is applied is a statement about a concept; it says how many things fall under it. But he also insisted that numbers are themselves objects which can be grouped into classes. He ends up defining numbers as classes, saying that for any concept F, the number of Fs is the class of classes which are equinumerous with the class of Fs. So, for example 1 becomes the class of all classes equinumerous with the class of things identical to 0.

With the numbers so defined Frege shows, using only his logical principles and definitions, that they will satisfy the axioms for the natural numbers, given earlier by Peano. This would justify his claim that the truths of arithmetic are really logical truths, expressible using only logical concepts such as identity, object, concept, and class together with logical operations, such as negation, conjunction, and the formation of universal and existential generalizations (expressed with his newly introduced quantifier/bound variable notation). Unfortunately, as is well-known, Frege's logic was shown by Russell to be inconsistent; it permits the existence of the class of all classes which do not belong to themselves and if this class either belongs or does not belong to itself, a contradiction results.

Russell's response (Whitehead and Russell 1910–13) was to place restrictions on the predicates which could be thought to determine classes. His vicious circle principle, used in developing the ramified theory of types, bans classes from being defined and formed by reference to more encompassing classes to which they would belong. So, for example, no class can be defined by referring to the totality of all classes, since it would itself belong to that totality. This principle insists that the 'parts,' or members, of a discrete whole must be definable independently of that whole. In addition Russell insists that classes are basically logical fictions, not genuine objects. In other words, statements about a class should in principle be expressible as statements about the members of that class. This would not be possible if the vicious circle principle were violated. His image is then very reductivistically foundational, with a vision of a universe of classes which can be built up successively from a given stock of individuals, and where the whole superstructure could in principle be shown to provide only a shorthand for making complex descriptions of that universe of individuals. This vision had great appeal to empiricists since it appeared to obviate the need to postulate the existence of any abstract objects in order to account for mathematical knowledge and its wide applicability.

The problem is that, as Russell himself was forced to recognize, this does not yield a theory of classes which meets mathematicians' requirements. If we remember that what mathematicians required was a unification of the theory of wholes given before their parts with that of wholes given after their parts, we can understand why Russell's

complex system, although much richer than anything achievable with traditional logic, will not serve, for it is constrained to a theory of wholes given after their independently specifiable parts and replaces set construction by logical construction of their defining predicates.

In order to have a theory rich enough to develop mathematics Russell had to add two specific axioms – Infinity, which says that there are infinitely many individuals, and Reduction, which basically allows the existence of all subclassess of a given class, no matter how defined, to be collected into a class. Both of these are existence axioms and cannot easily be claimed to be logical truths. Moreover their use raises once again the problem of consistency – how could one be sure that tacking these two axioms onto the system will not render it inconsistent?

An alternative response to the problems with Frege's logic was to axiomatize the theory of sets and then think about how to prove the axiomatized theory consistent. The axiomatization now regarded as standard is based on those of Zermelo and Fraenkel (hence called ZF). It includes operations for building up sets member by member, but also for an infinite set and for using predicates to mark off the subsets of an already given set. The totality of subsets of a given set is asserted to exist without any restriction which says that these have to be definable as the extensions of predicates. Moreover, in many cases an additional axiom, the Axiom of Choice is added, and this explicitly asserts the existence of sets as aggregates of objects for which there may well be no such definition. Gödel (1938) showed that it is possible to provide a model for the ZF axioms by restricting sets to those which are definable (the constructive universe). In this universe the axiom of choice and Cantor's continuum hypothesis would be true. However he and others have also argued that this universe is too restrictive for mathematical purposes. Subsequently Cohen (1966) proved that both the axioms of choice and the continuum hypothesis are independent of the remaining axioms of ZF set theory. This means that the basic ZF axioms remain neutral on whether set construction is reducible to logical construction, but to the extent that mathematics seems to require use of the axiom of choice and to presume a universe containing non-constructible sets, this reductive restriction is rejected.

The resulting relation between logic and set theory is complex. It is certainly not a matter of one providing a foundation for the other. ZF set theory is written in the language of classical first-order predicate logic and any results proved about theories written in such a language apply to set theory. Some of those results, however, are proved using set theory, since the semantic approach to the study of predicate logic, relies on the concept of a model, and models are defined as structured sets. Results about models are then proved in set theory. So there is a complex, symbiotic relation between axiomatic set theory and predicate logic.

Hilbert's (formalist) program was to develop finitary methods for theorizing about formally expressed axiomatic theories with the aim of proving whether or not they are consistent. The idea was that if it could be proved using only finitary methods that a theory of infinite sets was consistent (that no formal contradiction could be proved from the axioms) then it would be safe to use. Again this is a way of seeking to use a constructive base to legitimize something which goes beyond it.

Gödel (1962) contains his famous incompleteness results. His first incompleteness theorem showed that any consistent formal system capable of expressing arithmetic

would contain undecidable arithmetic sentences. On the assumption that any statement about numbers is either true or false, this would imply that there would always be some arithmetic truth that could not be proved in the particular formal system in question. This creates a problem for the logicist who wants to say that every arithmetic truth is a logical truth. It either has to be allowed that no formal system captures the notion of logical truth, or that the logicist claim is false, or that not every statement about numbers is determinately true of false. His second incompleteness result shows that the consistency of such a system cannot be proved by means formalizable within the system, which demonstrates that Hilbert's program for providing an ultimate consistency proof for infinitary methods by finitary means cannot be realized.

Where did this leave foundational programs? Although Gödel's results undercut the philosophical rationale for both logicist and formalist programs, foundational studies had taken on a life of their own. New branches of mathematics, and new ways of studying logics had been developed. There were plenty of things to be discovered about these new domains and work in all these areas for a while continued to fall under the title studies in the foundations of mathematics. Philosophers too needed to learn from the technical results to try to decipher their philosophical significance. The idea that mathematics has a foundation in logic could still be pursued by debating the boundaries of logic and the way in which a reduction to logic might be effected. However, that particular convergence of mathematics, set theory and logic required to reduce the construction of mathematical objects to logical construction (definition of predicates), which was central to the plausibility of the claim that mathematics could be provided with a foundation in logic, proved to be relatively short lived. By the late twentieth century logic, set theory and mathematics were developing on independent tracks, interacting in complex ways, but none serving as a bedrock on which to raise the others.

The weaker claim that any branch of mathematics can be given a logical foundation, by being written as an axiomatized theory in the language of first order logic, leads to a different way of saying that all mathematical truths are logical truths. One can then say that what mathematicians prove are logical truths of the form 'If P then A,' where P is some finite conjunction of axioms. If the axioms are inconsistent, all such statements are still be logical truths, given the materiality of the conditional in classical first-order logic. Unfortunately this gives a much too simplistic picture of mathematical practice. Take for example, Wyle's proof of Fermat's last theorem. This appeals to results in many branches of mathematics other than arithmetic. To even begin to represent his proof as establishing a logical connection one would include the axiomatizations of all these other bits of mathematics, and give a logical representation of the process of applying the results from one mathematical domain in another. Thanks to the work of the Bourbaki group in showing how to do mathematics within the framework of set theory, one might say that in principle this could be done within set theory; but others would question whether such a thing (a full formal proof) would be able to serve the functions of a proof – convincing people by helping them understand why the conclusion is true.

The focus of foundational studies was set in the nineteenth century at a time when it seemed that numbers of various kinds were the fundamental objects of mathematical investigation. In the twentieth century mathematics seemed to be equally concerned

with the investigation of structures and procedures. Structures can be characterized without saying how they can be built from objects. They can be characterized on the basis of the kinds of transformations under which they are preserved. This idea gave rise to a rival foundational bid from category theory, where objects are complex wholes given before their parts and internal structure is revealed through a study of the way they relate to other objects of their kind (category) through structure preserving mappings (morphisms).

The study of finitary procedures led to the theory of recursive and computable functions and to the developments of electronic computers. The extensive use and deployment of these computers has in turn been instrumental in undermining some of the presumptions which made foundational programs seem plausible. In particular development of the study of fractals, and complex systems, coupled with earlier results in nonstandard analysis, show that there is no more risk of contradiction associated with infinitesimals and the idea of structure all the way down, than with infinitely large sets.

Attempts to make computers into expert systems have stimulated the study of alternative logics, some of which (particularly non-monotonic and fuzzy logics) depart radically from the systems developed by Frege and Russell. In addition uses such as computer modeling have meant that there is continued interest in mathematics developed by constructivists, those who resisted both the move to reduce mathematics to logic and the use of infinitistic methods. Since computer memories are decidedly finite, computer representations of the continuous have to be based on finitary, approximative methods.

So we are once again in a context where it is not at all clear what a logical foundation for mathematics would look like, nor is it clear that logic is the place to look for foundations or even that foundations are what we need to be looking for.

References

Aristotle (1996) *Physics* (Robin Waterfield, trans.). Oxford and New York: Oxford University Press.

Berkeley, George (1992) *"De Motu" and "The Analyst": A Modern Edition with Introduction and Commentary* (Douglas M. Jesseph, trans.). Dordrecht and Boston: Kluwer Academic (original works published 1734).

Cohen, P. J. (1966) *Set Theory and the Continuum Hypothesis*. New York: W. A. BenJamin.

Dauben, Joseph Warren (1979) *Georg Cantor: His Mathematics and Philosophy of the Infinite*. Cambridge, MA, and London: Harvard University Press.

Descartes, René (1925) *The Geometry of René Descartes* (D. E. Smith and M. L. Latham, trans.). Chicago and London: Open Court (original work published 1637).

Descartes, René (1960) *Discourse on Method and Meditations* (Laurence J. Lafleur, trans.). New York and London: Macmillan (original work published 1637).

Euclid (1926) *The Thirteen Books of Euclid's Elements* (T. L. Heath, trans.). Cambridge: Cambridge University Press.

Frege, Gottlob (1950) *The Foundations of Arithmetic* (J. L. Austin, trans.). Oxford: Blackwell (original work published 1884).

Gödel, Kurt (1938) The consistency of the axiom of choice and the generalized continuum hypothesis. *Proc. Nat. Acad. Sci. USA*, 24, 556–7.

Gödel, Kurt (1962) *On Formally Undecidable Propositions of* Principia Mathematica *and Related Systems* (B. Meltzer, trans.). Edinburgh and London: Oliver and Boyd (original work published 1931).

Hilbert, David (1971) *Foundations of Geometry* (Leo Unger, trans.). La Salle, IL: Open Court (original work published in 1899).

Kant, Immanuel (1996) *Critique of Pure Reason* (Werner S. Pluhar, trans.). Indianapolis and Cambridge: Hackett (original published in 1787).

Newton, Isaac (1999) *The Principia* (I. Bernard Cohen and Anne Whitman, trans.). Berkeley, CA: University of California Press (original published in 1687).

Whitehead, A. N. and Russell, B. (1910–13) *Principia Mathematica*, 3 vols. Cambridge: Cambridge University Press.

Further Reading

Benacerraf, Paul and Putnam, Hilary (eds.) (1983) *Philosophy of Mathematics: Selected Readings*. Cambridge: Cambridge University Press.

Devlin, Keith (1997) *Goodbye, Descartes: The End of Logic and the Search for a New Cosmology of the Mind*. New York: Wiley.

Dummett, Michael (1991) *Frege: Philosophy of Mathematics*. London: Duckworth.

Hart, W. D. (ed.) (1996) *The Philosophy of Mathematics*. Oxford: Oxford University Press.

Lachterman, D. R. (1989) *The Ethics of Geometry: A Genealogy of Modernity*. New York and London: Routledge.

Shanker, S. G. (ed.) (1988) *Gödel's Theorem in Focus*. London, New York, Sydney: Croom Helm.

Tiles, Mary (1989) *The Philosophy of Set Theory*. Oxford: Blackwell.

Tiles, Mary (1991) *Mathematics and the Image of Reason*. London and New York: Routledge.

Tymoczko, Thomas (ed.) (1998) *New Directions in the Philosophy of Mathematics*. Princeton, NJ: Princeton University Press.

25

Property-Theoretic Foundations of Mathematics

MICHAEL JUBIEN

1 Introduction

The main goal of this essay is to show how a certain comparatively weak theory of properties is adequate to provide a 'foundation' for classical mathematics. Theories of properties have of course been enlisted in foundational efforts in the past. The most prominent example is surely Whitehead and Russell's (1910–13) theory of 'propositional functions' (where these entities are taken as properties and relations in intension). George Bealer (1982) has provided a more contemporary example. These theories are very far-reaching and intricate, and they are also very different. Whitehead and Russell's is a 'ramified type theory' in which each propositional function appears at a certain level in a complex infinite hierarchy, but Bealer's properties aren't 'stratified' at all.

In the middle decades of the twentieth century 'intensional' entities such as *properties* and *propositions* were generally either regarded with great suspicion or else rejected outright. This tendency often reflected the powerful influence of W. V. Quine, who argued that 'extensional' entities – notably *sets* – are more respectable philosophically and are capable of doing any work that might have seemed to require intensional entities (see Quine 1960). During this same period set theory flourished in its own right as an autonomous branch of mathematics and came to be regarded very generally as the ultimate foundation of classical mathematics. In this atmosphere the idea of property-theoretic foundations was so far from most people's minds that even Whitehead and Russell's effort was typically thought of as a foundation within a system of *set theory*, with the apparent dependency of sets upon intensional entities ignored or forgotten (see Parsons 1967).

But things have changed. In the past few decades intensional entities have come to enjoy a great deal of attention along with greatly revived respectability. To a large extent this traces to a surge of interest in modal logic, its semantics, and the philosophical discussion of alethic modality, especially the question of 'essentialism.' (Prominent contributors to these developments include Ruth Barcan Marcus and Saul Kripke.) The resuscitation of such a seemingly paradigmatic intensional notion as *modality* brought in its wake a renewed interest in all matters intensional, including of course intensional *entities*. This in turn prompted a reexamination of Quine's criticism of these entities,

and many have concluded that his criticism fails (e.g. Jubien 1996). Further, in a stunning reversal, philosophers have begun to argue that *sets* – so recently celebrated as paragons of clarity among *abstracta* – are in fact deeply obscure and mysterious (e.g. Bealer 1982; Jubien 1989). David Lewis (1991) also argues that these entities are fundamentally mysterious, but then throws up his hands and accepts impure nonempty sets (which he prefers to call *classes*) because he thinks they provide mathematics with a much-needed foundation.

At the present moment we, therefore, find some philosophers accepting properties and rejecting sets because, ironically, they think properties are clear while sets are obscure. And we even find philosophers who accept sets doing so with real reluctance because they concede the obscurity of the notion. As a very general comment, the stock in properties has been on a decades-long winning streak, while the stock in sets, though perhaps solid, has sustained some long-term decline and has some jittery holders.

So now is a good moment to revisit property-theoretic foundations of mathematics. The foundation I will offer here may have an appeal that others lack. For it's based on a very 'sparse,' epistemically conservative theory of properties, one that postulates no properties that aren't *intrinsic* to their instances. In fact it doesn't depend on postulating many properties or sorts of properties at all. It gets much of its foundational power from a dose of mereology that will be discussed later.

2 On Foundations

What do we really mean by 'foundations of mathematics'? Does mathematics even *need* a foundation? (Putnam (1967) argues that it does not.) If it does, must it have only one or is there room for many? The work that has historically been classified as 'foundations of mathematics' is remarkably diverse, ranging from the purely logical to the purely philosophical, and as a whole it provides no clear answers to these questions. (Parsons (1967) provides a very meticulous survey of the topic.) I won't make any effort at a complete discussion, but I will try to distinguish two very broad senses in which one might speak of foundations. These two notions are easily confused because the concept of the reducibility of one formal theory to another plays a central role in each. But in one of these senses there could only be one foundation of mathematics, while in the other there are potentially many different, equally acceptable foundations.

The *apparent* subject matter of classical mathematics includes such seemingly diverse entities as points, lines, natural numbers, real and imaginary numbers, ordered n-tuples, infinite sequences, functions, vector spaces, topological spaces, groups, rings, and so forth. Working mathematicians have typically proceeded as if these are independently existing 'Platonic' entities, and they haven't worried too much about whether this presupposition is ontologically defensible. But philosophers have often worried about it, and they have explored many positions, most of which may be seen as versions either of realism (Platonism), conceptualism, or nominalism, or as hybrids involving the reduction of entities of some kinds to others. We may think of such philosophical positions as *ontological* foundations, for at bottom they are claims about what mathematical entities really exist, along with claims about their ultimate natures.

For example, noting the reducibility of first-order versions of the various branches of classical mathematics to set theory, it might be held that it is really only *sets* that comprise the true subject matter of mathematics. Sets would be seen as abstract entities of their own special sort, and the idea that there are *also* such varieties of entities as numbers, vectors, and functions (etc.) would be abandoned. In this way sets (and set theory) would be seen as the ontological foundation of mathematics. A proponent of this position would owe a principled reason for preferring sets over any other sort of entities (such as properties) to which the apparent mathematical objects might also happen to be reducible. One ingredient of the reason could be the recent ascendency of sets themselves as apparent mathematical entities, but on its own this wouldn't be very convincing. For if the other apparent entities are up for grabs, why should sets be any different? The important point here is that since any proposed ontological foundation incorporates a claim about the real subject matter of mathematics, it is incompatible with each of its ontological alternatives.

Although I cannot try to establish it here, I think it is far from clear that the original motivation for seeking foundations required them to be *ontological*. Very roughly, I think the motivation was the worry that certain concepts of analysis involving the infinite were not well understood and threatened paradox. What is needed to address this kind of concern is a way to arrive at an adequate grasp of the concepts, one that increases our confidence that they aren't inherently incoherent or paradoxical. There is no good reason to suppose that this couldn't be done without making controversial ontological commitments. Nor is there any good reason to suppose it could only be done one way. Because the primary goal of foundations in this sense is improved understanding, we may think of such foundations as *epistemological*.

One way to arrive at a better understanding of a given concept is to 'model' it with other concepts. When specific theories are in hand, we may do this *syntactically* by carrying out a formal reduction of one theory to the other. If we initially understand the structural features of the reducing concepts better than those of the given concept, we automatically improve our grasp of that concept simply by doing the reduction. In the case of analysis and set theory, very intricate notions are readily 'modeled' in a theory whose sole primitive – the binary membership relation – is remarkably simple and, at least structurally, very accessible.

The reduction of one theory to another is also, in effect, a relative consistency proof: if the reducing theory is consistent, then so is the reduced theory. Of course this means the reducing theory is, logically speaking, at least as strong as the reduced theory. Despite this, it is possible to gain confidence in the coherence of the reduced theory as a result of the reduction. For our initial, intuitive conviction that the reducing theory is coherent may be substantially greater than our initial confidence in the coherence of the reduced theory, perhaps as a result of a greater accessibility of its primitive concepts. So the reduction can have the effect of boosting our confidence in the reduced theory rather than undermining our confidence in the reducing theory. (Of course we know from Gödel's work that the consistency of any first-order theory strong enough to reduce classical mathematics is in effect a matter of faith rather than proof.)

The reduction of analysis and the rest of classical mathematics to set theory should be seen as a great conceptual advance whether one accepts the existence of sets or not. For it shows that the various mathematical concepts are structurally related to the set

concept in certain ways regardless of whether that concept actually has instances. Since, under the epistemological conception of foundations, there would be no claim that sets comprise the ultimate subject matter of mathematics, it may not even matter whether they exist or not. We may be able to attain sufficiently improved understanding and confidence simply as a result of the formal reduction.

But it is also possible that the foundational gain would be even greater if we were convinced that the reducing entities really did exist. For example, someone who already believes the real ontology of mathematics is just its apparent Platonic ontology is likely to find it more satisfying to think that there actually exist *non-mathematical* entities that display the structural complexity of the putative mathematical objects. For then the foundation would not only enhance clarity and the conviction of coherence, it might also make Platonic mathematical objects seem more plausible by providing an independent precedent for Platonic entities of that level of complexity.

I will stop short of endorsing an ultimate ontology for mathematics. Instead I'll just mention what I think are the two most plausible candidates. They reflect a pair of ontological convictions. One is that the philosophical difficulties of sets are so overwhelming that sets should be rejected. The other is that the case for Platonic properties is very strong and doesn't rest on considerations about mathematics. Given these convictions, one candidate, inspired by Whitehead and Russell and by Frege, is that mathematics does have a genuinely ontological foundation in properties (and property theory). On this view, properties are the ultimate (and exclusive) subject matter of mathematics. Despite its historical moorings, this view would surely be seen as revisionary. The other candidate is more of a 'face-value' position. It's the view that mathematics really has no ontological foundation, that its ultimate subject matter is its original apparent ontology (of course, not including sets), and that an epistemological foundation in properties supports this ontology (as suggested above) while achieving the original foundational goals of clarity and coherence. I believe there is a good deal that can be said in favor of either of these positions. Of course they both require property-theoretic foundations, even if for one they are ontological and for the other they are not.

3 Properties, Sums, Plurality, and Reality

The theory to be offered here is basically the outgrowth of three simple ideas, two metaphysical and one purely logical. I believe both metaphysical notions are extremely compelling intuitively, and that the logical notion has a solid grounding in ordinary language. The first idea is a Platonic principle about properties: that properties 'constitute' things as being how they are. For a thing to be green just *is* for it to instantiate the property of *being green*, nothing more, nothing less. A corollary of this principle of constitution is that for any entity at all, since it is a *specific* entity, there is a property of *being that entity*. Intuitively, any such property must be *intrinsic* (and also *essential*) to its unique instance, but it must not be a *part* of that instance. The theory will postulate properties of this kind and also the property of *being self-identical*, which

certainly is also intrinsic to each of its instances. The intrinsicalness of the postulated properties (along with the fact that they are all instantiated) ensures that the theory is compatible with what is commonly called a 'sparse' conception of properties. And because it postulates only these very basic sparse properties, the theory is conservative epistemically.

The second idea is that any entities have a mereological sum, regardless of their individual natures. So not only are there sums of physical objects, there are also sums of abstract entities, and sums of entities some of which are concrete and others of which are abstract. The sum of any entities has each of them as a part, and has no part that isn't a part of some of those entities. (For a defense of the idea that mereology extends to the realm of the abstract, see Lewis (1991).)

The logical notion is 'plural' quantification. As it happens, plural quantifiers are quite common in ordinary English (but the phenomenon is still relatively unexplored in logic). Plural quantifiers range over the various pluralities of things, but without presupposing that the pluralities are individual entities like sets (or 'totalities' of any other kind). These quantifiers are not reducible to ordinary (singular) objectual quantifiers. A classic example of an ordinary English sentence with a plural quantifier is 'Some critics admire only each other.' It is clear what this sentence means, but that meaning cannot be captured with ordinary first-order quantifiers. Another example is the first sentence of the previous paragraph. (For more on plural quantification, see Boolos (1984).)

Unfortunately, these three very appealing ideas cannot be adopted in full generality, for then there would be a sum of *everything*, say S, and the property of *being S* would have to be a part of S. This is harsh reality. To avoid the problem we have to restrict at least one of the three ideas, and it is a further reality that no way of doing this can be quite as elegant and simple as the unadulterated combination of ideas. But this kind of reality has a familiar precedent, for the simplest and most elegant theory of *sets* – that based on unrestricted comprehension – proved to be self-contradictory. Our approach will be to limit the formation of sums and leave the other two notions alone. Other approaches are also possible, but we will not debate the relative merits of the various possibilities here. Our overall goal is merely to show that a great deal can be done even with an epistemically conservative conception of properties.

4 Mereological Property Theory

One of the central concepts in this theory (MPT) is the binary relation of *intrinsicalness*, which we will express with the symbol 'μ.' Thus 'xμy' means that y is intrinsic to x. Another key concept is expressed by the unary function symbol 'β,' which maps any entity x to the 'individuality' property of *being x* ('βx'). A third basic notion is the part–whole relation of mereology, which we will express with the binary predicate letter 'ρ,' so that 'xρy' means x is part of y. The theory will also employ identity and both singular and plural (objectual) quantification. Plural quantifiers will be enclosed within square brackets (e.g. '[∀x]' and '[∃x]') and singular quantifiers are left unenclosed ('∀x' and '∃x'). An occurrence of '[x]y' (etc.) within the scope of a plural quantifier on x is

understood to mean that y is an x (that is, y is one of the x's). MPT has four nonlogical axioms and an axiom scheme. (Here we view the axioms for identity and the functionality of β as logical axioms.)

1. *Self-Identity*: $\exists! x \forall y((y \mu x \leftrightarrow y = y) \& (y \rho x \rightarrow y = x))$.

This simply says that there is a unique property of which everything is an instance and which has no proper parts. We will call this property 'a' in the metalanguage. A consequence is that a instantiates itself. Of course there may be many distinct properties of which everything is an instance, but in practice we only need one, and we can think of it as the property of being self-identical. (An 'impure' version of MPT could be obtained by replacing Axiom 1 with an axiom postulating the existence of a concrete entity.)

2. *Individuality*: $\forall x \exists! y(y = \beta x \& \neg y \rho x \& \forall z((z \mu y \leftrightarrow z = x) \& (z \rho y \rightarrow z = y)))$.

For any entity x there is a unique property of *being x*, which is never a part of (and so never identical with) x itself, which has x as its only instance, and which has no proper parts. We will call any property that is a βx for some x a 'β-property.' Notice that the 'atomicity' of the β-properties ensures that any sum of given β-properties can have no other β-properties as parts. The β-properties are therefore in a certain sense independent of each other. We now adopt an axiom to ensure that the part–whole relation of MPT conforms to that of standard mereology.

3. *Mereology*: $\forall x \forall y \forall z(x \rho y \ \& \ y \rho z \rightarrow x \rho z) \ \&$
$\forall x \forall y \exists z[x \rho z \ \& \ y \rho z \ \& \ \forall w(w \rho z \rightarrow \exists v(v \rho w \ \& \ (v \rho x \vee v \rho y)))]$

The first conjunct ensures the transitivity of the part–whole relation and the second entails that any finite number of entities have a unique mereological sum. Hereafter we will not hesitate (in the metalanguage) to write 'x + y' for the sum of x and y. (Obviously + is commutative and associative.) The next axiom provides us with an entity that has infinitely many non-overlapping parts.

4. *Infinity*: $\exists x[\exists w(\forall y(y \mu w \leftrightarrow y = y) \ \& \ \beta w \rho x) \ \&$
$\forall y(y \rho x \ \& \ \neg(y = x) \ \& \ \forall z(\beta z \rho y \rightarrow (z \rho y \vee z = w))) \rightarrow \beta y \rho x)]$.

The entity x therefore has βa as a part and, for any part y other than x itself, it has βy as a part provided that whenever a property βz is part of y, either z = a or z is also part of y. We may think of sums of β-properties that meet this last condition as 'β-transitive.' Since βa is β-transitive, ββa must be part of x. But since ββa isn't β-transitive, the axiom doesn't require βββa to be part of x. On the other hand, βa + ββa *is* β-transitive, so its β-property is part of x. And since βa + ββa + β(βa + ββa) is *also* β-transitive, *its* β-property is part of x, and so on. Next we introduce an axiom scheme that, in combination with Axioms 1 and 4, 'generates' a vast universe of mereological sums.

5. *The Scheme of Generation*: All formulas
$\forall x \exists! y \mathbf{A}(x,y). \rightarrow . \forall z[\forall w][\forall u([w]u \rightarrow u\rho z) \rightarrow$
$\exists v[\forall x \forall y((([w]x \ \& \ \mathbf{A}(x,y)) \rightarrow y\rho v) \ \&$
$\forall s(s\rho v \rightarrow \exists x \exists y \exists t([w]x \ \& \ \mathbf{A}(x,y) \ \& \ t\rho y \ \& \ t\rho s))]],$

where x and y are free in the formula **A**(x,y). Although it looks rather complex, this scheme merely says, for each 'functional' formula **A**(x,y) (with parameters as required), that for any object z, and any w's, if the w's are parts of z, then there is a unique sum of the **A**-images of those w's. As a simple application, let z be any infinite sum conforming to Axiom 4. Then βa, $\beta\beta a$, $\beta(\beta a + \beta\beta a)$, and so on are all parts of z. Let 'the w's' be the parts of z that are β-properties, and let **A**(x,y) be the formula '$\exists s(x = \beta s \ \& \ \exists t(trs \ \& \ \neg(t = s) \ \& \ y = x) \vee y = u)$', where u is parametrized to $\beta(\beta a + \beta\beta a)$. **A**(x,y) is easily seen to be functional, for it maps each $w = \beta s$ to itself if s has a proper part, and maps every other w to $\beta(\beta a + \beta\beta a)$. So the axiom delivers a sum of which $\beta(\beta a + \beta\beta a)$ is a part but βa and $\beta\beta a$ are not.

For another example, let **A**(x,y) be the formula '$y = \beta x$' (which is obviously functional). Again take z to be any sum conforming to Axiom 4, and let 'the w's' be all parts of z. Then the resulting v is the sum of all the β-properties of its parts. So, for example, $\beta(\beta a + \beta(\beta a + \beta\beta a))$ is part of v and so is βz.

Although it generates a boundless universe, MPT is very 'minimalistic' from a property-theoretic perspective. In a nutshell, it says that there is a property (*self-identity*) that everything instantiates, that for anything at all there is the (mereologically atomic) 'individuality' property of *being that thing*, that there is a sum having infinitely many β-properties among its parts, and that the (intra-theoretically describable) functional *relata* of any parts of any entity has a sum. So, from the perspective of the theory, the only properties that need exist are *self-identity* and various β-properties, and the only nontrivial sums are those given by Axiom 4 and generated by Scheme 5. But even this 'minimal model' would constitute a remarkably lavish universe of entities – easily enough to provide an interpretation of ZF.

5 Foundations of Mathematics

Let's begin by making the (uncontroversial) assumption that classical mathematics is formally reducible to ZF. Then it suffices for our purposes if ZF, in turn, is formally reducible to MPT. In other words, MPT is an adequate foundation provided that '$\in$' is definable in the language of MPT in such a way that the translations of the ZF axioms under the definition are theorems of MPT. A careful demonstration that this is indeed the case would be lengthy and tedious. So in this section I will address the question semantically rather than axiomatically, and will do so in an informal and rather incomplete way. The goal is just to provide the construction that delineates a 'domain' of a 'model' of ZF within an arbitrary 'model' of MPT, along with an 'interpretation' of '$\in$'. Verification that the result actually is a 'model' will be omitted.

For this project to make sense we must employ a conception of *model* (and *interpretation* generally) that departs ontologically from the usual set-theoretic one. (Hence the quotation marks in the previous paragraph.) The main reason, as I see it, for exploring

property-theoretic foundations is the conviction that there really are no sets (whether pure or impure) and it is an immediate consequence that the sort of set-theoretic constructions that are usually thought to be the objects of model theory simply do not exist. If there are no sets, then model theory, as it is typically understood, has no subject matter. So we need to reinterpret it, and various approaches are possible.

The approach we will adopt avoids thinking of models as *individual entities* at all. Instead, the fundamental notion will be that of (plurally, now!) *some entities* modeling a theory by virtue of relations they bear to each other. As an example, suppose we have a first-order theory with a single binary relation symbol. In a typical model-theoretic approach, an *interpretation* of the theory is an ordered pair consisting of a nonempty set (the domain) and a set of ordered pairs of members of the domain (or some close variation on this idea). Such an interpretation is a *model* if the axioms are true when the quantifiers range over the domain and the relation symbol is interpreted as having the set of ordered pairs as its extension. So one model might be the ordered pair of the set of people who live in Detroit and the set of ordered pairs $\langle x,y \rangle$ of Detroiters x and y such that x is a parent of y. But, convenient though they may be, we don't need the sets and ordered pairs at all, because we understand perfectly well what it takes for the axioms of the theory to be true of the people in Detroit when the relation symbol is understood as expressing the parent-of relation. (Similarly, if there really are natural numbers, then *they* – with their characteristic relations – obey the axioms of number theory whether or not there exist any sets or ordered n-tuples of them from which to concoct standard model-theoretic interpretations.)

Now, using plural quantifiers in the metalanguage, we can say what it means to *specify an interpretation* without having to think of interpretations as specific *entities* of any sort whatever. We succeed in 'specifying an interpretation' whenever we give a clear specification of *some entities* (the intuitive *domain*) together with an appropriate association of *properties* and *relations* with the nonlogical symbols of the theory. If the theory's axioms happen to hold for those entities under those associations, then we have 'specified a *model*.' That there is no actual entity available to call 'the model' is at worst a mild inconvenience (and indeed is one that could be avoided in a full-blown *property-theoretic* model theory). The other side of this coin is that interpreting formal theories *via* set-theoretic objects is just a minor convenience, and is in no way a special source of mathematical or semantical rigor. In a successful case, the *specification* of the various sets that comprise a set-theoretic interpretation of a given theory would have exactly the same informal-though-precise status as the specifications of 'interpretations' in the present sense.

So now let's suppose that we have an arbitrary model of MPT. Then it contains entities we may view as ordinal numbers. In fact Axiom 4 very conveniently delivers an infinitude of such entities. We will define the ordinals in two stages, the first of which replicates the intuitive notion of β-transitivity mentioned above.

DEFINITION 1 A sum T of β-properties is β-*transitive* if for all x, if βx is part of T, then either x is part of T or x = a.

DEFINITION 2 A β-transitive sum T of β-properties is an *ordinal* if for any distinct x and y, if βx and βy are parts of T, then either βx is part of y or βy is part of x.

As an example, consider $\beta a + \beta\beta a + \beta\beta\beta a$. It is β-transitive, but it isn't an ordinal. For notice that if $x = a$ and $y = \beta\beta a$, then both βx and βy are parts of the sum, but βx is not part of y nor is βy part of x. On the other hand, $\beta a + \beta\beta a + \beta(\beta a + \beta\beta a)$ is easily seen to be an ordinal.

Because we have chosen not to view a as an ordinal, all ordinals are sums of β-properties. The ordinals begin like this:

$$\beta a;\ \beta a + \beta\beta a;\ \beta a + \beta\beta a + \beta(\beta a + \beta\beta a);\ \ldots$$

It is easy to see that the ordinals (up to any point) are well-ordered by the part–whole relation. We may define a *limit* ordinal as any ordinal other than βa that has no part βx such that every other β-part is part of x, and a *successor* ordinal as one that is neither βa nor a limit. (The existence of a limit ordinal follows from Axiom 4 and Scheme 5. To see this, note that the definitions of 'ordinal' and 'limit ordinal' may be captured by formulas of MPT, say '$\mathbf{O}(x)$' and '$\mathbf{L}(x)$.' Now let $\mathbf{A}(x,y)$ be the formula '$(\mathbf{O}(x) \,\&\, \forall y((y\rho x \rightarrow \neg\mathbf{L}(y)) \,\&\, y = x) \vee y = u)$,' with u parametrized to βa. Then $\mathbf{A}(x,y)$ is functional. Now let the z of Scheme 5 be any sum that conforms to Axiom 4, and let 'the w's' be all parts of v. The object v delivered by the resulting instance of the scheme is then the sum of all 'finite' ordinals and is easily seen to be a limit ordinal itself.) We may also define surrogates for the natural numbers by setting $0 = \beta a$ and the successor of n = the sum of n and βn.

We now employ transfinite induction to give the construction that will provide the basis for the domain of our model of ZF. The rough idea is to mimic the power set operation at successor stages and to mimic unions at limit stages. The result is a hierarchy of objects, each one a β-property, and hence each one a property with exactly one instance. From this hierarchy we may 'filter out' the domain of the model, and then define the surrogate of the membership relation. The first two stages of the construction are given explicitly by:

$S(0) = 0$ (i.e. βa) and $S(1) = \beta 0$ (i.e. $\beta\beta a$).

Next, for any ordinal α greater than 0, let

$$S(\alpha + 1) = \beta(x + \Sigma\beta y{:}y\rho x),\ \text{where } S(\alpha) = \beta x,$$

where '$\Sigma\beta y{:}y\rho x$' denotes the sum of the β's of the parts of x, guaranteed to exist by Scheme 5. (Note that, on the left side, '+1' means ordinal successor, not mereological sum.) So, at a given stage α, $S(\alpha)$ is some β-property, say βx. Then $S(\alpha + 1)$ is the result of applying β to the result of summing x with the sum of all objects βy, where y is a part of x. Finally, if λ is a limit ordinal, let

$$S(\lambda) = \beta\Sigma S(Y){:}Y\rho\lambda.$$

$S(\lambda)$ is therefore obtained by applying β to the sum of all earlier stages (which sum must exist by an application of Scheme 5 to the ordinal parts of λ *via* a functional formula

reflecting the method of construction). Now, given the entire hierarchy of S(δ)s, we are able to describe the domain of the model: an object is in the range of the ZF quantifiers iff it is a β-property that is a part of an instance of an S(δ). To illustrate, consider the first four stages:

$S(0) = \beta a;$
$S(1) = \beta\beta a;$
$S(2) = \beta(\beta a + \beta\beta a);$ and
$S(3) = \beta(\beta a + \beta\beta a + \beta\beta\beta a + \beta(\beta a + \beta\beta a)).$

So, for example, $\beta(\beta a + \beta\beta a)$ is in the domain because it is a β-property that is a part of the instance of S(3). But, for example, because $\beta a + \beta\beta a$ is *not* a β-property, it isn't in the domain even though it is a part of the instance of S(3) (and, for that matter, also a part of the instance of S(2)). That each $S(\alpha)$ is a β-property contributed to the domain by $S(\alpha + 1)$ is also illustrated here. For example, we have:

$S(3) = \beta(S(0) + S(1) + \beta S(1) + S(2)).$

Notice that every member of the domain is the result of applying the β-operator to a or to a sum of β-properties. Intuitively, βa will represent the null set, and every other member of the domain will represent the set whose members are precisely the sets represented by the β-properties that are parts of the sum. What makes this work is the 'independence of individuality' that is guaranteed by the atomicity of β-properties: no sum of any specific β-properties can have a β-property as a part that isn't one of those specific β-properties. It is now easy to provide the interpretation of '∈' in the given domain:

$x \in y$ iff $\exists z(y = \beta z \,\&\, x\rho z)$.

Notice that there is no need for a clause stating that x is a β-property since only β-properties fall in the range of the quantifiers in the first place. Thus it is clear that βa is the unique object from the domain that has no members – it's the surrogate of the empty set. There are no deep difficulties in verifying the other axioms of (pure) ZF. Thus we have shown how to construct a model of ZF from materials available in any model of MPT. It follows that MPT has at least as much foundational punch as ZF.

References

Bealer, G. (1982) *Quality and Concept*. Oxford: Clarendon Press.

Boolos, G. (1984) To be is to be the value of a variable (or to be some values of some variables). *Journal of Philosophy*, 81, 430–49.

Jubien, M. (1989) Straight talk about sets. *Philosophical Topics*, 17, 91–107.

Jubien, M. (1996) The myth of identity conditions. *Philosophical Perspectives*, 10, 343–56.

Lewis, D. (1991) *Parts of Classes*. Oxford: Basil Blackwell.

Parsons, C. (1967) Mathematics, foundations of. In P. Edwards (ed.), *The Encyclopedia of Philosophy*. (Vol. 5, pp. 188–213). New York: Macmillan and the Free Press.

Putnam, H. (1967) Mathematics without foundations. *Journal of Philosophy*, 64, 5–22.
Quine, W. V. O. (1960) *Word and Object*. Cambridge, MA: MIT Press.
Whitehead, A. N. and Russell, B. A. W. (1910–13) *Principia Mathematica*. Cambridge: Cambridge University Press (2nd edn., 1925–7).

This page intentionally left blank

Part IX

MODAL LOGICS AND SEMANTICS

This page intentionally left blank

26

Modal Logic

JOHAN VAN BENTHEM

1 Enriching Extensional Logic with Intensional Notions

When Frege wrote *Begriffsschrift*, he intentionally left out the key intensional notions of traditional logic before him. On one telling page he enumerates a list of things for which he sees no need – and readers of some erudition will recognize this anonymous enemy as Kant's famous "Table of Categories", including Modality. Nevertheless, in this century modal notions made their way back onto the logical agenda, leading to extensions of classical systems with operators of necessity, possibility, entailment, and other metaphysically inspired notions. These formalisms were influential as a tool for analyzing philosophical arguments. I still recall the shudder when reading my first sequences of symbols claiming to be a proof of God's existence 'out of a box.' But also, the semantics of modal logics in terms of possible worlds has formed a powerful philosophical union with the ontologies of Kripke and Lewis. These motivations also provided a watershed from mathematical logic, whose practitioners disliked modal logic instinctively, even though they are willing to countenance such deviations as intuitionistic or quantum logic. But worse than that, by the early 1980s, modal logic had also acquired powerful enemies within philosophy, preaching its imminent demise. I remember sneaking through corridors in those days, avoiding encounters with energetic colleagues who might be tempted to lend a helping hand to Historical Necessity. But modal logic did not die, its enemies never managed to invent an equally powerful substitute, its content and uses rather multiplied, and Handbooks wisely still include the subject.

2 Changing Views of Modal Logic

In what follows, I present a modern account of modal logic – not as a metaphysical system of any sort, but as a logical 'fine-structure formalism' for talking about actions, knowledge, and many other concrete things all around us. This view is very different from the original motivation in 'philosophical logic,' and I do not claim that it is uncontroversial in that field, especially among the *ancien regime*. But it is about time that a broader community learns what is really going on.

We have come a long way since the 1960s, because of two separate developments. First, what happened is a familiar phenomenon in science: originally non-intended applications of a theory take over. In the case of modal logic, these started with temporal and epistemic logic, then we had spatial logics, dynamic logics of action, and by now also modal logics of grammatical derivation, generalized quantifiers, games, or concept descriptions in AI. And this expansion is going on all the time. These applications provide new impetus to modal logic, at a time when it seems fair to say that ontology is no longer a live source of inspiration. A second influence came from inside modal logic. The mathematical theory of the subject that began to flourish in the 1970s yielded (as abstract mathematics should) surprising new viewpoints on what makes modal languages tick, which generated different perspectives – and in the end, a startling *inversion*. Viewed in one way, modal logics are typically extensions of classical logic with new operators. Viewed in another, and perhaps ultimately more insightful way, modal logics are *fragments* of classical logical languages, that serve as milestones in a natural 'fine-structure hierarchy' of expressiveness and reasoning.

Out of this historical panorama, we choose three notions as our major themes, *viz. fine-structure, information,* and *dynamics*. These will be introduced by looking at the basic modal language: propositional logic with box $\Box$ and diamond $\Diamond$. But first, let us mention another characteristic of much modal research: its exotic landscapes of different logics, such as K, T, S_4, S_5, or more bizarre code names. This seems a huge difference with a monotheistic religion like classical logic, which has only one set of validities – and hence many people associate modal logic with heathen botany. Now, the mathematical theory of the 1970s did create more unity in what has been called the 'jungle of modal logics.' Powerful meta-theorems appeared establishing properties like decidability, interpolation, frame-correspondence or completeness for whole families of 'modal logics' at once, or locating systematic failures – using methods from universal algebra and model theory. Nevertheless, and more controversially than our stance so far, we think this diversity is not a *fundamental* characteristic of the modal way of life – even though it is certainly one of its useful conveniences. Such 'logics' are different modal theories of special types of accessibility relation, comparable to special theories formulated on top of classical predicate logic. Our exposition will therefore concentrate on *modal base languages* and their properties, with an occasional excursion into the special frame classes of this other dimension of research.

3 A Précis of Basic Modal Logic

Language and interpretation

The basic modal language is very simple, and yet it has been a useful laboratory for new basic techniques. We interpret formulas in so-called possible worlds models – the grand name is still popular for its nostalgic mood – $\mathbf{M} = (W, R, V)$, according to the well-known truth definition:

$\mathbf{M}, s \models \Diamond A$ iff for some t with Rst, $\mathbf{M}, t \models A$
$\mathbf{M}, s \models \Box A$ iff for all t with Rst, $\mathbf{M}, t \models A$

It helps to think of the worlds as 'states' of some kind, while accessibility encodes possible moves that can be made to get from one state to another. But there are many other useful concrete views of these, in essence, 'decorated graphs' (figure 26.1).

Invariance for Bisimulation

The expressive power of this language is measured by a suitable notion of similarity between different models.

DEFINITION A *bisimulation* between two models **M**, **N** is a binary relation E between their states m, n s.t. whenever m E n, then (a) m, n satisfy the same proposition letters, (b1) if m R m′, then there exists a world n′ with n R n′ and m′ E n′, (b2) the same 'zigzag clause' holds in the opposite direction.

Together, this 'atomic harmony' and the two zigzag clauses make bisimulation a natural notion of 'process equivalence' – and indeed it was independently discovered in computer science. Example (disregarding proposition letters): the two black worlds in **M**, **N** are connected by the drawn bisimulation, consisting of all the matches indicated by dotted lines – but there is no bisimulation which includes a match between the black worlds in **N** and **K** (figure 26.2).

INVARIANCE LEMMA If E is a bisimulation between **M** and **N**, and m E n , then m, n satisfy the same modal formulas.

That is, modal formulas are *invariant for bisimulation*. Thus, we can see the above failure of bisimulation by noting that the model in the middle satisfies the formula

$$\Diamond\Diamond\Box\bot$$

in its root, whereas the one to the right does not. The converse to the Lemma only holds for a modal language with *arbitrary infinite* conjunctions and disjunction – or for the plain modal language over special models. For instance:

PROPOSITION If m, n satisfy the same modal formulas in *finite* models **M**, **N**, then there is a bisimulation E between these with m E n.

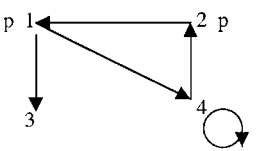


Figure 26.1

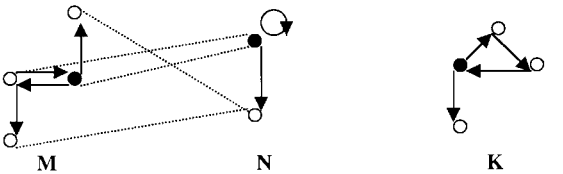


Figure 26.2

But there are still stronger definability results. For example, for any model **M**, s with a designated world s, there is an infinitary modal formula $\phi^{\mathbf{M},s}$ true in only those models **N**, t which are 'bisimilar' to **M**, s (i.e. some bisimulation links t to s).

Validity and proof systems

Universal validity is axiomatized in Hilbert-style by the so-called *minimal modal logic*:

1. all laws of propositional logic
2. a definition of $\Diamond\phi$ as $\neg\Box\neg\phi$
3. the modal distribution axiom $\Box(\phi \rightarrow \Psi) \rightarrow (\Box\phi \rightarrow \Box\Psi)$
4. the necessitation rule 'if $\vdash\phi$, then $\vdash\Box\phi$'

This looks like a standard axiomatization of first-order logic (with $\Box$ as $\forall$, and $\Diamond$ as $\exists$), but leaving out axioms with tricky side conditions on freedom and bondage: $\forall x\phi \rightarrow [t/x]\ \phi$ and $\phi \rightarrow \forall x\phi$. Modal deduction, either axiomatic or in other proof formats (sequents, natural deduction), is simple reasoning in perspicuous notation.

Modal logic games

Not intrinsic to modal logic, but a pleasant dynamic trend is this. All our notions have fine-structure as *games*. In an *evaluation game*, players Verifier (V) and Falsifier (F) disagree about a formula. Disjunction is a choice for V, conjunction for F, negation is role switch, $\Diamond$ makes V pick a successor of the current world, $\Box$ does the same for F. A game p is won by Verifier if the atom p holds in the current state, otherwise by Falsifier. A player also wins the game if the other player must move, for a modality, but cannot.

FACT **M**, $s \models \phi$ iff Verifier has a *winning strategy*
for the ϕ-game in **M** starting from s.

For example, our first model picture induces the following game tree for formula $\Diamond\Box\Diamond p$ starting from state 1, with bold-face indicating the winning positions for Verifier (figure 26.3).

In this game, V has two winning strategies: 'left,' and 'right,' ⟨'right,' 'down'⟩. These are indeed the two possible successful ways of verifying formula $\Diamond\Box\Diamond p$ in the given

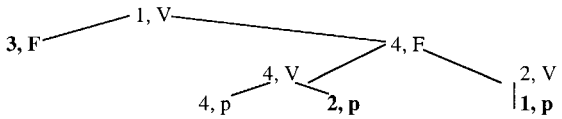


Figure 26.3

model **M** at world 1. Likewise, there are *model comparison games* between Duplicator (maintaining an analogy) and Spoiler (claiming a difference), playing over pairs (m, n) in two models **M**, **N**. These provide a fine-structured way of checking for the earlier bisimulation. In each round Spoiler chooses a state x in one model which is a successor of m or n, Duplicator responds with a corresponding successor y in the other model. If x, y are different in their atomic properties, Spoiler wins – if Duplicator cannot find a matching successor: likewise.

For example, in the non-bisimulation example **N**, **K** in figure 26.2, starting from a match between the two black worlds, Spoiler needs 3 rounds to win: forcing Duplicator in 2 rounds into a match where one world has no successor, while the other does.

FACT (a) Spoiler's winning strategies in a k-round game between **M**, s, **N**, t match the modal formulas of *operator depth* k on which s, t disagree. (b) Duplicator's winning strategies over an *infinite* round game between **M**, s , **N**, t match the bisimulations between them linking s to t.

One winning strategy for Spoiler in the preceding example exploits the earlier 'difference formula' $\Diamond\Diamond\Box\bot$. Many other logical notions can be 'gamified'. In particular, there are *construction games* determining if a given formula has a model, or *proof games* finding a derivation of it through a dialogue between two players.

Decidability and complexity

The basic modal language is a *decidable* 'miniature' of first-order logic. There are many decision methods for validity or satisfiability, exploiting special features of modal formulas – each with their virtues in generalization. Well-known methods are 'selection,' 'filtration,' and 'reduction.'

The deeper underlying issue is the precise computational *complexity* of various key tasks for a logic. These include not just satisfiability or validity testing, but also model checking and model comparison. Here are some landmark observations.

MODEL CHECKING Given a finite model **M**, s and a modal formula ϕ, checking if **M**, s $\models$ ϕ takes polynomial time in length(ϕ) + size(**M**) .

This is better than for first-order logic, where the same task takes polynomial space.

SATISFIABILITY Checking if a given modal formula has a model takes *polynomial* space in the size of the given formula.

For propositional logic the same task takes just non-deterministic polynomial time. For the full first-order language, of course, it is undecidable.

MODEL COMPARISON Checking if there is a bisimulation between given finite **M**, s, **N**, t takes polynomial time in the size of these models.

This may look surprising, but simple algorithms exist throwing out successive pairs of worlds that cannot make any bisimulation. These benchmark complexity outcomes may differ as modal languages are varied, allowing us to detect 'jumps.' Complexity awareness is a new feature of increasing importance in logic.

Model theory

The model theory of basic modal logic is much like that of first-order logic: with classical highlights such as Craig interpolation, Los–Tarski preservation theorem for universal modal formulas, etc. The analogy gets lost for many special modal logics, where for example interpolation is much scarcer.

Translation

Modal operators behave much like first-order quantifiers. The following translation T takes all modal formulas ϕ to first-order formulas $T(\phi)$ with one free variable x having the same truth conditions on models **M**, s:

(a) $T(p) = Px$,
(b) T commutes with all the Boolean operators,
(c) $T(\Diamond\phi) = \exists y\,(Rxy \,\&\, [y/x]T(\phi))$, $T(\Box\phi) = \forall y\,(Rxy \rightarrow [y/x]T(\phi))$

With some care, only two variables x, y are needed in all these first-order translations (free or bound). E.g. $\Box\Diamond\Box p$ translates faithfully into the formula

$$\forall y\,(Rxy \rightarrow \exists x(Ryx \,\&\, \forall y(Rxy \rightarrow Py))).$$

Here is the essential semantic feature that makes these translated modal formulas special inside the full first-order language over R^2, P^1, Q^1, . . .

MODAL INVARIANCE THEOREM The following assertions are equivalent for all first-order formulas $\phi = \phi(x)$: (a) ϕ is equivalent to a translated modal formula, (b) ϕ is invariant for bisimulations.

The 'modal fragment' is a small fragment of FOL, sharing its 'nice' properties, but remaining decidable. What you get for free on this view are 'universal' properties of first-order formulas, such as the Löwenheim–Skolem Theorem. Not for free is, for example, the Interpolation Theorem: modal consequences might have non-modal first-order interpolants: honest work is required to show that indeed *modal* interpolants exist. The fragmentist perspective is general: many other modal languages live inside first-

order logic or other standard logics, under some translation transcribing their standard semantics. We will see later what makes these fragments so well-behaved.

Landscapism

On top of the minimal logic, there are uncountably many different 'modal logics.' This landscape has two major highways: because of this

THEOREM Every normal modal logic is either a subset of the logic *Id* (with characteristic axiom $\phi \leftrightarrow \Box\phi$) or of *Un* (axiom $\Box\bot$).

On the former road lie the usual systems like T, S_4, S_5, on the latter, for example, 'Gödel–Löb logic' of arithmetical provability axiomatized by $\Box(\Box\phi \rightarrow \phi) \rightarrow \Box\phi$. Modal logics in this landscape can be studied by proof-theoretic or semantic methods, with a flourishing industry of completeness theorems providing bridges between the two.

Completeness

A typical modal completeness theorem runs like this.

THEOREM A formula is provable in K4 (K plus the axiom $\Box\phi \rightarrow \Box\Box\phi$) iff it is true in all models whose accessibility relation is *transitive*.

There are many techniques for proving such results, ranging from simple inspection of the Henkin model of all complete theories in the logic to drastic 'model surgery.' The motivation for completeness theorems can come from two directions. Either one has a pre-existing logic given by axioms and rules (such as the above cases), and seeks a useful corresponding model class – or one has a natural model class (say, some interesting space–time structure), and wishes to axiomatize its modal validities for 'simple reasoning.' The literature is replete with both. In this survey, we do not pursue this completeness line, since it gets so much exposure anyway.

Correspondence

The preceding correspondence between modal axioms and properties of the accessibility relation is a major attraction of modal logic. It can also be studied directly in the semantics, calling a modal formula true in a *frame* (a model stripped of its valuation) if it holds under all valuations. This line of research has produced two key results of a model-theoretic nature:

THEOREM A modal formula defines a first-order frame-property iff it is preserved under taking 'ultrapowers' of frames.

THEOREM A first-order frame-property is modally definable iff it is preserved under taking (a) 'generated subframes,' (b) 'p-morphic frame images,' (c) 'disjoint unions,' and (d) 'inverse ultrafilter extensions.'

Known non-first-order modal principles are the McKinsey Axiom $\Box\Diamond p \rightarrow \Diamond\Box p$, and the earlier-mentioned Gödel–Löb Axiom. Useful in practice is the *Sahlqvist Theorem*, describing an effective method for constructing first-order equivalents for most widely used modal axioms, which has by now reached the world of automated theorem proving. It proceeds by substituting first-order descriptions of 'minimal valuations' into a modal axiom to get a natural first-order equivalent (if available).

EXAMPLE The above K4 axiom $\Box p \rightarrow \Box\Box p$ has a first-order translation $\forall y\,(Rxy \rightarrow Py) \rightarrow \forall y\,(Rxy \rightarrow \forall z\,(Ryz \rightarrow Pz))$. A minimal valuation for p making the antecedent true is $Pu := Rxu$. Substituting this into our formula, and dropping the then tautologically true antecedent, we are left with a consequent of the syntactic form $\forall y\,(Rxy \rightarrow \forall z\,(Ryz \rightarrow Rxz))$, which is precisely frame *transitivity*.

In a sense this whole mathematical theory is a study of simple modal fragments of the complex realm of *second-order logic*, a perspective we will not pursue here.

The basic modal language has limited expressive power. But it has been the main mathematical laboratory for notions, techniques, and results. In what follows here, we look at some modern extensions, and new basic issues to which these give rise.

4 The Major Applications

Contemporary 'applications' of modal logic are not routine uses of existing notions and techniques: they add things not dreamed of in the original framework. This short article cannot really do justice to the variety of developments here. Here are some major directions that are arguably most influential in the 'drift' of the field.

Epistemic logic

Propositional attitudes like knowledge show logical behavior like that of ontological modalities. In particular, the epistemic operator

$K_i\phi$ 'agent **i** knows that ϕ' is like a universal modality stating that ϕ is true *in all of i's epistemically indistinguishable situations*.

And the same is true to some extent for other epistemic propositional attitudes, such as 'belief.' On this view, accessibilities are often equivalence relations for each agent – though alternatives exist. Languages like this express basic epistemic statement patterns that we often use in natural discourse, such as

$K_i\phi \vee K_i\neg\phi$ 'agent i knows whether ϕ is the case'

and modal axioms acquire a new flavor:

$K_i\phi \rightarrow K_i K_i\phi$ 'positive introspection'
$\neg K_i\phi \rightarrow K_i\,K_i\neg\phi$ 'negative introspection'

But the major new theme in this epistemic setting is a 'social one.' It is not the Lonely Thinker that is essential to understanding cognition, but *interaction* between different agents in a *group:* $K_i K_j \phi$ or $K_i \neg K_j \phi$. What I know about your knowledge or ignorance is crucial, both to my understanding and to my actions. (For example, I might empty your safe tonight if I think you don't know that I know the combination.) Some forms of 'group knowledge' even transcend simple iterations of individual knowledge assertions. The central example here is *common knowledge*: if everyone knows that your partner is unfaithful, but no more, you have private embarrassment – if it is common knowledge, you have public shame and perhaps a need for violent action. Technically, common knowledge works as follows:

$C_G\phi$ ϕ holds at every world reachable via any finite chain of uncertainty relations for actors in G.

For example, in the picture in figure 26.4, where p holds in the current world (the black dot), in the black world, (a) agent Q does not know if p is the case: $\neg K_Q p \ \& \ \neg K_Q \neg p$; (b) agent A does know that p is the case: $K_A p$; while (c) it is common knowledge in the group {Q, A} that A knows *whether* p is the case: $C_{\{Q, A\}}\ (K_A p \vee K_A \neg p)$. Incidentally, this might be a good situation for Q to ask A a *question* about p: but more on epistemic *actions* below.

Dynamic logic

Accessibilities can also be viewed as *transitions* for actions that change states. In 'dynamic logic' – originally designed to describe the execution of computer programs, but now used as a general logic of action, we have

$[\pi]\phi$ says that *after every successful execution of action* π, ϕ *holds.*

Thus, modal statements relate actions to 'postconditions' describing their effects (and also to 'preconditions' for their successful execution). A concrete model of this are *games*, where actions are moves available to players. For example in the tree shown in figure 26.5, player E has a *strategy* for achieving an outcome satisfying p.

This strategic assertion is captured by the dynamic modal formula $[a \cup b]\langle c \cup d\rangle p$. Again we get a minimal modal logic for universal validity here, this time set up as a *two-level system* treating propositions and actions (transition relations) on a par. This joint set-up allows for a logical analysis of important action constructions, encoded in valid principles for general operations as well as specific actions:

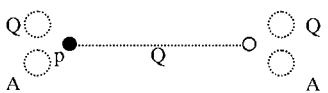


Figure 26.4

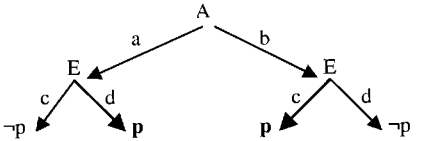


Figure 26.5

$[\pi;\pi']\phi \leftrightarrow [\pi][\pi']\phi$	sequential composition
$[\pi \cup \pi']\phi \leftrightarrow [\pi]\phi \,\&\, [\pi']\phi$	choice
$[(\phi)?]\psi \leftrightarrow (\phi \rightarrow \psi)$	test for proposition ϕ

A major new feature here is unbounded finite *repetition* of actions: π^*. This is typical for computation, and it is not first-order definable. This shows in axioms

$[\pi^*]\phi \leftrightarrow (\phi \,\&\, [\pi][\pi^*]\phi)$	fixed-point axiom
$(\phi \,\&\, [\pi^*](\phi \rightarrow [\pi]\phi)) \rightarrow [\pi^*]\phi$	induction axiom

Thus, dynamic logics resemble *infinitary* fixed-point extensions of classical logic, but they do retain the 'modal stamp': being bisimulation-invariant, and decidable. Fixed-point definitions are ubiquitous in computer science, but also in mathematics or linguistics, because many natural notions involve a kind of 'implicit' recursion. An elegant current system of this kind for actions is a generalization of dynamic logic allowing arbitrary fixed-point definitions: the so-called 'μ-calculus.'

Temporal and spatial logic

A more traditional, but very lively application area of modal logic concerns 'physical' rather than 'human' nature. We mention this as a counterpoint to our cognitive slant. One concrete interpretation of modal models is as *flows of time*, accessibility being 'earlier than.' The universal modality will then say 'everywhere in the future,' which comes with an obvious dual 'everywhere in the past.' Temporal logics are prominent in computer science and AI, where they show a great diversity beyond this basic modal point of departure. In particular, they can live over different primitive entities: duration-less points, or extended 'periods.' Usually, the vocabulary of temporal languages is much richer than the basic modal language. A typical example are operators allowing us a view of what goes on *during* the successful execution of a program or plan:

UNTIL $\phi\psi$ *at some point later than now* ϕ *holds,*
while at all intermediate points ψ *is true*

In this same physical arena, modal logics of *space* are also gaining importance, for example in knowledge representation. One of these revives an old mathematical idea. Let our models be topological spaces endowed with a valuation. Then the modality

$\Box\phi$ may be read as saying that the current point lies in the *topological interior* of the set $[[\phi]]$ of all points where ϕ holds.

Then, modal laws come to encode various topological facts about space, for example:

$\Box(\phi\&\psi) \leftrightarrow \Box\phi \,\&\, \Box\psi$ says that open sets are closed under intersections.

This style of analysis may be extended to modal fragments of geometry. It provides an alternative to our standard semantics quantifying over successors in some binary world-to-world relation. (Technically, it is a 'neighborhood semantics,' of a sort developed in the 1960s to explore landscapes below the minimal modal logic K.) Thus our spatial excursion also shows that the 'standard approach' is not sacrosanct.

AI, linguistics, mathematics

Modal logic has either been applied, or rediscovered, in such areas as artificial intelligence ('description languages,' 'context logics'), linguistics ('categorial grammar,' 'feature logics'), and indeed mathematics, with flourishing areas such as 'provability logic,' and in recent years also modal versions of set theory. This list is not complete (intuitionistic logic or relevant logic or linear logic are also similar in some of their key features), but it does show that modal structures occur naturally across a wide range of disciplines.

5 Fine-Structure of Expressive Power

Modal logic today shows several new general themes that cut across these various applications. We mention a few, though there is certainly no consensus on a simple synthesis out of the current research scene. One is *extension of expressive power*.

Logical extensions

Modal languages can be enriched over their original models. A popular 'logical extension' of this sort adds a *universal modality*

$U\phi$ saying that *ϕ is true at all worlds*, accessible or not.

This gives more expressive power, which one can use to state 'global facts,' such as the inclusion of one region of the model in another. But our standard techniques generalize, for example, the language of $\{\Box, U\}$ matches up with 'total bisimulations,' whose domains and ranges are the whole models being compared. And also: its minimal logic remains decidable – though the complexity of validity goes up to exponential time. (When added to more complex languages, indeed, U may push a decidable logic over the brink into undecidability.) In earlier years, extending the basic modal language was 'not done,' because it would change the rules of the game, and make life too easy. Here is another example. Having *names for specific worlds* would be a great convenience, both

in practice and in the modal metatheory, but the basic language does not allow it. For example, much has been made of the latter's inability to express the frame property of *irreflexivity* ($\forall x \neg Rxx$). But this is expressed quite simply by the following axiom in an extended modal language:

$i \rightarrow \neg \Diamond i$ where the 'nominal' i is a special proposition letter ranging over only *singleton sets* of worlds.

Nowadays the tendency is to add such devices freely, only subject to striking a good balance between increased expressive power and manageable complexity. Another example is the above operator 'Until' of temporal logic, where inhibitions as to enrichment have always been weaker. What keeps these extensions 'modal' is that they allow for bisimulation analysis, while staying decidable. Much is known by now about which added operator leads to which jump in decidable complexity for our benchmark tasks of satisfiability, model checking, and model comparison.

'Geometrical' extensions

By contrast to the preceding move, 'geometric extensions' enrich the similarity type of our models, adding modalities with *new accessibilities*, as in epistemic or dynamic logic, or in *polyadic* modal languages with n-ary alternative relations. For example an existential 'dyadic modality'

$\Diamond\phi\psi$ holds at s iff $\exists t, u$ *s.t.* R^3stu, ϕ *holds at* t, ψ *holds at* u

Concrete interpretations for such ternary accessibility relations R include:

s is the concatenation of two expressions t, u,
s is the merge of the two resources t, u.

Guarded fragment

One limit to which many extensions of both types tend is the so-called *Guarded Fragment* of first-order logic. This is defined inside the full first-order syntax by allowing only quantifications of the 'guarded' form

$\exists \mathbf{y}(G(\mathbf{x}, \mathbf{y}) \,\&\, \phi(\mathbf{x}, \mathbf{y}))$

where $\mathbf{x}$, $\mathbf{y}$ are tuples of variables, $G(\mathbf{x}, \mathbf{y})$ is an atomic formula whose variables occur in any order and multiplicity, and ϕ is a guarded formula having only variables from $\mathbf{x}$, $\mathbf{y}$ free. Many modalities are guarded in this syntactic sense:

$\Diamond p$ $\quad \exists y(Rxy \,\&\, Py)$
$\Diamond pq$ $\quad \exists yz(Rxyz \,\&\, Py \,\&\, Qz)$

This sublanguage of first-order logic, where groups of objects are only introduced 'under guards' still yields to modal analysis supporting a 'nice' meta-theory.

THEOREM The Guarded Fragment has a characteristic bisimulation.
THEOREM The Guarded Fragment is decidable in doubly exponential time.

These properties even transfer to certain extensions. Another interesting property exemplified in this setting is *robust decidability*: small modal languages sometimes bear the weight of expressive extensions that otherwise explode reasoning complexity. An example are fixed-point operators for inductive definitions. On top of first-order logic, these make the resulting language non-axiomatizable – when added to the Guarded Fragment, however, they do not increase complexity at all.

Two dimensions

The earlier 'landscape' of modal logic was really one-dimensional: it kept the basic language constant in expressive power, varying deductive strength of special theories expressed in it. But now we have a second dimension: systematic variation of expressive power. This new two-dimensional landscape has many 'thresholds of complexity' which are currently being charted.

6 System Combination: Action and Information

Other main themes in general modal logic today are *many agents*, *dynamics*, and *system combination*. The former has already occurred in our survey. As to the latter, many applications are 'multi-modal,' putting together various modal logics in one system: say of action, knowledge, and time. There are several ways of doing this, ranging from mere 'juxtaposition' to more intricate forms of interaction between the component logics. One then wants to predict expressiveness and complexity of the combination from those of its parts – plus the mode of combination used. There is an incipient general theory of relevant modes of combination, including new constructions of 'product' and 'fibering.' This style of thinking even shows in modern technical views of *modal predicate logic*. One can 'deconstruct' this famous system into a combination of two modal logics: a static one of world accessibility, and a dynamic one of object assignment to variables. The main challenge arising then are the unpredictable effects of various combinations. Disregarding further generalities concerning composition of logics, we describe two rather exciting recent special combinations of long-standing modal ideas.

Information update

Models of epistemic logic serve as information states for groups of agents. Epistemic formulas are then evaluated against worlds in such states, telling us what is true or not in them. But knowledge usually functions in *communication*: it is conveyed to others via speech acts, and influenced by theirs. To model such *cognitive actions*, we need to combine two earlier systems: epistemic logic and dynamic logic. In particular, a communicative action changes the current epistemic model! In the simplest case, this 'update' works as follows:

public announcement of a proposition ϕ to a group of agents eliminates all worlds in the current model $\boldsymbol{M}$ that satisfy ϕ

Suppose that in our earlier two-agent two-world picture Q asks A : "p?" and A then truthfully answers 'Yes.' Then the –p-world gets eliminated, and we are left with a one-world model where p has become common knowledge among {Q, A}. But more subtle cases are possible, even with very simple models of this sort. For example, a question itself may convey crucial information! Suppose that, by asking, Q conveys the information that she does not know the status of p. Even if A did not know the answer at the start, this may tell him enough to settle p, and thereby answer the question. Figure 26.6 shows one scenario where this happens.

But the modeling power of combined epistemic dynamics is still higher. For example, suppose neither Q nor A knew about p, but A asks expert R, who answers only to Q. Then A learns whether p, Q is no wiser about p, but it has become common knowledge that A knows whether p. This requires 'arrow elimination' (figure 26.7).

These simple pictures hide delightful subtleties. For example, one may check that, on this account, a public announcement that some formula ϕ is the case need not always result in our 'learning that ϕ', in the form of an updated model where ϕ holds! (For, truth value switches may happen when we process an announcement of *ignorance*.) Precise algorithms for performing updates associated with communicative acts, public or private, have been proposed in recent years – and these provide an entirely new use of our 'standard models.' Eventually, in this view of communication, one wants to describe not just information update, but also actions of 'withdrawal' or *revision*, triggered by propositions that contradict the content of our current information state. These cognitive actions require modal logics of *counterfactual conditionals* that feed into modern *belief revision theory*.

Game logics

There can be more than one attractive way of putting modal ideas together. Another interesting mix of 'epistemics' and 'dynamics' occurs in the analysis of *games*. As players move through a game tree, their information changes. Plain game trees are described in dynamic logic, as we saw in an earlier section, though realistic reasoning

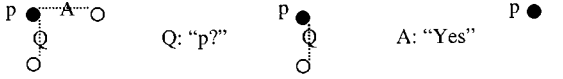


Figure 26.6

Figure 26.7

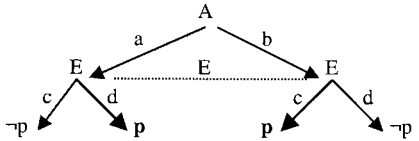


Figure 26.8

about future game actions also require a logic of players' *preferences*. Especially interesting, however, are *imperfect* information games, where players may not know the precise moves played by their opponents. Thus, in these games, the primary epistemic uncertainty is *between actions*, and only in a derived sense between the resulting game states. (Think of a card game where we cannot observe which initial hand Nature is dealing to our opponent, or where some mid-play moves by our opponents may be partially hidden.) An informative example is the earlier game tree, but now with an uncertainty link for player E at the second stage – she does not know the precise opening move played by A (figure 26.8).

We can view this as a model for an obvious combined dynamic-epistemic language, having both epistemic modalities K_i and dynamic ones [a], which may interact. In particular, half-way, player E knows 'de dicto' that she has a winning move

$$K_E(\langle c\rangle p \vee \langle d\rangle p)$$

but she does not know any particular winning move 'de re':

$$\neg K_E\langle c\rangle p \;\&\; \neg K_E\langle d\rangle p!$$

Indeed, this game is 'non-determined' in a natural sense: E cannot force an outcome p, but neither can A force outcome ¬p. The general logic of these game trees is the minimal propositional dynamic logic plus epistemic 'multi-S5.' But on top of that, the combined dynamic-epistemic language can also express *modes of playing games*. Take the game-theoretic notion of 'Perfect Recall.' This describes players whose *own* actions never introduce any uncertainties that they did not have before. Properly understood, this validates an interchange axiom

$$(turn_E \;\&\; K_E[a]\phi) \rightarrow [a]K_E\phi:$$

what we know about the result of our own game moves is still known to us after we perform them. (To understand this better, contrast the effects of non-'epistemically neutral' actions like drinking genever.) Thus, we can correlate modal logics in this epistemic-dynamic language with special styles of playing a game. Another mode is 'Bounded Memory' – whose treatment requires a universal modality. This simple example also illustrates a general point. Games are a nice target for logical analysis because they show cognition at work under well-defined 'laboratory circumstances.'

7 Back to the Heartland

Modal logic started as an epicycle on standard logic. And it is still viewed by most people as a 'nonstandard' topic beyond The Core. But latterly, it has started to influence the heartland itself. We conclude with two examples of this 1990s trend.

Modal foundations of predicate logic

Predicate logic *itself* is a form of modal or dynamic logic! The key truth condition for the existential quantifier reads

M, s $\models \exists x\phi$ iff there exists d in D^M s.t. **M**, s[x:=d] $\models \phi$

This has the modal pattern for evaluating an existential modality $\langle x \rangle$:

M, s $\models \exists x\phi$ iff there exists t s.t. $R^x st$ with **M**, t $\models \phi$

Viewed in this light, the usual set of 'valid laws' of first-order logic can be deconstructed into several layers: (1) Its decidable(!) core is the minimal modal logic, which contains such laws as Monotonicity: $\forall x(\phi \rightarrow \psi) \rightarrow (\forall x\phi \rightarrow \forall x\psi)$. This level makes no presuppositions whatsoever concerning the form of the models, which could have any kind of 'states' and 'variable shifts' R^x. (2) Next, there are laws recording universal effects of taking variable assignments for states, plus the special shift relation of 'agreeing up to the value for x.' For example $\forall x\phi \rightarrow \forall x\, \forall x\phi$ expresses the *transitivity* of R^x: indeed, all of S_5 holds. (3) Most 'specifically', some first-order laws express *existence* properties for states. Here is an example:

$\exists x\, \forall y\phi \rightarrow \forall y\, \exists x\phi$ expresses *confluence*: whenever s R^x t and s R^y u, then there also exists a state v s.t. t R^y v and u R^x v (figure 26.9).

Thus, modal analysis reveals unexpected 'fine-structure' in the class of what is usually lumped together as 'standard validities': they are valid for different reasons!

Moreover, on our general modal models, the predicate-logical language gets increased expressive power, because new distinctions come up. For example:

polyadic quantifiers $\exists xy\bullet$

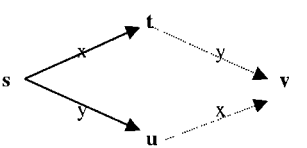


Figure 26.9

introducing two objects becomes different from iterations $\exists x \exists y \bullet$ or $\exists y \exists x \bullet$.
Summing up, we get a highly unorthodox view. The 'modal core' of standard logic is decidable, *pace* Church and Tarski – but piling up special (existential) model conditions makes state sets behave so much like *full function spaces* D^{VAR} that their total logic becomes undecidable.

Dynamic predicate logic

Another dynamic view on first-order logic rather emphasizes the *state change* implicit in evaluating an existential quantifier. We move to a new state containing a suitable 'witness value' for x. More generally, one can let first-order formulas denote *actions of evaluation*:

(a) atomic formulas are *tests* if the current state satisfies the relevant fact,
(b) an existential quantifier picks an object and assigns it to x (*random assignment*),
(c) a substitution operator [t/x] is a *definite assignment* x:=t,
(d) a conjunction is sequential action *composition*,
(e) a negation $\neg\phi$ is a test for the *impossibility* of succesfully executing the action ϕ.

The resulting 'dynamified' version of first-order logic has applications in the semantics of natural language – as anaphoric pronouns 'he,' 'she,' 'it,' show this kind of dynamic behavior. One nice illustration occurs with sentences like

$\exists x\, Kx \rightarrow Hx$ 'if you get a kick, it hurts'

The standard logical folklore must 'improve' natural language here to arrive at the universal first-order form $\forall x\, (Kx \rightarrow Hx)$. But with dynamic semantics, this meaning arises automatically, as any value assigned by the existential move in the antecedent will be bound to x when the consequent is processed. This system has also inspired programming languages for dynamic execution of specifications.

'Dynamic predicate logic' exemplifies a general paradigm of bringing out the implicit cognitive dynamics which underlies existing logical systems. This allows one to view natural language meanings in terms of updates of propositional content, perspective, and other parameters that determine the transfer of information.

8 Conclusion

This survey is different in spirit from standard wisdom in philosophical logic. We have presented modal logic as a tool for fine-structure analysis of the expressiveness and complexity of logical languages, including effects of their combinations, and the major applications (information, action) that drive abstract theory today. There is no uniform conclusion, or even a new definition of modal logic in the end: the modern field is just too rich for that. Our purpose with this short article will have been served if the reader experiences a culture-shock, seeing the differences between reality and the picture still painted by many 'standard textbooks.'

References

Technical surveys

Bull, R. and Segerberg, K. (1984) Basic modal logic. In D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic* (pp. 1–88). Dordrecht: Reidel.

The chapters on modal logic by M. Fitting ('Modal logic') and C. Stirling ('Modal and temporal logics') in S. Abramsky, D. Gabbay and T. Maibaum (eds.) (1991) *Handbook of Logic in Computer Science*. Oxford: Oxford University Press.

Benthem, J. van (1991) *Manual of Intensional Logic*. Stanford: CSLI Publications.

Chagrov, A. and Zakharyashev, M. (1997) *Modal Logic*. Oxford: Clarendon Press.

Consecutive volumes in the recent series *Advances in Modal Logic*. Dordrecht: Kluwer.

Modern textbooks

Popkorn, S. (1992) *First Steps in Modal Logic*. Cambridge: Cambridge University Press.

Blackburn, P., de Rijke, M. and Venema, Y. (2000) *Modal Logic*. Cambridge: Cambridge University Press.

Modal predicate logic

M. Fitting's chapter in this volume.

Temporal logic

Benthem, J. van (1996) Temporal logic. In D. Gabbay, C. Hoggar and J. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming* (vol. 4, pp. 241–350). Oxford: Oxford University Press.

Gabbay, D., Hodkinson, I. and Reynolds, M. (1994) *Temporal Logic*. Oxford: Oxford University Press.

Epistemic logic

Fagin, R., Halpern, J., Moses, Y. and Vardi, M. (1995) *Reasoning about Knowledge*. Boston, MA: MIT Press.

Dynamic logic

Goldblatt, R. (1986) *Logics of Time and Computation*. Stanford, CA: CSLI Publications.

Harel, D., Kozen, D. and Tiuryn, J. (2000) *Dynamic Logic*. Boston, MA: MIT Press.

Provability logic

Boolos, G. (1993) *The Logic of Provability*. Cambridge: Cambridge University Press.

Modal set theory

Barwise, J. and Moss, L. (1999) *Vicious Circles*. Stanford, CA: CSLI Publications.

Philosophical foundations

Zalta, E. (1993) A philosophical conception of propositional modal logic. *Philosophical Topics*, 21, 263–81.

Modal grammar

Blackburn, P. and Meyer-Viol, W. (1997) Modal logic and model theoretic syntax. In M. de Rijke (ed.), *Advances in Intensional Logic* (pp. 29–60). Dordrecht: Kluwer.

Chapters on feature logic by W. Rounds ('Feature logics') and on categorial grammar by M. Moortgat ('Categorical type logics') in J. van Benthem and ter Meulen, A. (eds.) (1997) *Handbook of Logic and Language*. Amsterdam: Elsevier.

Modal logic and AI

See the Electronic Discussion Newsletter, *Reasoning about Action and Change*: Linköping, http://www.ida.liu.se/ext/etai/rac/notes/.

Modal logic and game theory

Benthem, J. van (1999) *Logic and Games*. Lecture notes. Amsterdam and Stanford.

Logical dynamics

Benthem, J. van (1996) *Exploring Logical Dynamics*. Stanford, CA: CSLI Publications.

Gerbrandy, J. and Groeneveld, W. (1997) Reasoning about information change. In J. van Benthem and Y. Shoham (eds.), *Cognitive Actions in Focus, Journal of Logic, Language and Information*, vol. 6 (pp. 147–69).

Baltag, A. (1999) *A Logic of Communication*. Amsterdam: Centre for Mathematics and Computer Science.

Ditmarsch, H. van (2000) Knowledge games. Dissertation. Groningen: Department of Computer Science and Institute for Logic, Language and Computation, Amsterdam.

27

First-Order Alethic Modal Logic

MELVIN FITTING

1 Introduction

Propositional modal logic, with its possible world semantics, is now a standard part of a philosophical education, while first-order modal logic is less familiar. But there are several well-known problematic concepts that can be made more intelligible using a first-order modal semantics; among these are existence, designation, identity, synonymy, intension, and extension. I will address these and other issues. I will assume a general familiarity with propositional possible world semantics, and begin at the quantificational level; (Hughes and Cresswell 1996) is a standard reference. I will not attempt to move from semantics to proof procedures, length precludes that, but (Fitting and Mendelsohn 1998) contains tableau systems that are appropriate for what is presented here.

I should begin by saying something about the status of possible worlds. It is sometimes asked what they are, or even where they are. These are the wrong questions to be asking. Consider classical logic, for a moment. To say a formula Φ is valid is to say it is true in all models. One does not inquire where these models come from – we are talking formal mathematics, and they exist in the same sense that any mathematical structure exists. (I grant that questions of mathematical existence can be tricky too, but they are not what concern us now.) In addition, we occasionally apply classical logic to the actual world – we extract a formal model from 'reality.' When we do so we must stipulate the domain of quantification. This amounts to specifying what the 'things' of the real world are. Do they include numbers? Do they include concepts like beauty? Applying classical logic to the real world is not as straightforward as we often make it seem, but nonetheless, we do it.

Modal models involve possible worlds. Generalizing from classical logic, a formula is taken to be valid if it is true no matter what the domain and no matter what the interpretation of symbols, *and no matter at what possible world of a model we evaluate the formula.* This is a formal definition, just as in the classical case. Possible-world models are mathematical structures too.

We still must deal with the desire to apply modal notions in the actual world. The problem is much like that of applying classical logic to the actual world but now, in addition to stipulating domains and interpretations, we must also stipulate possible

worlds. They are not 'out there' to be found with a telescope. Intuitively, they represent how things might have been, and to a considerable extent, this is up to us. Is a situation in which Julius Caesar was a bottle of salad dressing really a way things could have been, or not? It does not seem to me that such a question has an answer independent of the asker, just as whether beauty is in the range of a quantifier or not probably depends on who is using the quantifier, and for what purpose. In short, as a piece of mathematics, possible world semantics is on the same footing as all mathematics. As a way of understanding discourse about the real world, the semantics goes a long way towards clarifying things, but there is considerable ambiguity or, if you prefer, flexibility.

In what follows I will sometimes be describing formal models, mathematical structures. But sometimes I will be using possible-world semantics informally, with some intuitive notion of possible worlds which I assume is sufficiently understood by both me and the reader to make the discussion mutual. In such situations, the real world will generally be assumed to be among the possible worlds, and the quantification domain will be assumed to include at least all real things. But keep in mind the discussion above as to what is a real thing. By keeping the discussion imprecise I am, in effect, allowing for a variety of different ways of understanding everyday modal discourse in terms of possible-world semantics.

2 Intensions

Let us say an adult is someone 21 or older. The property of being an adult has a certain *extension*: the set of people who are, in fact, 21 or older. At other times, or under imagined circumstances, the same property will have a different extension. The *intension* of the property is, in some indefinite sense, its meaning, and so determines its extension under various circumstances. Trying to formalize meaning is a formidable task, and reasonable people can differ about how this should be done. The common denominator among all such attempts is: the intension of a property should determine its extension, in every circumstance. If we ignore the issue of how, intensions simply become maps from situations to extensions.

In addition to properties, we also need to treat individuals and individual concepts. The number 9, and Bertrand Russell, are individuals, or individual objects. The number of the planets, or the junior author of the *Principia*, are individual concepts. As things are, they designate 9 and Bertrand Russell respectively, but under other circumstances they might not have done so. Once again, some notion of meaning is involved. And once again, however that notion of meaning is understood, an individual concept will associate an individual object with each circumstance. Formalized it will simply be a map from situations to objects.

This leads to the beginnings of a formal treatment – (Fitting and Mendelsohn 1998) contains a fuller version of what follows. I'll assume we have a first-order modal language with *relation symbols* of various arities. The equality symbol, =, is among them. (Since it's what we're used to, I'll write = in the conventional infix position.) There will also be *constant symbols* – typically c, d, ... And there will be *variables* – typically x, y, ... Relation symbols will be used to represent properties in *intension*, and constant

symbols will be used to represent *individual concepts*. Intensions determine extensions, which are sets of objects, and likewise individual concepts determine objects, so we need machinery for dealing with objects as well. I'll assume variables have individual objects as values. For an atomic formula, say $P(c)$, it will be taken to be true at a possible world if the individual object designated by c at that world is in the extensional property designated by P at that world.

There is yet one more piece of machinery that must be introduced, and it will be less familiar. In *classical* logic, if $\Phi(x)$ is a formula, we can think of it as determining the extensional property of being something that makes Φ true. But now we are trying to think intensionally. Using $\Box$ for the necessity symbol and P as a one-place relation symbol, how should we understand the formula $\Box P(x)$, that is, what *intensional* property does it determine? In particular, for a constant symbol c, how should we read the formula $\Box P(c)$? Should it be taken to say c has the P property necessarily, or that c has the necessary-P property? These are *not* synonymous. Suppose, for instance, that c is the richest-person-in-the-world individual concept, and P is the intensional property being-wealthy (both notions change with changing circumstances). It seems likely that $P(c)$ is true under all circumstances – the richest person in the world, whoever that is, is wealthy, however we measure wealth. Then $\Box P(c)$ should be taken to be valid, since $P(c)$ is always true. On the other hand, while c designates the richest person in the world currently, that person might be poor under other circumstances, so we cannot say, of c, that the person has the necessarily-wealthy property. But then $\Box P(c)$ should not be taken to be valid. What is needed is some way of distinguishing between these two interpretations of the single formula $\Box P(c)$.

I'll make use of a device called *predicate abstraction*. If Φ is a formula and x is a variable, $\langle \lambda x.\Phi \rangle$ is a predicate abstract. If t is a term – either a constant symbol or a variable – and $\langle \lambda x.\Phi \rangle$ is a predicate abstraction, $\langle \lambda x.\Phi \rangle(t)$ will be counted as a formula. Then $\Box\langle \lambda x.P(x) \rangle(c)$ and $\langle \lambda x.\Box P(x) \rangle(c)$ are both formulas, and obviously different. The semantics introduced below will give them different readings, corresponding to the two readings of $\Box P(c)$ above.

Now the class of *formulas* can be specified. It is built up in more-or-less the usual way, using propositional connectives $\wedge$, $\vee$, $\supset$, $\equiv$, and $\neg$, modal operators $\Box$ and $\langle\rangle$ quantifiers $\forall$ and $\exists$, and predicate abstraction. I skip details, as they are quite straightforward to supply. For simplicity, I'll abbreviate formulas like $\langle \lambda x.\langle \lambda y.\langle \lambda z.\Phi \rangle(e) \rangle(d) \rangle(c)$ by $\langle \lambda x, y, z.\Phi \rangle(c, d, e)$.

3 Models

A *frame* is a structure $\langle \mathcal{G}, \mathcal{R} \rangle$, where $\mathcal{G}$ is a (nonempty) set of *possible worlds* and $\mathcal{R}$ is a binary relation on $\mathcal{G}$ of *accessibility*. Intuitively, one thinks of the members of $\mathcal{G}$ as representing the way things are, and the various ways they could be – possible situations, say. The accessibility relation tells us which situations are relevant to which. It is, by now, common knowledge that placing natural restrictions on $\mathcal{R}$ produces well-known modal logics. In many ways, S_5 is the simplest of the modal logics, and the most natural if $\Box$ is to represent metaphysical necessity. For S_5 $\mathcal{R}$ is simply the universal relation, the

one that always holds. In what follows, I'll assume this is my choice for $\mathcal{R}$. Certain things are simpler and more natural with such a choice, though much of what is said applies more generally – see (Fitting and Mendelsohn 1998).

Certainly if things were different, different things might exist. An *extended frame* is a structure $\langle \mathcal{G}, \mathcal{R}, \mathcal{D} \rangle$ where $\langle \mathcal{G}, \mathcal{R} \rangle$ is a frame and $\mathcal{D}$ is a *domain function* mapping $\mathcal{G}$ to nonempty sets. If $\Gamma \in \mathcal{G}$, think of $\mathcal{D}(\Gamma)$ as the set of objects existing in Γ. Also, by the *domain of the frame* I mean the union of the domains of the various possible worlds. If we understand possible existence to mean actual existence under other circumstances, the domain of a frame consists of those things having actual or possible existence in our formal setting.

A *model* is a structure $\langle \mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{I} \rangle$ where $\mathcal{I}$ is an *interpretation* in the extended frame $\langle \mathcal{G}, \mathcal{R}, \mathcal{D} \rangle$. The *domain of the model* is the domain of the underlying frame. The interpretation must meet three requirements. First, it should associate with every constant symbol c a mapping, $\mathcal{I}(c)$, assigning to each possible world some member of the domain of the model. Second, it should associate with every n-place relation symbol R a mapping, $\mathcal{I}(R)$, assigning to each possible world some n-place relation on the domain of the model. Third, it should associate with the equality symbol, $=$, the constant mapping assigning to each possible world the equality relation on the domain of the model.

The notion of interpretation captures the informal idea expressed earlier. Associated with each relation symbol is a relation in intension – a map from possible worlds to relations in extension. Likewise, associated with each constant symbol is an individual concept. Say we have a constant symbol c and a possible world Γ. $\mathcal{I}(c)$ is a function on possible worlds, so $\mathcal{I}(c)(\Gamma)$ is some member of the domain of the model. There is no requirement that it be something that exists at Γ; that is, it need not be in $\mathcal{D}(\Gamma)$. It makes perfectly good sense to talk about Pegasus, who exists in a mythological world even though he does not exist in ours. Similarly a relation symbol, at a world, is some relation in extension, but there is no requirement that things in that relation in extension actually exist at that world. If there were such a requirement, we would be unable to say that Pegasus has the property of being mythological.

4 About Quantification

If I claim that everything has a certain property, what am I claiming? I could mean everything that *actually exists* has the property (actualist quantification). I could mean everything that *does or could exist* has the property (possibilist quantification). In our formal semantics, actualist quantifiers, at a world Γ, range over the domain of that world, $\mathcal{D}(\Gamma)$. Possibilist quantifiers range over the domain of the model. Both are natural, but for different purposes.

Here, possibilist quantification will be taken as basic, because there is an easy way to define actualist quantification from it. Introduce a special one-place relation symbol, E, and interpret it at each world as the set of things that actually exist there – an existence predicate, in other words. Formally, in a model $\langle \mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{I} \rangle$, we will require that $\mathcal{I}(\mathsf{E})$ be the function that maps each possible world Γ to $\mathcal{D}(\Gamma)$. Further, introduce rela-

tivized quantifiers: $(\exists^{\mathsf{E}}x)\Phi$ abbreviates $(\forall x)[\mathsf{E}(x) \supset \Phi]$ and $(\exists^{\mathsf{E}}x)\Phi$ abbreviates $(\exists x)[\mathsf{E}(x) \wedge \Phi]$. Intuitively speaking (since the full formal semantics has not been fully specified yet), if $(\forall x)$ and $(\exists x)$ are read in a possibilist way, quantifying over the domain of the model, then $(\forall^{\mathsf{E}}x)$ and $(\exists^{\mathsf{E}}x)$ correspond to actualist quantification, with things restricted to world domains.

5 Truth in Models

Now comes the key definition: truth of formulas at possible worlds of a model. Simultaneously, the meaning of predicate abstracts must also be defined. Formulas can contain free variables, and so we need machinery for giving them values. A *valuation* is a mapping from free variables to the domain of a model. Note that valuations do not depend on possible worlds – free variables are supposed to represent objects, not intensions. If v is a valuation and $\mathcal{I}$ is an interpretation, between them they supply meanings for all terms. I'll use the following notation. For a possible world Γ,

(1) If x is a variable, $(v * \mathcal{I})(x, \Gamma) = v(x)$.
(2) If c is a constant symbol, $(v * \mathcal{I})(c, \Gamma) = \mathcal{I}(c)(\Gamma)$

Thus for any term t, $(v * \mathcal{I})(t, \Gamma)$ is the object associated with t at possible world Γ. I'll also use the following notation. If v is a valuation, x is a variable, and d is an object in the domain of the model, $v[x/d]$ is the valuation that is like v except that it maps x to d. And I'll say a formula is an *atom* if it is of the form $R(t_1, \ldots, t_n)$ where R is an n-place relation symbol and $t_1, \ldots, t_n$ are terms, or if it is of the form $\langle\lambda x.\Phi\rangle(t)$ where $\langle\lambda x.\Phi\rangle$ is a predicate abstract and t is a term.

Now, the fundamental notion to be defined is symbolized $\mathcal{M}, \Gamma \Vdash_v \Phi$ and is read: formula Φ is true at possible world Γ of model $\mathcal{M}$ with respect to valuation v. Simultaneously meanings are assigned to predicate abstracts. Here is the definition.

Let $\mathcal{M} = \langle \mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{I} \rangle$ be a model.

(1) For atoms, $\mathcal{M}, \Gamma \Vdash_v R(t_1, \ldots, t_n)$ if $\langle (v * \mathcal{I})(t_1), \ldots, (v * \mathcal{I})(t_n)\rangle \in \mathcal{I}(R)(\Gamma)$.
(2) $\mathcal{M}, \Gamma \Vdash_v (X \wedge Y)$ if $\mathcal{M}, \Gamma \Vdash_v X$ and $\mathcal{M}, \Gamma \Vdash_v Y$, and similarly for the other propositional connectives.
(3) $\mathcal{M}, \Gamma \Vdash_v \Box X$ if $\mathcal{M}, \Delta \Vdash_v X$ for every $\Delta \in \mathcal{G}$ such that $\Gamma \mathcal{R} \Delta$, and similarly for $\Diamond X$.
(4) $\mathcal{M}, \Gamma \Vdash_v (\forall x)X$ if $\mathcal{M}, \Gamma \Vdash_{v[x/d]} X$, for every x in the domain of the model $\mathcal{M}$, and similarly for $(\exists x)$.
(5) The interpretation $\mathcal{I}$ is extended to predicate abstracts as follows. $\mathcal{I}(\langle\lambda x.\Phi\rangle)$ is the map that assigns to possible world Γ the set $\{d \mid \mathcal{M}, \Gamma \Vdash_{v[x/y]} \Phi\}$.

The definition is technical, but the content is intuitive. Item 1 says an atom is true at a world if the individual objects associated with the subject terms, at that world, are in the extension of the predicate, at that world. Item 5 says the intension of a predicate abstract, in a model, is determined in the obvious way by the behavior of the formula being abstracted. The other items are essentially standard.

6 Equality

Now that the technical definitions have been given, it is time to see how things behave. I'll begin with equality, whose interaction with necessity has always been considered a bit tricky. Recall, $\mathcal{R}$ is taken to hold between any two worlds in our discussion, so the underlying logic is S_5.

Suppose c and d are constant symbols, so that $c = d$ is a formula. To say it is true at a possible world of a model is to say the interpretations of c and d, at that world, are in the interpretation of $=$ at that world. Since the interpretation of $=$ is the equality relation at every possible world, this amounts to saying that c and d designate the same object at the world. Formally we have the following: $\mathcal{M}, \Gamma \Vdash_v c = d \Leftrightarrow \mathcal{I}(c)(\Gamma) = \mathcal{I}(d)(\Gamma)$.

What about necessary equality? If we say of two individual concepts that they are necessarily equal, are we saying their equality is necessary, or are we saying they have a 'necessarily equal' property. That is, are we asserting $\Box\langle\lambda x, y.x = y\rangle(c, d)$ or are we asserting $\langle\lambda x, y.\Box(x = y)\rangle(c, d)$? The two are not synonymous.

Consider first $\Box\langle\lambda x, y.x = y\rangle(c, d)$, or equivalently $\Box(c = d)$. To say this is true at possible world Γ of a model is to say $c = d$ is true at every world. By the analysis above, this amounts to saying c and d designate the same object at every world. This is a strong requirement, and it really amounts to saying c and d are *synonymous*. It is easy to produce formal models in which $(c = d) \supset \Box(c = d)$ is not valid, that is, in which $(c = d) \supset \Box\langle\lambda x, y.x = y\rangle(c, d)$ is not valid.

Now consider the other version, $\langle\lambda x, y.\Box(x = y)\rangle(c, d)$. Suppose that c and d happen to designate the same object at possible world Γ of a model. Certainly, at every world, that object is identical to itself. But this is just what it takes for $\langle\lambda x, y.\Box(x = y)\rangle(c, d)$ to be true at Γ. Thus $(c = d) \supset \langle\lambda x, y.\Box(x = y)\rangle(c, d)$ is simply a valid formula.

The difference between the two versions is striking, at least until one realizes that different things are really being said. We might read $\Box\langle\lambda x, y.x = y\rangle(c, d)$ as asserting it is necessary *that* c and d be equal. This is an assertion about their intensions and, as noted above, really asserts synonymy. Likewise we might read $\langle\lambda x, y.\Box(x = y)\rangle(c, d)$ as asserting the necessary equality *of* c and d, that is, of the objects designated by c and d. Well, if *objects* are equal under any circumstances, they cannot be otherwise and so we have, *of* c and d, that their equality implies their necessary equality.

Suppose we apply these observations to a few well-known problematic cases, discussed in Quine (1953a). Say we have a model in which the possible worlds include the actual one and various alternatives to it – representing how things could have been. Let 'n' be a constant symbol intended to be interpreted as *the number of the planets*, which can vary in different possible worlds of our model. Also let '9' be a constant symbol interpreted as the number 9 at every possible world. (This assumes that numbers are in the domain of our model, of course.) Now what about the assertion, 'necessarily the number of the planets is nine'? If we read it as $\langle\lambda x, y.\Box(x = y)\rangle(n, 9)$ it is true – the number of the planets is, in fact, 9, and 9 is 9 no matter what. But if we read it as $\Box\langle\lambda x, y.x = y\rangle(n, 9)$, it is quite different. This amounts to asserting synonymy, and is false.

Or again, say 'm' and 'e' are intended to denote the morning and evening stars respectively – in the actual world they denote the same object, but in other situations

they need not do so. In the actual world, $m = e$ is the case, and hence $\langle \lambda x, y.\Box(x = y)\rangle(m, e)$ is true. But $\Box\langle \lambda x, y.x = y\rangle(m, e)$ is not so. In words, it is true, of the morning star and of the evening star, that they are identical, and this identity is necessary (as identity between objects is always necessary, if true). But it is not true that the morning star and the evening star are necessarily identical, that is, it is not true that the terms are synonymous.

7 Rigidity

In an example above I used a constant symbol, '9', which was interpreted to designate the same object in all possible worlds – the number 9. This is an example of a *rigid* term. For S_5, rigidity can be expressed quite simply: a term c is rigid in a model just in case the formula $\langle \lambda x.\Box(x = c)\rangle(c)$ is valid in the model. A little thought will make it clear it is asserting that, whatever c designates at a world, it designates the same thing at all worlds – in other words, its interpretation is a constant function.

Kripke and others have made the case that names in ordinary language are used rigidly (Kripke 1980). According to this theory, a name like 'Moses' received its initial designation at some point in the past and, by a complex process, some version of that designation has been passed down to us. This contrasts with definite descriptions. According to the Biblical account, Moses led the Israelites out of Egypt, but we can still make sense of a claim that he might not have done so. 'Moses' designates rigidly, but 'the person who led the Israelites out of Egypt' does not. Definite descriptions will be discussed later in this chapter.

8 De Re/De Dicto

Suppose we say 'The British monarch is necessarily the head of the British government.' This can be read in two different ways. On the one hand, we might be asserting the necessity of a particular statement, 'the British monarch is the head of the British government.' In this case, the necessity operator is used in a *de dicto* way, applying to a sentence (dictum). On the other hand, we might be ascribing a certain necessary property, 'necessarily being the head of the British government,' to an object, in this case the person who happens to be the British monarch. Such a usage of necessity is *de re*, ascribing a necessary property to a thing (res). In the present example, the *de dicto* version is correct, since the British monarch is defined to be the formal head of the British government. But the *de re* version is not correct since a British monarch could abdicate, and so no longer be government head.

To formalize the notions of the previous paragraph, suppose we introduce a constant symbol 'm,' intended to designate the 'British monarch' individual concept. That is, at each possible world it designates whoever is British monarch under those circumstances. And suppose we introduce a one-place relation symbol 'H,' intended to designate the intensional notion of being the head of the British government. It is easy to see that the *de re* version formalizes as $\langle \lambda x.\Box H(x)\rangle(m)$, while the *de dicto* version becomes

$\Box\langle\lambda x.H(x)\rangle(c)$. These certainly look different, and one can easily produce models in which they are not equivalent.

It does sometimes happen that, for certain terms, *de re* and *de dicto* usages coincide. Let us say that *de re* and *de dicto* are equivalent for a constant symbol c, in a model, provided $\langle\lambda x.\Box\Phi\rangle(c)$ and $\Box\langle\lambda x.\Phi\rangle(c)$ are equivalent at every world of that model, for every formula Φ. The question is, when does such a event occur? And the answer is quite simple: *de re* and *de dicto* are equivalent for c in a model if and only if c is rigid in that model. In particular, *de re* and *de dicto* are equivalent for names, assuming the Kripke et al. thesis. A proof of all this is not difficult, but I omit it here – one can be found in Fitting and Mendelsohn (1998).

9 Partial Designation

I've been assuming that terms always designate, but this is simplistic. A name, for instance, takes on a designation at a certain time, and before that it designates nothing. Definite descriptions provide another example. 'The present King of France' does not designate, though there were times when it did. To treat such things the notion of model must be somewhat expanded.

From now on the definition of interpretation is modified. If $\langle\mathcal{G}, \mathcal{R}, \mathcal{D}\rangle$ is an extended frame, an *interpretation* $\mathcal{I}$ is a mapping that behaves as before on relation symbols, but that assigns to each constant symbol c a mapping from *some* set of possible worlds (not necessarily all of them) to the domain of the frame. If a possible world Γ is in the domain of $\mathcal{I}(c)$, I'll say c *designates* at Γ. The definition of $(v * \mathcal{I})$ must also be modified: if c does not designate at Γ, $(v * \mathcal{I})(c, \Gamma)$ is undefined, and otherwise things are as they were.

Of course the definition of truth in a model must be modified as well. Partial truth assignments might be introduced – a formula could be true, false, or lack a truth value, at a possible world. This is an interesting direction, but it is not what is done here. I will simply assume that any ascription of a property to a term that does not designate is false. Formally, item 1 of the definition of $\mathcal{M}, \Gamma \Vdash_v \Phi$ is replaced by the following. (Recall, atoms can involve relation symbols or predicate abstracts.)

1. For an atom $R(t_1, \ldots, t_n)$,
   (a) if any of $t_1, \ldots, t_n$ do not designate at Γ then $\mathcal{M}, \Gamma \nVdash_v R(t_1, \ldots, t_n)$,
   (b) if all of $t_1, \ldots, t_n$ designate at Γ then $\mathcal{M}, \Gamma, \Vdash_v R(t_1, \ldots, t_n)$ just in case $\langle(v * \mathcal{I})(t_1), \ldots, (v * \mathcal{I})(t_n)\rangle \in \mathcal{I}(R)(\Gamma)$.

The rest of the definition remains the same.

I are *not* assuming any formula involving a non-designating term is false – only atoms. Among atoms, one in particular stands out: $\langle\lambda x.x = x\rangle$. In a model, at a world, if c fails to designate, $\langle\lambda x.x = x\rangle(c)$ will be false. But if c does designate, $\langle\lambda x.x = x\rangle(c)$ obviously must be true. Thus this abstract can serve as a convenient 'designation' predicate, and we give it that official role: $\mathbf{D}$ abbreviates $\langle\lambda x.x = x\rangle$. Now, c designates at a world if and only if $\mathbf{D}(c)$ is true at that world. If c does not designate, $\mathbf{-D}(c)$ will be true. This illustrates what was said above: there are true sentences involving non-designating terms.

10 Designation and Existence

Recall that earlier an existence relation symbol, E, was introduced. Using it, a pair of interesting abstracts can be defined.

$\mathbf{E}$ abbreviates $\langle\lambda x.\mathsf{E}(x)\rangle$.
$\overline{\mathbf{E}}$ abbreviates $\langle\lambda x.\ \neg\ \mathsf{E}(x)\rangle$.

Strictly speaking, $\mathbf{E}$ behaves the same as E and so is not really needed, but having it makes a nice symmetry with $\overline{\mathbf{E}}$. At a possible world Γ, $\mathbf{E}(x)$ is true just when x has as value an individual object that exists at Γ. Likewise, at Γ, $\overline{\mathbf{E}}(x)$ is true just when x has as value an individual object that is in the domain of the model but not in the domain of Γ, in other words, an object having possible but not actual existence at Γ. Since our possibilist quantifiers range over the domain of the model, we have the validity of $(\forall x)[\mathbf{E}(x) \vee \overline{\mathbf{E}}(x)]$ – quantifiers range over what has actual or possible existence. We also have that $(\forall x)[\overline{\mathbf{E}}(x) \equiv \neg\mathbf{E}(x)]$ is valid.

Constants are a different story, since they have intensional objects as values, and such objects might be partial. If c does not designate at a possible world Γ, neither $\mathbf{E}(c)$ nor $\overline{\mathbf{E}}(c)$ will be true at Γ, by part 1a of the definition of truth. On the other hand, if c does designate at Γ, it must designate something that actually or possibly exists. Putting all this together, we have the validity of $\mathbf{D}(c) \equiv [\mathbf{E}(c) \vee \overline{\mathbf{E}}(c)]$ for constant symbols.

All this is a little reminiscent of Meinong (1889). Think of $\mathbf{D}(c)$ as analogous to asserting that *c has being*. If c has being, it might or might not actually exist. In this sense 'the golden mountain' has being, does not actually exist, but could. Where the present treatment diverges from that of Meinong is, strictly interpreted, 'the round square' cannot designate at any possible world since the conditions are contradictory, and hence we cannot even say it has being. This point is related to the fact that, while a pair of abstracts $\mathbf{E}$ and $\overline{\mathbf{E}}$ was introduced, there was no companion for $\mathbf{D}$. An abstract $\langle\lambda x.\ \neg\ (x = x)\rangle$ could be considered, of course. But, for every constant symbol c, $\langle\lambda x.\ \neg\ (x = x)\rangle(c)$ will always be false. If c does not designate, it is false because no abstract correctly applies to a nondesignating term. If c does designate, it is false because the object designated must be self-identical. Roughly speaking, non-being is a property, but an uninteresting one since it never correctly applies to any term.

Going a little further, suppose c does not designate at Γ. Then $\mathbf{E}(c)$ will not be true at Γ, so $\neg\mathbf{E}(c)$ will be true. Of course $\overline{\mathbf{E}}(c)$ will not be true since c does not designate. It follows that $[\overline{\mathbf{E}}(c) \equiv \neg\mathbf{E}(c)]$ does not hold. This looks like a clash with the validity of $(\forall x)[\overline{\mathbf{E}}(x) \equiv \neg\mathbf{E}(x)]$, but recall that quantifiers range over individual objects, while constant symbols represent intensional objects, and may fail to designate. In fact universal generalization, $(\forall x)\Phi \supset \langle\lambda x.\Phi\rangle(c)$, is not valid – it fails when c does not designate. What we have instead is the validity of $(\forall x)\Phi \supset [\mathbf{D}(c) \supset \langle\lambda x.\Phi\rangle(c)]$.

11 Definite Descriptions

Definite descriptions, such as 'the King of France,' can be translated away into the primitives of our language, or they can be treated as primitives themselves. I'll straddle the fence, so to speak, and present both approaches.

To treat them as primitives the language must be enlarged, so that if x is a variable and Φ is a formula, then $\iota x.\Phi$ is a term with free variables those of Φ, except for x. The term $\iota x.\Phi$ is read, 'the x such that Φ,' or more briefly, 'the Φ.' The expanded definition of the term uses formulas, but the definition of formula uses terms, so it no longer is the case that terms can be defined first, and then formulas – the two must be defined simultaneously. This complicates things, but the obvious mutually recursive definition works fine. I'll skip over the details.

Next, the definition of designation for terms must be extended to include them. Suppose $\mathcal{M} = \langle \mathcal{G}, \mathcal{R}, \mathcal{D}, \mathcal{I} \rangle$ is a model and $\iota x.\Phi$ is a definite description. I'll say this definite description *designates* at possible world $\Gamma \in \mathcal{G}$ just in case there is *exactly one* d in the domain of the model such that $\mathcal{M}, \Gamma \Vdash_{v[x/d]} \Phi$. Take $\mathcal{I}(\iota x.\Phi)$ to be the mapping whose domain is the set of possible worlds at which $\iota x.\Phi$ designates, and assigns to a possible world Γ in its domain the unique d such that $\mathcal{M}, \Gamma \Vdash_{v[x/d]} \Phi$.

According to this definition, if $\iota x.\Phi$ does not designate at possible world Γ, then $\langle \lambda y.\Psi \rangle(\iota x.\Phi)$ is simply false at that world, for any formula Ψ. In particular, $\langle \lambda y.\Phi \rangle(\iota x.\Phi)$ will be false. On the other hand, if $\iota x.\Phi$ does designate at world Γ, it is immediate from the definition that $\langle \lambda y.\Phi \rangle(\iota x.\Phi)$ is true at Γ. We thus have the simple principle $\mathbf{D}(\iota x.\Phi) \equiv \langle \lambda y.\Phi \rangle(\iota x.\Phi)$. In the present world it is not true that 'the King of France is King of France,' because the definite description 'the King of France' does not designate.

Russell (1905) showed that definite descriptions could be translated away in context, essentially saying that while they have the appearance of terms, formulas containing them are really abbreviations for more complex constructions. Stating Russell's translation in present notation, $\langle \lambda y.\Psi \rangle(\iota x.\Phi)$ is taken as abbreviating the formula $(\exists z)\{(\forall w)[\langle \lambda x.\Phi \rangle(w) \equiv (w = z)] \wedge \langle \lambda y.\Psi \rangle(z)\}$. That is, we have a formula asserting exactly one object has the property $\langle \lambda x.\Phi \rangle$, and that object also has the property $\langle \lambda y.\Psi \rangle$. It is not hard to see that a Russell approach is equivalent to the approach taking definite descriptions as primitive. The same formulas are validated either way.

Ontological arguments provide interesting examples of definite descriptions at work. Let's begin with one in the Descartes style. Suppose we define God to be the necessarily existent being – take g to be short for $\iota x.\Box\mathbf{E}(x)$. A definite description has its defining property if and only if it designates, so we have the validity of $\mathbf{D}(g) \equiv \langle \lambda y.\Box\mathbf{E} \rangle(g)$. But in this case we can do better – we also have $\mathbf{D}(g) \equiv \Box\mathbf{E}(g)$. This is not because of general principles about definite descriptions, but because of the particular form involved, $\iota x.\Box\mathbf{E}(x)$, and the fact that the underlying logic is S_5. (Proof of validity takes some work – give it a try.)

Continuing: for any term c, $\mathbf{D}(c) \equiv [\mathbf{E}(c) \vee \overline{\mathbf{E}}(c)]$. It follows that $\mathbf{E}(g) \supset \mathbf{D}(g)$ is valid. Combining things, we have the validity of $\mathbf{E}(g) \supset \Box\mathbf{E}(g)$. From this, by standard modal logic manipulation, we get the validity of $\Diamond\mathbf{E}(g) \supset \Diamond\Box\mathbf{E}(g)$. Since our modal logic is S_5, $\Diamond\Box X \supset \Box X$, and so we have the validity of $\Diamond\mathbf{E}(g) \supset \Box\mathbf{E}(g)$. This is a crucial step in Descartes' argument: God's existence is necessary, if possible. To complete the proof, we must establish the validity of $\Diamond\mathbf{E}(g)$. Unfortunately, at this point Descartes simply assumed it to be the case. I'll leave it to you to verify that $\Diamond\mathbf{E}(g)$, $\Box\mathbf{E}(g)$, and $\mathbf{E}(g)$ all turn out to be equivalent, so the Descartes assumption really begs the question.

The first ontological argument, historically, was that of Anselm. I'll conclude this section with a very informal discussion of it. This time, define God to be the maximally conceivable being. That is, God is the being such that I can conceive of nothing greater. Now let g abbreviate the informal definite description, 'the maximally conceivable being.' We have $\mathbf{D}(g) \equiv [\mathbf{E}(g) \vee \overline{\mathbf{E}}(g)]$, so if we assume that g designates (in the actual world), we have either $\mathbf{E}(g)$ or $\overline{\mathbf{E}}(g)$. Making reasonable assumptions, if we had $\overline{\mathbf{E}}(g)$ we would have a contradiction, because I can conceive of an existing God, and this would be greater than a nonexisting God. Consequently we must have $\mathbf{E}(g)$. This part of the argument is very imprecise, but we don't need to make it sharper because it is clear that it all depends on the initial assumption that g designates, and that was never verified. In short, the Anselm argument makes a plausible case that 'the maximally conceivable being' cannot designate a nonexistent being, but it does not establish that it designates anything.

12 What Next?

I've sketched the semantics for a first-order modal logic, and showed how it could be used to elucidate several topics of interest to philosophers. But the logic was, by design, a limited modal logic. There were quantifiers over individual objects, and constant symbols for individual concepts. One can complete the set by adding quantifiers over individual concepts and constant symbols for individual objects (Fitting 2000a). One can then consider whether or not to simply identify individual objects with individual concepts that are rigid. Technical issues are one thing, philosophical implications another. But this is beyond what we do here.

Going still further, one can introduce higher-type notions. We already have intensional relations – we could allow quantification over them, then add relations of relations, quantify over them, and so on. The intensional/extensional split that we have already seen continues upward through all these levels, and things become quite complex. Gödel devised an ontological argument of genuine interest, but to study it formally requires some machinery of this sort (Fitting 2000b).

Of course the more complicated things get, the less immediate our intuitions. The modal logic presented here is complex enough for many purposes, yet simple enough for us to grasp informally. Further exploration can be left to the intrepid.

References

Chisholm, R. M. (ed.) (1960) *Realism and the Background of Phenomenology*. New York: Free Press.

Fitting, M. C. (2000a) Modality and databases. LNCS, Tableaux 2000. Heidelberg: Springer.

Fitting, M. C. (2000b) *Types, Tableaux, and Gödel's God*. Available on my web site: comet.lehman.cuny.edu/fitting.

Fitting, M. C. and Mendelsohn, R. (1998) *First-Order Modal Logic*. Amsterdam: Kluwer.

Hughes, G. E. and Cresswell, M. J. (1996) *A New Introduction to Modal Logic*. London: Routledge.

Kripke, S. (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Meinong, A. (1889) On the theory of objects. Reprinted in Chisholm (1960).

Quine, W. V. O. (1953a) Reference and modality (pp. 139–59). In Quine (1961).

Russell, B. (1905) On denoting. *Mind*, 14, 479–93. Reprinted in Robert C. Marsh (ed.), *Logic and Knowledge: Essays 1901–1950, by Bertrand Russell*. London: Allen & Unwin 1956.

Further Reading

Marcus, R. B. (1992) *Modalities*. New York: Oxford University Press.

Parsons, T. (1969) Essentialism and quantified modal logic. *Philosophical Review*, 78, 35–52.

Parsons, T. (1985) *Nonexistent Objects*. New Haven, CT: Yale University Press.

Quine, W. V. O. (1948) On what there is. *Review of Metaphysics*. Reprinted in Quine (1961).

Quine, W. V. O. (1953b) Three grades of modal involvement. In *The Ways of Paradox and Other Essays*. (pp. 156–174) New York: Random House.

Quine, W. V. O. (1961) *From a Logical Point of View*, 2nd edn. New York: Harper & Row.

Smullyan, A. F. (1948) Modality and description. *Journal of Symbolic Logic*, 13, 31–7.

Thomason, R. H. (ed.) (1974) *Formal Philosophy, Selected Papers of Richard Montague*. New Haven and London: Yale University Press. In particular, see "Pragmatics" (95–118), "Pragmatics and intensional logic" (119–47), "On the nature of certain philosophical entities" (148–87).

28

Proofs and Expressiveness in Alethic Modal Logic

MAARTEN DE RIJKE AND
HEINRICH WANSING

1 Introduction

> *Alethic modalities* are the necessity, contingency, possibility or impossibility of something being true. Alethic means "concerned with truth."
>
> (Lacey 1976: 132)

The above dictionary characterization of alethic modalities states the central notions of alethic modal logic: necessity, and other notions that are usually thought of as being definable in terms of necessity and Boolean negation: impossibility, contingency, and possibility. The syntax of modal propositional logic is inductively defined over a denumerable set of sentence letters $p_0, p_1, p_2, \ldots$ as follows:

$$A ::= p \mid \neg A \mid (A \vee B) \mid \Box A$$

The other Boolean operations ($\top$, $\bot$, $\wedge$, $\supset$ and $\equiv$) are defined as usual. A formula is read as follows:

'it is necessary that A'	is expressed as	$\Box A$
'it is impossible that A'	is expressed as	$\Box\neg A$
'it is contingent that A"	is expressed as	$\neg\Box A$
'it is possible that A'	is expressed as	$\neg\Box\neg A$

Usually, 'it is possible that A' is abbreviated $\Diamond A$, and if $\Diamond$ is primitive, $\Box A$ is abbreviated $\neg\Diamond\neg A$. Although one would expect that $\Box A$ implies A, the weakest system of normal modal propositional logic does not have $\Box A \supset A$ as a theorem. This is understandable from the point of view of the most prominent formal semantics for modal logic. The basic semantic intuition behind alethic modal logic is that $\Box A$ is true at a state ('possible world') s if and only if (iff) A is true at every state accessible from s. What exactly is meant by accessibility of t from s is deliberatively left open, to make room for various readings, like 't is compatible with the physical laws of s,' 't is a conceptually possible alternative of s,' 't lies in the future of s,' or 't is an output-state of a terminating per-

formance of some generic action in s.' Clearly, if $\Box A$ is true at s iff A is true always in the future of s, the unprovability of $\Box A \supset A$ is intuitively correct.

Modal reasoning has been discussed by Aristotle already, and the idea of necessary truth as truth in all possible worlds is due to Leibniz, while its modern mathematical rendition goes back to Kripke. Over the past century modal logic has been used extensively to conceptualize and reason about a wide variety of modal and modal-like notions, some of which were mentioned above. To stay within the number of pages allotted to us, we have had to impose very drastic restrictions. First of all, our treatment is mainly logical or even mathematical. Second, we have decided to focus on two topics that, we think, are of relevance to anyone wanting to use modal logic for modeling and analyzing informal notions: *expressive power* (what can we say with the logic?) and *reasoning methods* (what are the implications of what we are saying?). In both cases we will focus on propositional modal logic; however, many interesting philosophical and mathematical phenomena and problems arise in modal predicate logic, and we will briefly touch on some of them.

More concretely, we will survey the model theory of normal modal propositional logic and present basic notions and results of completeness and correspondence theory. Moreover, we indicate various ways of enhancing the expressive power of the language of alethic modal logic. We present an overview of two important types of proof systems for normal modal logics, namely labeled tableau systems and display calculi. The last part of the chapter is concerned with several problems arising in modal predicate logic. We conclude this chapter with pointers to important survey articles and volumes on modal logic.

2 Model Theory

'Revolutionary' is an overused word, but no other word adequately describes the impact relational semantics (i.e. the concepts of frames, models, satisfaction, and validity that we are about to introduce) has had on the study of modal logic. Somewhere around 1960 modal logic was reborn as a new field, through the work of authors such as Hintikka, Kanger, and Kripke. Below we recall the basic concepts that came with these changes, and we discuss one of the key issues to which the new era gave rise: expressive power.

Basics

A *relational structure* is simply a tuple $(W,R_1,R_2,\ldots)$ consisting of a domain W and relations R_1, R_2, ... on this domain. A *frame* for the propositional modal logic introduced in Section 1 is a relational structure $\mathcal{F} = (W,R)$ equipped with a single binary relation. A frame (W,R) is turned into a *model* $\mathcal{M} = (W,R,V)$ by equipping it with a *valuation* V, that is a function mapping proposition letters in the language to subsets of the domain W; note that models can be viewed as relational structures in a natural way, namely as structures of the form $(W, R, V(p_0), V(p_1), \ldots)$, consisting of a domain, a single binary relation R, and the unary relations given by V.

In spite of their mathematical kinship, frames and models are *used* very differently. Frames are essentially mathematical pictures of ontologies or structural properties that are more or less invariant across situations, while the unary relations provided by valuations decorate frames with contingent information.

DEFINITION 1 Let w be a state in a model $\mathcal{M}=(W,R,V)$. We inductively define the notion of a formula A being *true* in $\mathcal{M}$ at w (notation: $\mathcal{M},w \models A$) as follows:

$\mathcal{M},w \models p$ iff $w \in V(p)$
$\mathcal{M},w \models \neg A$ iff not $\mathcal{M},w \models A$
$\mathcal{M},w \models A \vee B$ iff $\mathcal{M},w \models A$ or $\mathcal{M},w \models B$
$\mathcal{M},w \models \Box A$ iff for all $v \in W$ with wRv we have $\mathcal{M},v \models A$

It follows from this definition that $\mathcal{M},w \models \Diamond A$ iff for some $v \in W$ with wRv we have $\mathcal{M},v \models A$. Note also that the notion of truth is *local*: formulas are evaluated at some particular state w. Moreover, $\Box$ and $\Diamond$ both work locally: only states R-accessible from the current one can be explored by our operators.

A formula A is *globally* or *universally true* in a model $\mathcal{M}$ (notation: $\mathcal{M} \models A$) if it is true at all states in $\mathcal{M}$.

Finally, these notions can also be lifted to sets of formulas Σ: $\mathcal{M},w \models \Sigma$ if $\mathcal{M},w \models A$ for every $A \in \Sigma$; and $\mathcal{M} \models \Sigma$ if $\mathcal{M} \models A$ for every $A \in \Sigma$.

One often finds the word 'world' (or 'possible world') being used for the entities in W; this use derives from our intended alethic reading of the modal language. The machinery of frames, models, and truth which we have defined is essentially an attempt to capture – by mathematical means – the view (often attributed to Leibniz) that *necessity* means *truth in all possible worlds*, and that *possibility* means *truth in some possible world*.

The truth definition stipulates that $\Diamond$ and $\Box$ check for truth not at *all* possible worlds (that is, at all elements of W) but only at R-accessible possible worlds. This may seem a weakness of the truth definition – but in fact, it is its greatest source of strength. Varying R is a mechanism which gives us a firm mathematical grip on the pretheoretical notion of access between possible worlds. For example, by stipulating that $R = W \times W$ we can allow all worlds access to each other; this corresponds to the Leibnizian idea in its purest form. Going to the other extreme, we might stipulate that *no* world has access to any other. Between these extremes there is a wide range of options to explore. Should interworld access be reflexive? Should it be transitive? What impact do these choices have on the notions of necessity and possibility? For example, if we demand symmetry, does this justify certain principles, or rule others out?

Another philosophical issue concerns the ontological status of the states in possible worlds models. Do possible worlds exist? If they exist, are they concrete or abstract entities? Lewis (1986) has been widely criticized for his concretist possible worlds realism; a well-known defender of the existence of abstract possible worlds is Plantinga (1974). Possible worlds anti-realists like Chihara (1998) try to explain away metaphysical commitments of quantification over possible worlds in the metalanguage of modal logic. It

seems fair to say that normally modal logicians do not feel hampered in their work by these ontological disputes.

Recall that models are composite entities consisting of a frame (our ontology) and contingent information (the valuation). We often want to ignore the effects of the valuation and get a grip on the more fundamental level of frames. The concept of *validity* lets us do this.

DEFINITION 2 A formula A is *valid at a state w in a frame $\mathcal{F}$* (notation: $\mathcal{F},w \models A$) if A is true at w in every model $(\mathcal{F},V)$ based on $\mathcal{F}$; A is *valid in a frame $\mathcal{F}$* (notation: $\mathcal{F} \models A$) if it is valid at every state in $\mathcal{F}$.

For instance, $\Box(A \supset B) \supset (\Box A \supset \Box B)$ is valid on all frames. In contrast, $\Diamond\Diamond p \supset \Diamond p$ is not valid on all frames, while it is valid on all transitive frames.

What does *logical consequence* mean for modal languages? Just like we have local and global notions of truth and validity, we have two consequence relations for modal formulas. A piece of terminology: if $\mathbf{S}$ is a class of models, then a *model from* $\mathbf{S}$ is simply a model $\mathcal{M}$ in $\mathbf{S}$; if $\mathbf{S}$ is a class of frames, then a model from $\mathbf{S}$ is a model based on a frame in $\mathbf{S}$.

DEFINITION 3 Let $\mathbf{S}$ be a class of models or a class of frames. Let Σ and A be a set of modal formula and a single formula. We say that A is a *(local) semantic consequence* of Σ over $\mathbf{S}$ (notation: $\Sigma \models_{\mathbf{S}} A$) if for all models $\mathcal{M}$ from $\mathbf{S}$, and all states w in $\mathcal{M}$, if $\mathcal{M},w \models \Sigma$, then $\mathcal{M},w \models A$.

As an example, suppose that we are working with **Tran**, the class of frames (W,R) in which R is a transitive relation. Then $\{\Diamond\Diamond p\} \models_{\mathbf{Tran}} \Diamond p$, but $\Diamond p$ is not a local consequence of $\{\Diamond\Diamond p\}$ over the class of all frames.

DEFINITION 4 Let A, Σ and $\mathbf{S}$ be as in Definition 3. Then A is a *global semantic consequence* of Σ over $\mathbf{S}$ (notation: $\Sigma \models^{g}_{\mathbf{S}} A$) if for all structures (i.e. models or frames) S in $\mathbf{S}$, if $S \models \Sigma$ then $S \models A$.

The local and global notions are different, yet there is a systematic connection between them. One can show that, for Σ a set of formulas and $\mathbf{F}$ a class of frames, $\Sigma \models^{g}_{\mathbf{F}} A$ is equivalent to $\{\Box^n B \mid B \in \Sigma, n \in \omega\} \models_{\mathbf{F}} A$.

Table 28.1 Some axioms

Name	*Formula*
D	$\Box p \supset \Diamond p$
T	$\Box p \supset p$
B	$p \supset \Box\Diamond p$
4	$\Box p \supset \Box\Box p$
5	$\Diamond A \supset \Box\Diamond A$

Completeness

During the first years after the arrival of possible worlds semantics, the topic of *axiomatic completeness* formed the bridge linking the new era with the previous syntactic era. The core notion here is that of a *normal modal logic*, which is simply a set of formulas satisfying certain syntactic conditions. The system **K** (after Kripke) is the minimal (or 'weakest') system for reasoning about frames; stronger systems are obtained by adding extra axioms.

A *normal modal logic* Λ is a set of formulas that contains all tautologies, $\Box(p \supset q) \supset (\Box p \supset \Box q)$, and $\Diamond p \equiv \neg\Box\neg p$, and that is closed under the following three rules

1 *Modus ponens*: given A and $A \supset B$, prove B.
2 *Uniform substitution*: given A, prove C, where C is obtained from A by uniformly replacing proposition letters in A by arbitrary formulas.
3 *Generalization*: given A, prove $\Box A$.

We write $\vdash_\Lambda A$ to denote that $A \in \Lambda$. If $\Gamma \cup \{A\}$ is a set of formulas, then A is Λ-*deducible* from Γ if either $\vdash_\Lambda A$ or there are formulas $B_1, \ldots, B_n \in \Gamma$ such that $\vdash_\Lambda (B_1 \wedge \ldots \wedge B_n) \supset A$. We call the smallest normal modal logic **K**, and a formula A is ***K*-*provable*** if $\vdash_{\mathbf{K}} A$. **K** is the minimal modal logic in the following sense: its axioms are all valid on all frames, and all three rules of inference preserve validity, hence all **K**-provable formulas are valid.

For many purposes **K** is too weak. For instance, if we are interested in transitive frames, we would like a proof system which reflects this. For example, we know that $\Diamond\Diamond p \supset \Diamond p$ (or equivalently $\Box p \supset \Box\Box p$) is valid on **Tran**, the class of all transitive frames, so we want a proof system that generates this formula. **K** does not do this, for $\Diamond\Diamond p \supset \Diamond p$ is not valid on all frames.

We can extend **K** to cope with many such semantic restrictions by adding extra axioms. Given a set of formulas Γ, we can add them as extra axioms to **K**, thus forming the axiom system **K**Γ. Table 28.1 contains some familiar axioms with their traditional names.

There is a precise sense in which **K** and its extensions **K**Γ capture frame classes. A normal modal logic Λ is *sound* with respect to a class of frames **F** if for all formulas A, $\vdash_\Lambda A$ implies $\mathcal{F} \models A$ for any $\mathcal{F} \in$ **F**. Λ is *strongly complete* with respect to **F** if for any set of formula $\Gamma \cup \{A\}$, if $\Gamma \models_{\mathbf{F}} A$ then $\Gamma \vdash_\Lambda A$. Λ is *(weakly) complete* with respect to **F** if for any formula A, if $\mathbf{F} \models A$, then $\vdash_\Lambda A$. Table 28.2 lists a number of well-known modal logics together with classes of frames for which they are sound and strongly complete.

One of the most powerful methods for proving (strong) completeness results is based on *canonical models*. Given a normal logic Λ, one proves its strong completeness with respect to a class of frames **F** by showing that every Λ-consistent set of formulas can be satisfied in a model based on a frame in **F**. The canonical model method builds this model out of maximal Λ-consistent sets of formulas and uses Λ's axioms to show that the underlying frame is in **F**. More precisely, a set Γ is maximal Λ-consistent if it is Λ-consistent (i.e. $\Gamma \nvdash_\Lambda \bot$) and any set of formulas properly containing Γ is not Λ-consistent. By Lindenbaum's Lemma, any Λ-consistent set can be extended to a maximal consistent one. The set of maximal consistent sets forms the domain of a

canonical model; the accessibility relation R in the canonical model is defined by wRv if for all formulas A, $A \in v$ implies $\Diamond A \in w$. Finally, the valuation V of the canonical model is defined by $V(p) = \{w \mid p \in w\}$.

Throughout the 1960s canonical models were the key tools used to analyze modal logics. They seem to have first been used by Makinson (1966) and Cresswell (1967), and in Lemmon and Scott (1977) (originally written in the mid-1960s) they appear fully-fledged in the form that has become standard. For a long time it was thought that every normal modal logic was complete with respect to some class of frames, and that the canonical model method could be used to prove this. The matter was resolved in 1974, when Fine (1974) and Thomason (1974) published examples of incomplete normal modal logics. We refer the reader to Chagrov and Zakharyaschev (1997) for a modern perspective and state-of-the-art account of the canonical model method.

Measuring expressive power

After the discovery of the incompleteness result, and because of an increase in interest from other disciplines to use modal logic as a *description language* for describing, for example process graphs or syntactic structures, attention shifted in part to *expressive power*. If we *are* using modal logic as a description language for talking about relational structures, which properties can we express? Which properties escape our description language? How can we overcome such limitations?

Before we can start answering such questions, we need to make a few things clear. First of all, recall that there are two levels at which we can use modal languages as description languages: the level of *models* and the level of *frames*, hence, the questions above can also be posed at two levels. Second, to be able to specify properties of models or frames that a modal language may or not may be able to express, we need some kind of 'background language.' For modal languages as languages for describing models we use a language of first-order logic which has unary predicate symbols P_0, P_1, P_2, ...

Table 28.2 Some logics and their associated accessibility conditions

Logic	*Conditions on accessibility*
K	none
KD	seriality ($\forall x \exists y\ xRy$)
KT	reflexivity
KB	symmetry
KDB	seriality, symmetry
KTB	reflexivity, symmetry
K4	transitivity
K5	Euclidicity ($\forall x \forall y \forall z$ $((xRy \wedge xRz) \supset yRz)$)
KD4	seriality, transitivity
S4 (= **KT4**)	reflexivity, transitivity
S5 (= **KTB4** = **KT5**)	universal

corresponding to the proposition letters in our modal language, as well as a single binary predicate symbol R.

How are this background language and the modal language defined at the beginning of this chapter related? Both can be used to talk about models of the kind used in Definition 1. For the modal language we already know this, while the only things we need to interpret the first-order language are a binary relation to interpret R (but the models of Definition 1 have that) and unary predicates to interpret P_0, P_1, P_2, ... (and, again, our models provide those, through the valuation). The modal truth definition provides the bridge between the two languages. To see this, let x be a first-order variable. The *standard translation* ST_x taking modal formulas to first-order formulas is defined as follows:

$$\begin{aligned} ST_x(P) &= Px, \\ ST_x(\neg\phi) &= \neg ST_x(\phi), \\ ST_x(\phi \vee \psi) &= ST_x(\phi) \vee ST_x(\psi), \\ ST_x(\Box\phi) &= \forall y\,(xRy \supset ST_y(\phi)), \end{aligned}$$

where y is a fresh variable (that is, a variable that has not been used so far in the translation). Note that the standard translation is nothing but a transcription of the modal truth definition in first-order logic.

As an example, $ST_x(\Diamond\Box p \supset p)$ is $\exists y\,(xRy \wedge \forall z\,(yRz \supset Pz)) \supset P(x)$.

PROPOSITION 1 On models, modal formulas are equivalent to their standard translations. More precisely, let A be a modal formula. Then:

1. For all models $\mathcal{M}$ and states w of $\mathcal{M}$: $\mathcal{M}, w \models A$ iff $\mathcal{M} \models ST_x(A)[w]$.
2. For all models $\mathcal{M}$, $\mathcal{M} \models A$ iff $\mathcal{M} \models \forall x\, ST_x(A)$.

(For a first-order formula $A(x)$, the expression $\mathcal{M} \models A(x)[w]$ means that $A(x)$ is true in $\mathcal{M}$ under the assignment of w to the free variable x in $A(x)$.)

Proposition 1 may be interpreted as saying that, on models, the modal language is nothing but a fragment of the first-order language that we have specified above. But which fragment? The key notion required to answer this question is that of a *bisimulation*, introduced by van Benthem (1976, 1983) in the course of his work on definability and expressive power of modal logics.

Let $\mathcal{M} = (W,R,V)$ and $\mathcal{M}' = (W',R',V')$ be two models. A nonempty binary relation $Z \subseteq W \times W'$ is called a *bisimulation between* $\mathcal{M}$ and $\mathcal{M}'$ if the following conditions are satisfied:

1. If wZw' then w and w' satisfy the same proposition letters.
2. If wZw' and wRv, then there exists v' (in $\mathcal{M}'$) such that vZv' and $w'R'v'$ (the *forth condition*).
3. The converse of 2: if wZw' and $w'R'v'$, then there exists v (in $\mathcal{M}$) such that vZv' and wRv (the *back condition*).

Two states w and w' that are linked by a bisimulation are called *bisimilar*.

PROPOSITION 2 Modal formulas cannot distinguish between bisimilar states. That is, for all models $\mathcal{M}$ and the $\mathcal{M}'$ and all states w of $\mathcal{M}$ and w' of $\mathcal{M}'$, if there is a bisimulation Z relating w to w', then $\mathcal{M},w \models A$ iff $\mathcal{M}',w' \models A$, for all modal formulas A.

What does Proposition 2 mean for our discussion on expressive power? By the proposition, if some property X is true of a state w and false of some w' that is bisimilar to it, then X cannot be expressed by means of a modal formula. Let us make this more concrete: consider the models $\mathcal{M}$ and $\mathcal{M}'$ shown in figure 28.1. There exists a bisimulation between the models; it is given by the following relation Z: Z = {(1,*a*), (2,*b*), (2,*c*), (3,*d*), (4,*e*), (5,*e*)}. Condition 1 of the definition of a bisimulation is obviously satisfied: Z-related states make the same propositional letters true. Moreover, the back and forth conditions are satisfied too: any move in $\mathcal{M}$ can be matched by a similar move in $\mathcal{M}'$, and conversely.

There are some obvious differences between, for instance, the state 3 in $\mathcal{M}$ and the state d in $\mathcal{M}'$, despite the fact that they are bisimilar. For instance, the property $\exists y \exists z (xRy \wedge xRz \wedge y \neq z \wedge P(y) \wedge P(z))$ is true of 3 in $\mathcal{M}$ but not of d in $\mathcal{M}'$. Hence, by Proposition 2, this property is not expressible by a modal formula.

But we can get more out of bisimulations. By a famous result due to van Benthem, the inability to distinguish between bisimilar states is characteristic of the modal fragment:

THEOREM 1 (VAN BENTHEM CHARACTERIZATION THEOREM) Let $A(x)$ be a first-order formula (over a vocabulary consisting of R, P_0, P_1, P_2, ...). Then $A(x)$ is equivalent to the standard translation of a modal formula iff it cannot distinguish between bisimilar states.

The above result was first proved by van Benthem (1976) in his PhD thesis; (see also van Benthem 1983). Analogous bisimulation-based characterizations have since been given for a wide variety of modal and modal-like languages; consult Blackburn et al. (2001) for an overview.

We now turn to a brief discussion of the expressive power of the modal language as a language for talking about frames. We start by explaining why frame definability is intrinsically second-order, and give examples of frame classes that are modally definable but not first-order definable. Recall that validity is defined as quantifying over all states of the domain of a frame and over all possible valuations. But a valuation assigns a *subset* of a frame to each proposition letter, and this means that when we quantify across all valuations, we are implicitly quantifying across all subsets of the frame. We can make this more precise in the following manner: we saw that at the level of models, the modal language can be translated in a truth-preserving way into a first-order language – but we can view the predicate symbols P_0, P_1, P_2, ... that correspond to the proposition letters p_0, p_1, p_2, ... as monadic second-order variables that we can quantify over. If we do this, we are in effect viewing the standard translation as a way of translating into a second-order language with a binary relation symbol, and monadic predicate variables P_0, P_1, P_2, ... This leads to the following result:

PROPOSITION 3 Let A be a modal formula. Then the following holds for any frame $\mathcal{F}$ and any state w of $\mathcal{F}$:

1. $\mathcal{F}, w \models A$ iff $\mathcal{F} \models \forall P_1 \ldots \forall P_n\, ST_x(A)[w]$
2. $\mathcal{F} \models A$ iff $\mathcal{F} \models \forall P_1 \ldots \forall P_n\, ST_x(A)$

As a concrete example, it can be shown that a formula as simple as the McKinsey formula $\Diamond\Box p \supset \Box\Diamond p$ is essentially a second-order formula when interpreted on frames: there is an uncountable frame $\mathcal{F}$ on which the McKinsey formula is valid, while it is *in*valid on each of $\mathcal{F}$s countable elementary subframes, thus showing that the McKinsey formula violates the downward Löwenheim–Skolem Theorem, one of the essential model-theoretic properties of first-order logic.

There are many modal formulas that define first-order conditions on frames. Tables 28.1 and 28.2 provide examples. Given that we have just seen that frame definability is a second-order notion, this is a surprising result. It turns out that in many cases the (often difficult to decipher) second-order condition produced by second-order translation is equivalent to a much simpler first-order condition. There exists an algorithm, called the *Sahlqvist–van Benthem* algorithm, that computes a corresponding first-order condition for a large class of modal formulas; this is the celebrated *Sahlqvist Correspondence Theorem*.

To be able to define the class of formulas for which the Sahlqvist–van Benthem algorithm works, we need the following shorthand: a *boxed atom* is a formula of the form $\Box \ldots \Box p$; in the case where the number of boxes preceding p is 0, the boxed atom $\Box \ldots \Box p$ is just the proposition letter p. Next, a *negative* formula is one in which all occurrences of proposition letters are in the scope of an odd number of negation signs. Furthermore, a *Sahlqvist antecedent* is a formula built up from $\top$, $\bot$, boxed atoms, and negative formulas, using $\wedge$, $\vee$ and $\Diamond$. A *Sahlqvist implication* is an implication $A \supset B$ in which B is positive and A is a Sahlqvist antecedent. Finally, then, a *Sahlqvist formula* is a formula that is built up from Sahlqvist implications by freely applying boxes and conjunctions, and by applying disjunctions only between formulas that do not share any proposition letters.

Examples of Sahlqvist formulas include $\Box(p \supset \Diamond p)$, and the axioms D, T, B, 4, and 5 from table 28.1. Typically forbidden combinations in Sahlqvist antecedents are 'boxes over disjunctions,' and 'boxes over diamonds,' as illustrated by the McKinsey formula.

THEOREM 2 (Sahqlvist Correspondence Theorem) Let A be a Sahlqvist formula. Then, on frames, A is equivalent to a first-order condition $C_A(x)$ that is effectively computable from A.

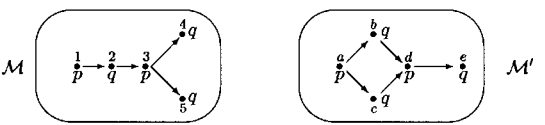


Figure 28.1 Bisimilar models

The key idea underlying the proof of the above result is the following: strip off the initial block of monadic second-order universal quantifiers in $\forall P_1 \ldots \forall P_n ST_x(A)$, thus reducing it to a first-order formula. The obvious way of getting rid of universal quantifiers is to perform universal instantiation, but the key point underlying the proof of the Sahlqvist Correspondence Theorem is that, in the case of Sahlqvist formulas, instantiations can be chosen in such a way that the resulting first-order formula is equivalent to (and not just implied by) the original second-order formula.

To illustrate the point, consider the Sahlqvist formula $(p \wedge \Diamond \neg p) \supset \Diamond p$. Its second-order translation is

$$\forall P\,(Px \wedge \exists y\,(Rxy \wedge \neg Py) \supset \exists z\,(Rxz \wedge Pz)).$$

Pulling out the existential quantifier produces

$$\forall P \forall y\,(Px \wedge Rxy \wedge \neg Py \supset \exists z\,(Rxz \wedge Pz)),$$

and moving the negative part $\neg Py$ to the consequent we get

$$\forall P \forall y\,(Px \wedge Rxy \supset Py \vee \exists z\,(Rxz \wedge Pz)). \tag{1}$$

The minimal instantiation to make Px true is one that assigns P to an object u iff $u = x$. After instantiation we obtain

$$\forall y\,(Rxy \supset y = x \vee \exists z\,(Rxz \wedge z = x)),$$

and it can be shown that this is actually equivalent to (1). The latter can of course be simplified to $\forall y\,(Rxy \wedge x \neq y \supset Rxx)$.

The Sahlqvist Correspondence Theorem comes together with a Sahlqvist Completeness Theorem: not only does every Sahlqvist formula correspond to a first-order property of frames, but when we use one as an axiom in a normal modal logic, that logic is guaranteed to be *complete* with respect to the class of frames defined by the first-order property! Moreover, the completeness result can be proved using the canonical model method; see Blackburn et al. (2001) for details.

To conclude our discussion of Sahlqvist formulas, we want to mention a result due to Kracht (1993, 1999), who has isolated the first-order formulas that are the correspondents of Sahlqvist formulas in, as an application of his so-called calculus of internal describability. Unfortunately, the details are too technical to be included here; see also Blackburn et al. (2001).

While Kracht's result gives us insight into the first-order frame properties definable by means of Sahlqvist formulas, it does not provide us with a complete description of the modally definable properties. For this, we have to turn to the Goldblatt–Thomason Theorem. The result characterizes the expressive power of modal languages on frames in terms of four fundamental frame constructions: *disjoint unions*, *generated subframes*, *bounded morphic images*, and *ultrafilter extensions*. Here, the disjoint union $\mathcal{F}$ of two frames $\mathcal{F}_1$ and $\mathcal{F}_2$ simply has the disjoint union of the domains of $\mathcal{F}_1$ and $\mathcal{F}_2$ as its

domain, while its relation is the disjoint union of the relations for $\mathcal{F}_1$ and $\mathcal{F}_2$. Moreover, $\mathcal{F}_1 = (W_1, R_1)$ is a generated subframe of $\mathcal{F}_2 = (W_2, R_2)$ if W_1 is a subset of W_2 that is closed under the addition of R_2-related states, while R_1 is simply the restriction of R_2 to W_1. A bounded morphism is nothing but a functional version of our earlier notion of bisimulation, adapted to the case of frames. And, finally, the ultrafilter extension of a frame is a kind of completion of the original frame; they are built by using the ultrafilters over a given frame as the states of a new frame, and defining an appropriate relation between them; see Blackburn et al. (2001) for formal and informal explanations.

THEOREM 3 (GOLDBLATT–THOMASON THEOREM) Let K be a class of frames that is defined by a first-order sentence. That is, let K be such that for some first-order sentence A, we have that, for all frames $\mathcal{F}$, $\mathcal{F} \in \mathsf{K}$ iff $\mathcal{F}$ satisfies A. Then K is definable by means a modal formula iff it is closed under bounded morphic images, generated subframes, disjoint unions while it reflects ultrafilter extensions in the sense $\mathcal{F} \in \mathsf{K}$ whenever the ultrafilter extension $\mathcal{F}$ is in K.

The Goldblatt–Thomason Theorem was actually proved by Goldblatt. His original result was stronger than the one we have given, applying to any frame class that is closed under elementary equivalence; this result was published in a joint paper with S. K. Thomason (1974).

3 Proof Theory

Although modern alethic modal logic started as a syntactic enterprise, its proof theory was somewhat neglected after the advent of possible worlds semantics. An exception is the development of semantic tableau calculi for modal logic. Tableau proof systems amount to rules for the construction of countermodels and take into account the relational patterns of possible worlds models. We will first consider semantic tableaux and then 'display logic', a generalization of Gentzen's sequent calculus based on the idea of residuation and Galois connection.

Tableau calculi

Tableau calculi incorporating the accessibility relation of possible worlds models were first introduced by Kripke (1963) and were later 'linearized' by various authors, notably Fitting (1983, 1993) and Mints (1992). The basic declarative unit of these calculi is not just a formula A, but rather a *formula plus label* (σ, A). In general, the label σ is a nonempty finite sequence of positive integers. A simplification is possible for **S5**. Since **S5** is characterized by the class of all frames with a universal accessibility relation R, R can be neglected, and the label σ may just be a single positive integer. A comprehensive survey on tableau methods for modal and tense logics is Goré (1999). The use of labels allows one to formulate tableau calculi for certain extensions of the minimal normal modal logic **K** by imposing constraints on accessibility and on occurrences and the shape of labels on tableau branches.

A Gentzen sequent is an expression $\Delta \rightarrow \Gamma$, where Δ and Γ are finite sets of formulas, and $\Delta \rightarrow \Gamma$ is to be understood as the claim that $\wedge \Delta \supset \vee \Gamma$ is provable. In (extensions of) classical logic, the latter formula is valid iff the set $\{A \mid A \in \Delta\} \cup \{\neg B \mid B \in \Gamma\}$ fails to be satisfiable. Rules for manipulating the sequent $\Delta \rightarrow \Gamma$ can therefore also be stated as rules for manipulating the finite set $\{A \mid A \in \Delta\} \cup \{\neg B \mid B \in \Gamma\}$. Although tableau calculi are often presented using the set notation, we here prefer a sequent notation. We will use bold letters $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ (possibly with primes or subscripts) to denote arbitrary finite sets of labeled formulas. A *sequent* is an expression of the form $\mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ is called the *antecedent* and $\mathbf{Y}$ is called the *succedent* of this sequent. We use $\mathbf{s}$, $\mathbf{s}_1$, $\mathbf{s}_2$, ... to denote sequents and the 'turnstile' $\vdash$ to denote derivability between single sequents and finite sets of sequents.

Tableau calculi are given by (finite) sets of derivation rules of the form $\mathbf{s} \vdash \mathbf{s}_1, \ldots, \mathbf{s}_n$. A *tableau* for a given sequent $\mathbf{s}$ is a tree of sequents rooted in $\mathbf{s}$, such that every node of the tree is an instantiation of one of the derivation rules of the tableau calculus under consideration. A tableau for $\mathbf{s}$ is *closed* if every leaf of any branch of the tableau has the form $(\sigma, B) \rightarrow (\sigma, B)$. We assume a binary relation of 'accessibility' between labels. This relation may satisfy certain conditions, and a number of such conditions is defined in table 28.3.

The logic **K** and various extensions of it that can be dealt with by means of labeled tableaux, require certain properties of accessibility between labels. Table 28.4 lists such systems together with the required properties of accessibility between labels. A label τ

Table 28.3 Conditions on accessibility

Name of condition	*Definition*
general	for every n, σn is accessible from σ
symmetry	for every n, σ is accessible from σn
reflexivity	σ is accessible from σ
transitivity	if σ is a proper initial segment of τ, then τ is accessible from σ
universal	any label is accessible from any label

Table 28.4 Some logics and their associated label accessibility conditions

Logic	*Conditions on accessibility*
K, KD	general
KT	general, reflexivity
KB, KDB	general, symmetry
KTB	general, reflexivity, symmetry
K4, KD4	general, transitivity
S4 (= **KT4**)	general, reflexivity, transitivity
S5 (= **KTB4**)	universal

occurring on a tableau branch is said to be a *simple, unrestricted extension* of a label σ iff (1) τ is the result of extending σ on the right with a single positive integer and (2) τ is not an initial segment of any label occurring on the branch. The label τ is *available* on a branch if it occurs on that branch. Right and left introduction rules for $\Box$ can now be stated in such a way that variations among the systems listed in table 28.4 can be accounted for by side conditions on the left rule (cf. Fitting 1993: 402). These rules are stated in table 28.5 together with the tableau rules for disjunction and negation.

For every logic Λ from table 28.4, let $T\Lambda$ be its tableau calculus. Here is an example (1) of a closed tableau for $\phi \rightarrow (1, \neg\Box A \vee \Box\Box A)$ in $T\mathbf{K4}$ and an example (2) of a closed tableau for $\phi \rightarrow (1, \neg A \vee \Box\neg\Box\neg A)$ in $T\mathbf{KB}$.

(1)
$\phi \rightarrow (1, \neg\Box A \vee \Box\Box A)$
$\phi \rightarrow (1, \neg\Box A), (1, \Box\Box A)$
$(1, \Box A) \rightarrow (1, \Box\Box A)$
$(1, \Box A) \rightarrow (\langle 1, 2\rangle, \Box A)$
$(1, \Box A) \rightarrow (\langle 1, 2, 3\rangle, A)$
$(\langle 1,2, 3,\rangle, A) \rightarrow (\langle 1, 2, 3\rangle, A)$

(2)
$\phi \rightarrow (1, \neg A \vee \Box\neg\Box\neg A)$
$\phi \rightarrow (1, \neg A), (1, \Box\neg\Box\neg A)$
$(1, A) \rightarrow (1, \Box\neg\Box\neg A)$
$(1, A) \rightarrow (\langle 1, 2\rangle, \neg\Box\neg A)$
$(1, A), (\langle 1, 2\rangle, \Box\neg A) \rightarrow \phi$
$(1, A), (1, \neg A) \rightarrow \phi$
$(1, A) \rightarrow (1, A)$

THEOREM 4 A modal formula A is a theorem of a logic Λ from table 28.4 iff there is a closed tableau for $\phi \rightarrow (1, A)$ in $T\Lambda$.

Display calculi

The display calculus (Belnap 1992) is a generalization of Gentzen's sequent calculus. We will present display logic only to the extent needed to treat normal modal logics. A more comprehensive presentation of display logic and its application to modal and non-

Table 28.5 Tableau rules

Name	*Rule*
L$\vee$	$\mathbf{X}, (\sigma, A \vee B) \rightarrow \mathbf{Y} \vdash \mathbf{X}, (\sigma, A) \rightarrow \mathbf{Y} \quad \mathbf{X}, (\sigma, B) \rightarrow \mathbf{Y}$
R$\vee$	$\mathbf{X} \rightarrow (\sigma, A \vee B), \mathbf{Y} \vdash \mathbf{X} \rightarrow (\sigma, A), (\sigma, B), \mathbf{Y}$
L$\neg$	$\mathbf{X}, (\sigma, \neg A) \rightarrow \mathbf{Y} \vdash \mathbf{X} \rightarrow (\sigma, A), \mathbf{Y}$
R$\neg$	$\mathbf{X} \rightarrow (\sigma, \neg A), \mathbf{Y} \vdash \mathbf{X}, (\sigma, A) \rightarrow \mathbf{Y}$
L$\Box$	$\mathbf{X}, (\sigma, \Box A) \rightarrow \mathbf{Y} \vdash \mathbf{X}, (\tau, A) \rightarrow \mathbf{Y}$ for any τ accessible from σ provided (i) for **K**, **KB**, and **K4**, τ must be available on the branch; (ii) for **KD**, **KT**, **KDB**, **KTB**, **KD4**, **S4**, and **S5**, τ must either be available on the branch or be a simple, unrestricted extension of σ
R$\Box$	$\mathbf{X} \rightarrow (\sigma, \Box A), \mathbf{Y} \vdash \mathbf{X} \rightarrow (\tau, A), \mathbf{Y}$ provided τ is a simple, unrestricted extension of σ

classical logics can be found in Belnap (1982, 1990), Goré (1998), Kracht (1996), Restall (1998) and Wansing (1998). The modal display calculus is based on the observation that the operators ◆ ('sometimes in the past,' i.e. the possibility operator with respect to the inverse $R^{\smile}$ of the accessibility relation R) and □ form a residuated pair. The following definition is taken from Dunn (1990: 32):

DEFINITION 5 Let $\mathcal{A} = (\mathsf{A}, \leq)$ and $\mathcal{B} = (\mathsf{B}, \leq')$ be partially ordered sets with functions f: $\mathsf{A} \to \mathsf{B}$ and g: $\mathsf{B} \to \mathsf{A}$. The pair (f,g) is called

- *residuated* iff $(fa \leq' b$ iff $a \leq gb)$;
- a *Galois connection* iff $(b \leq' fa$ iff $a \leq gb)$;
- a *dual Galois connection* iff $(fa \leq' b$ iff $gb \leq a)$;
- a *dual residuated pair* iff $(b \leq' fa$ iff $gb \leq a)$.

Obviously, (◆,□) forms a residuated pair with respect to the (local) semantic consequence relation $\models$ with respect to classes of Kripke frames. These ideas of residuation and Galois connection can be generalized, but for our purposes we have all we need to formulate introduction sequent rules for the modal operators. The polyvalent comma as a structure connective in Gentzen's sequent calculus is replaced by a number of structure connectives: $\boldsymbol{I}$ (nullary), $*$ (unary), $\bullet$ (unary), $\circ$ (binary). Every formula A is a structure, and we will use X, Y, and Z as variables for structures. The structures are defined by:

$$X ::= A \mid \boldsymbol{I} \mid {*X} \mid {\bullet X} \mid X \circ Y.$$

A *display sequent* is an expression $X \to Y$; X is called the *antecedent* and Y the *succedent* of $X \to Y$. The intended meaning of the structure connectives can be made explicit by a translation $\mathbf{t}(X \to Y) := \mathbf{t}_1(X) \supset \mathbf{t}_2(Y)$ of sequents into formulas, where $\mathbf{t}_i(A) = A$ $(i = 1,2)$, and:

$$\begin{array}{llll}
\mathbf{t}_1(\boldsymbol{I}) & = \top & \mathbf{t}_2(\boldsymbol{I}) & = \bot \\
\mathbf{t}_1(*X) & = \neg\, \mathbf{t}_2(X) & \mathbf{t}_2(*X) & = \neg\, \mathbf{t}_1(X) \\
\mathbf{t}_1(\bullet X) & = \blacklozenge \mathbf{t}_1(X) & \mathbf{t}_2(\bullet X) & = \Box\, \mathbf{t}_2(X) \\
\multicolumn{2}{l}{\mathbf{t}_1(X \circ Y) = \mathbf{t}_1(X) \wedge \mathbf{t}_1(Y)} & \multicolumn{2}{l}{\mathbf{t}_2(X \circ Y) = \mathbf{t}_2(X) \vee \mathbf{t}_2(Y).}
\end{array}$$

Under the **t**-translation, the following basic structural rules are valid in every normal modal logic:

(1) $X \circ Y \to Z \dashv\vdash X \to Z \circ {*Y} \dashv\vdash Y \to {*X} \circ Z$
(2) $X \to Y \circ Z \dashv\vdash X \circ {*Z} \to Y \dashv\vdash {*Y} \circ X \to Z$
(3) $X \to Y \dashv\vdash {*Y} \to {*X} \dashv\vdash X \to {**Y}$
(4) $X \to \bullet Y \dashv\vdash \bullet X \to Y$.

Here, $X_1 \to Y_1 \dashv\vdash X_2 \to Y_2$ is an abbreviation of $X_1 \to Y_1 \vdash X_2 \to Y_2$ and $X_2 \to Y_2 \vdash X_1 \to Y_1$. If two sequents are interderivable by means of (1)–(4), they are said to be *structurally* or *display equivalent*. The name 'display logic' is due to the fact that any substructure of a given display sequent **s** may be displayed as the entire antecedent or

Table 28.6 Introduction rules for the logical operations

Name	*Rule*
$(\rightarrow \neg)$	$X \rightarrow {*}A \vdash X \rightarrow \neg A$
$(\neg \rightarrow)$	${*}A \rightarrow X \vdash \neg A \rightarrow X$
$(\rightarrow \vee)$	$X \rightarrow A \circ B \vdash X \rightarrow A \vee B$
$(\vee \rightarrow)$	$A \rightarrow X \quad B \rightarrow Y \vdash A \vee B \rightarrow X \circ Y$
$(\rightarrow \Box)$	$\bullet X \rightarrow A \vdash X \rightarrow \Box A$
$(\Box \rightarrow)$	$A \rightarrow X \vdash \Box A \rightarrow \bullet X$

Table 28.7 Additional structural rules

Name	*Rule*
(I)	$X \rightarrow Z \dashv\vdash \boldsymbol{I} \circ X \rightarrow Z, \quad X \rightarrow Z \dashv\vdash X \rightarrow \boldsymbol{I} \circ Z$
(A)	$X_1 \circ (X_2 \circ X_3) \rightarrow Z \dashv\vdash (X_1 \circ X_2) \circ X_3 \rightarrow Z$
(P)	$X_1 \circ X_2 \rightarrow Z \vdash X_2 \circ X_1 \rightarrow Z$
(C)	$X \circ X \rightarrow Z \vdash X \rightarrow Z$
(M)	$X \rightarrow Z \vdash X \circ Y \rightarrow Z$
(MN)	$\boldsymbol{I} \rightarrow X \vdash \bullet \boldsymbol{I} \rightarrow X$

succedent of a structurally equivalent sequent s'. In order to state this fact precisely, we define the notion of antecedent and succedent part of a sequent. An occurrence of a substructure in a given structure is called positive (negative), if it is in the scope of an even (odd) number of *'s. An antecedent (succedent) part of a sequent $X \rightarrow Y$ is a positive occurrence of a substructure of X or a negative occurrence of a substructure of Y (a negative occurrence of a substructure of X or a positive occurrence of a substructure of Y).

THEOREM 5 (DISPLAY THEOREM, BELNAP 1992) For every display sequent s and every antecedent (succedent) part X of s there exists a display sequent s' structurally equivalent to s such that X is the entire antecedent (succedent) of s'.

The structure connectives *, $\boldsymbol{I}$ and $\circ$ give rise to introduction rules for the Boolean connectives, and $\bullet$ permits formulating introduction rules for $\Box$. These introduction rules are presented in table 28.6. Table 28.7 collects further structural rules that together with the introduction rules ensure the classical and normal modal behavior of the logical operations. A richer inventory of structural rules (and another choice of structure connectives) is called for in display calculi for substructural logics, see Goré (1998). In addition to structural rules and introduction rules, every display calculus contains two distinguished *logical* (structural) rules, namely identity for atoms and cut:

(identity) $\vdash \mathrm{p} \rightarrow \mathrm{p}$
(cut) $\mathrm{X} \rightarrow \mathrm{A}, \quad \mathrm{A} \rightarrow \mathrm{Y} \vdash \mathrm{X} \rightarrow \mathrm{Y}$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{A \to A}{\Box A \to \bullet A}}{\Box(\neg A \vee B) \circ \Box A \to \bullet A}\ (bs)}{*A \to * \bullet (\Box(\neg A \vee B) \circ \Box A)}}{\neg A \to * \bullet (\Box(\neg A \vee B) \circ \Box A) \quad B \to B}}{\neg A \vee B \to * \bullet (\Box(\neg A \vee B) \circ \Box A) \circ B}}{\Box(\neg A \vee B) \to \bullet(* \bullet (\Box(\neg A \vee B) \circ \Box A)) \circ B}}{\Box(\neg A \vee B) \circ \Box A \to \bullet(* \bullet (\Box(\neg A \vee B) \circ \Box A)) \circ B}\ (bs)}{\bullet(\Box(\neg A \vee B) \circ \Box A) \circ \bullet(\Box(\neg A \vee B) \circ \Box A) \to B}}{\bullet(\Box(\neg A \vee B) \circ \Box A) \to B}}{\Box(\neg A \vee B) \circ \Box A \to \Box B}\ (bs)}{*\Box B \circ \Box(\neg A \vee B) \to *\Box A}}{*\Box B \circ \Box(\neg A \vee B) \to \neg\Box A}\ (bs)}{\Box(\neg A \vee B) \to \neg\Box A \vee \Box B}$$

Figure 28.2 A derivation in **DK**

It can be shown by induction on formulas A that $\vdash A \to A$. The display calculus **DK** consists of (id), (cut), the basic and additional structural rules and the introduction rules for $\neg$, $\vee$, and $\Box$. As an example, figure 28.2 depicts a cut-free derivation of $\Box(\neg A \vee B) \to \neg\Box A \vee \Box B$, where (*bs*) indicates the repeated application of some basic structural rules.

Using induction on the complexity of X, it can be shown that in every extension of **DK** by structural rules, $\vdash X \to \mathbf{t}_1(X)$ and $\vdash \mathbf{t}_2(X) \to X$. This observation is used in the proof of the characterization theorem.

THEOREM 6 In ***DK***, $\vdash X \to Y$ iff $\mathbf{t}_1(X) \supset \mathbf{t}_2(Y)$ is provable in ***K***.

A display sequent system is said to be a *proper* display calculus, if it satisfies certain conditions C1–C8 first stated by Belnap (1992). A logic is said to be *properly displayable*, if it can be presented as a proper display calculus. Every proper display calculus enjoys cut-elimination (Belnap 1992) and even strong cut-elimination (Wansing 1998). In this case, strong cut-elimination means that there is a set of reduction steps for turning a given sequent proof into a cut-free proof of the same sequent such that – *modulo certain mild restrictions* – every sufficiently long sequence of applications of these reduction steps to a proof Π will return a cut-free proof Π' of the same sequent. The class of all properly displayable extensions of the smallest normal temporal logic has been characterized by Kracht (1996).

Here we will just consider display calculi for extensions of **K** by the familiar and important axiom schemata *D*, *T*, 4, *B* and 5 that correspond to the seriality, reflexivity, transitivity, symmetry, and Euclidicity, respectively, of the accessibility relation R. It turns out that these axiom schemata can be captured by the purely structural rules

Table 28.8 Axiom schemata and corresponding structural rules

Schema	*Structural rule*
D	$* \bullet * \boldsymbol{I} \rightarrow Y \vdash \boldsymbol{I} \rightarrow Y$
T	$X \rightarrow \bullet Y \vdash X \rightarrow Y$
4	$X \rightarrow \bullet Y \vdash X \rightarrow \bullet \bullet Y$
B	$* \bullet *(X \circ * \bullet * Y) \rightarrow Z \vdash Y \circ * \bullet * X \rightarrow Z$
5	$* \bullet * X \rightarrow Y \vdash \bullet * \bullet * X \rightarrow Y$

stated in table 28.8. Let $\theta \subseteq \{D, T, 4, B, 5\}$, $\overline{\theta} = \{\alpha' \mid \alpha \in \theta\}$. Let $\mathbf{K}\theta$ be the result of adding the axiom schemata from θ to $\mathbf{K}$, and let $\mathbf{DK}\theta'$ be the result of adding the structural rules from θ' to $\mathbf{DK}$.

THEOREM 7 In $\boldsymbol{DK}\theta$, $\vdash X \rightarrow Y$ iff $\mathbf{t}_1(X) \supset \mathbf{t}_2(Y)$ is provable in $\mathbf{K}\theta'$.

4 Modal Predicate Logic

While propositional modal logic has become a highly developed discipline with a broad spectrum of choices as regards expressive power and reasoning methods, in some cases the added modeling power of modal *predicate* logic is called for. Below we briefly discuss some of the philosophical and mathematical issues involved with this choice.

In modal predicate logic there are various junctions where metaphysics, philosophy of language and formal logic meet. Let $\mathcal{F} = (W, R)$ be a frame. If to every state $s \in W$ a domain $D = d(w)$ is associated, there are at least the following, well-known options:

1. $(\forall s \in W)$, $d(s) \neq \phi$; (varying domains);
2. $(\forall s, t \in W)$, $d(s) \neq \phi$ and if sRt, then $d(s) \subseteq d(t)$ (increasing domains);
3. $(\forall s, t \in W)$, $d(s) \neq \phi$ and $d(s) = d(t)$ (constant domains).

Is every individual present in every state? What are the effects a state transition can have on a domain? It seems natural to assume that if sRt, individuals not already present in s may appear in t or individuals present in s may disappear in t. With a fixed set of individual constants, the assumptions of varying and increasing domains permit non-designating ground (that is, variable-free) terms. In addition to the semantical problem of interpreting non-designating ground terms and formulas containing such terms, the metaphysical question arises, whether an individual may or not possess properties in a state where the individual does not exist. The assumption of constant domains corresponds to the validity of the Barcan formula $\forall x \Box A \supset \Box \forall x A$ and the assumption of increasing domains corresponds to the validity of the converse Barcan formula $\Box \forall x A \supset \forall x \Box A$. We refer to the recent Fitting and Mendelsohn (1998) for an overview of discussions of these and related matters.

There is a whole web of mathematical questions related to the Barcan formula and its variations. As to proof-theoretical aspects, the standard ordinary sequent calculus for **K** uses Gentzen sequents $\Delta \to \Gamma$ and comprises just one introduction rule for $\Box$, namely

$$\Delta \to A \vdash \{\Box B \mid B \in \Delta\} \to \Box A$$

If this calculus is enlarged by the familiar introduction rules for the universal quantifier, the Barcan formula and its converse are derivable. This fact supports the idea that modal logic requires a generalized notion of sequent.

It has often been observed that $\Box$ is a universal quantifier over possible worlds in the metalanguage of modal logic. In display logic, a universal quantifier prefix $\forall x$ can be treated like the necessity operator, by associating with $\forall x$ a structure operation $\bullet_x$ and a binary relation R_x such that in succedent position $\bullet_x A$ is interpreted as $\forall x A$ and in antecedent position as $\exists x^{\smile}$, the 'possibility' operator with respect to the converse relation $R_x^{\smile}$ of R_x. The Barcan formula and its converse then correspond to additional structural rules, for details see Wansing (1998):

$$X \to \bullet_x \bullet Y \vdash X \to \bullet\bullet_x Y; \quad X \to \bullet\bullet_x Y \vdash X \to \bullet_x \bullet Y.$$

Tableaux calculi for modal predicate logics with and without the Barcan formula can be found in Mayer and Cerrito (2000).

Just like the identity of individuals gives rise to many philosophical questions in modal predicate logic, it also gives rise to many deep mathematical questions. As a result, various alternative semantic frameworks were developed for modal predicate logic during the 1990s, including the Kripke bundles of Shehtman and Skvortsov (1990) and the category-theoretic semantics proposed by Ghilardi (1991).

The notion of (axiomatic) completeness is another source of interesting mathematical questions in modal predicate logic. It turns out that the minimal predicate logical extension of many well-behaved and complete propositional modal logics need not be complete. The main (negative) result in this area is that among the extensions of **S4**, propositional modal logics **L** whose minimal predicate logical extension is complete must have either $\mathbf{L} \supseteq \mathbf{S5}$ or $\mathbf{L} \subseteq$ S4.3. This excludes completeness results for predicate logical extensions for logics such as S4.1 and S4.3Grz. Positive completeness results are known only for some boundary cases: the predicate logical extensions of S4, S4.2, S4.3, and S5 and its extensions; see Cresswell (2001) for a recent overview.

Still further mathematical questions come up in the search for algorithmically well-behaved fragments of modal predicate logics; very powerful results were recently obtained by Hodkinson, Wolter, and Zakharyaschev (2000).

References

Belnap, N. D. (1982) Display logic. *Journal of Philosophical Logic*, 11: 375–417. Reprinted as §62 of A. R. Anderson, N. D. Belnap, and J. M. Dunn, *Entailment: The Logic of Relevance and Necessity*, vol. 2. Princeton, NJ: Princeton University Press, 1992.

Belnap, N. D. (1990) Linear logic displayed. *Notre Dame Journal of Formal Logic*, 31, 14–25.
Benthem, J. van (1976) Modal Correspondence Theory. PhD thesis, Mathematisch Instituut and Instituut voor Grondslagenonderzoek, University of Amsterdam.
Benthem, J. van (1983) *Modal Logic and Classical Logic*. Bibliopolis.
Blackburn, P., de Rijke, M. and Venema, Y. (2001) *Modal Logic*. Cambridge: Cambridge University Press. See also http://www.mlbook.org.
Bull, R. and Segerberg, K. (1984) Basic modal logic. In D. M. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*. vol. II (pp. 1–88). Dortrecht: Reidel.
Chagrov, A. and Zakharyaschev, M. (1997) *Modal Logic*, vol. 35 of *Oxford Logic Guides*. Oxford: Oxford University Press.
Chihara, C. (1998) *The Worlds of Possibility*. Oxford: Oxford University Press.
Cresswell, M. J. (1967) A Henkin completeness theorem for T. *Notre Dame Journal of Formal Logic*, 8, 186–90.
Cresswell, M. J. (2001) How to complete some modal predicate logics. In *Advances in Modal Logic*, vol. 2.
Dunn, J. M. (1990) Gaggle theory: an abstraction of galois connections and residuation with applications to negation and various logical operations. In J. van Eijck (ed.), *Proceedings JELIA 1990* (pp. 31–51). Heidelberg: Springer.
Fine, K. (1974) An incomplete logic containing *S4*. *Theoria*, 40, 23–9.
Fitting, M. (1983) *Proof Methods for Modal and Intuitionistic Logics*. Dordrecht: Reidel.
Fitting, M. (1993) Basic modal logic. In D. M. Gabbay et al. (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 1 (pp. 365–448). Oxford: Oxford University Press.
Fitting, M. and Mendelsohn, R. L. (1998) *First-order Modal Logic*. Dordrecht: Kluwer Academic Publishers.
Ghilardi, S. (1991) Incompleteness results in Kripke semantics. *Journal of Symbolic Logic*, 56: 517–38.
Girle, R. (2000) *Modal Logics and Philosophy*. Acumen.
Goldblatt, R. (2000) Mathematical modal logic: a view of its evolution. To appear. Draft available at http://www.vuw.ac.nz/~rob.
Goldblatt, R. I. and Thomason, S. K. (1974) Axiomatic classes in propositional modal logic. In J. Crossley (ed.), *Algebra and Logic* (pp. 163–73). Heidelberg: Springer.
Goré, R. (1998) Substructural logics on display. *Logic Journal of the IGPL*, 6, 451–504.
Goré, R. (1999) Tableau methods for modal and temporal logics. In M. D'Agostino, D. M. Gabbay, R. Hähnle and J. Posegga (eds.), *Handbook of Tableau Methods*. Dordrecht: Kluwer Academic Publishers.
Hodkinson, I. M., Wolter, F. and Zakharyaschev, M. (2000) Decidable fragments of first-order temporal logics. *Anals of Pure and Applied Logic*, 106, 85–134.
Hughes, G. and Cresswell, M. J. (1996) *A New Introduction to Modal Logic*. London: Routledge.
Kracht, M. (1993) How completeness and correspondence theory got married. In M. de Rijke (ed.), *Diamonds and Defaults* (pp. 175–214).
Kracht, M. (1996) Power and weakness of the modal display calculus. In H. Wansing (ed.), *Proof Theory of Modal Logic* (pp. 93–121). Dordrecht: Kluwer Academic Publishers.
Kracht, M. (1999) *Tools and Techniques in Modal Logic*. Number 142 in Studies in Logic. Amsterdam: Elsevier.
Kripke, S. (1963) Semantical analysis of modal logic I: Normal modal propositional calculi. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 9, 67–96.
Lacey, A. R. (1976) *A Dictionary of Philosophy*. London: Routledge & Kegan Paul.
Lemmon, E. J. and Scott, D. S. (1977) *The "Lemmon Notes": An Introduction to Modal Logic*. Oxford: Blackwell.
Lewis, D. (1986) *On the Plurality of Worlds*. Oxford: Blackwell.

Makinson, D. C. (1966) On some completeness theorems in modal logic. *Zeitschrift für mathemathische Logik und Grundlagen der Mathematik*, 12, 379–84.
Mayer, M. Cialdea and Cerrito, S. (2000) Variants of first-order modal logic. In R. Dyckhoff (ed.), *Proceedings Tableaux 2000*. Heidelberg: Springer.
Mints, G. (1992) *A Short Introduction to Modal Logic*. Stanford, CA: CSLI Publications.
Plantinga, A. (1974) *The Nature of Necessity*. Oxford: Oxford University Press.
Popkorn, S. (1992) *First Steps in Modal Logic*. Cambridge: Cambridge University Press.
Restall, G. (1998) Displaying and deciding substructural logics 1: Logics with contraposition. *Journal of Philosophical Logic*, 27, 179–216.
Shehtman, V. and Skvortsov, D. (1990) Semantics of non-classical first-order predicate logics. In *Mathematical Logic* (pp. 105–16). New York: Plenum Press.
Thomason, S. K. (1974) An incompleteness theorem in modal logic. *Theoria*, 40, 150–8.
Wansing, H. (1998) *Displaying Modal Logic*. Dordrecht: Kluwer Academic Publishers.
Zakharyashev, M., Segerberg, K., de Rijke, M. and Wansing, H. (2000) The origins of modern modal logic. In M. Zakharyashev et al. (eds.), *Advances in Modal Logic*. Vol. 2. Stanford, CA: CSLI Publications.

Further Reading

We conclude this chapter with some pointers to the literature on modal logic. First, details on the history of modern modal logic are available, for instance, in Blackburn et al. (2001); Bull and Segerberg (1984); Goldblatt (2000); Zakharyashev et al. (2000). Second, there are several survey papers in recent and not so recent handbooks that can serve as valuable starting points for further studies; these include Bull and Segerberg (1984). Third, there is a broad range of modern textbooks on modal logic, ranging from the philosophically oriented: Girle (2000); Hughes and Cresswell (1996) to the more mathematically inclined: Blackburn et al. (2001); Popkorn (1992). Finally, the Advances in Modal Logic initiative, with its accompanying workshops, volumes, and web site at www.aiml.net is a rich source of information.

29

Alethic Modal Logics and Semantics

GERHARD SCHURZ

1 Introduction

The first axiomatic development of modal logic was untertaken by C. I. Lewis in 1912. Being anticipated by H. McCall in 1880, Lewis tried to cure logic from the 'paradoxes' of *extensional* (i.e. *truthfunctional*) implication $\supset$ (cf. Hughes and Cresswell 1968: 215). He introduced the stronger notion of *strict* implication $<$, which can be defined with help of a necessity operator $\Box$ (for 'it is neessary that:') as follows: $A < B$ *iff* $\Box(A \supset B)$; in words, A strictly implies B iff A necessarily implies B (A, B, . . . for arbitrary sentences). The new primitive sentential operator $\Box$ is *intensional* (*non-truthfunctional*): the truth value of A does *not* determine the truth-value of $\Box$A. To demonstrate this it suffices to find two particular sentences p, q which agree in their truth value without that $\Box$p and $\Box$q agree in their truth-value. For example, it p = 'the sun is identical with itself,' and q = 'the sun has nine planets,' then p and q are both true, $\Box$p is true, but $\Box$q is false. The dual of the necessity-operator is the possibility operator $\Diamond$ (for 'it is possible that:') defined as follows: $\Diamond A$ *iff* $\neg\Box\neg A$; in words, A is possible iff A's negation is not necessary. Alternatively, one might introduce $\Diamond$ as new primitive operator (this was Lewis' choice in 1918) and define $\Box A$ as $\neg\Diamond\neg A$ and $A < B$ as $\neg\Diamond(A \wedge \neg B)$.

Lewis' work cumulated in Lewis and Langford (1932), where the five axiomatic systems S1–S5 were introduced. S1–S3 are weaker that the standard systems of § 2.2, but S4 and S5 coincide with standard S4 and S5 (for details on Lewis' systems cf. Hughes and Cresswell 1968: ch. 12; Chellas and Segerberg 1996). Lewis' pioneer work was mainly syntactic-axiomatic, except for the modal matrix-semantics (for details in the 'algebraic' tradition, started by Lukasiewicz, cf. Bull and Segerberg 1984: 8ff). The philosophically central semantics for modal logic is *possible world semantics*. It goes back to ideas of Leibniz, was first developed by Carnap and received broadest acceptance through the later work of Kripke. The *actual* world, in which we happen to live, is merely one among a multitude of other possible worlds, each realizing a different but logically complete collection of facts. The basic idea of possible world semantics as expressed by Carnap (1947: 9f, 174f) is:

(1) $\Box$A is true in the actual world iff A is true in all possible worlds.
$\Diamond$A is true in the actual world iff A is true in some possible world.

Thus, the truth valuation of sentences is *relativized* to possible worlds (or just: *worlds*). In order to obtain a recursive definition, the truth of modalized sentences (e.g. $\Box A$, $\Diamond A$) must also be determined relative to possible worlds. Let W be a given set of worlds w, $w_1 \ldots \in W$; then (1) is rephrased as follows:

(1*) $\Box A$ [or: $\Diamond A$] is true in a given $w \in W$ iff A is true in all [or some, resp.] $w \in W$

What is the *ontological status* of possible worlds? Forbes (1985: 74) distinguishes between three philosophical positions: (1) According to *absolute* realism, possible worlds exist and are entities sui generies. Lewis (1973: 84ff) has defended this position. (2) For *reductive* realism, possible worlds exist but can be reduced to more harmless (e.g. conceptual or linguistic) entities. (3) For *anti-realism*, possible worlds don't exist; so possible-world-sentences are either false or meaningless. While (1) and (3) are extreme positions, some variant of position (2) is the most common view. Kripke (1972: 15), for example, denies the 'telescope view' of worlds and conceives possible worlds as possible (counterfactual) *states* or *histories* of the actual world. (The 'possible state' – versus 'possible history' – interpretation is a further important choice; cf. Schurz 1997: 40f.) Those who still regard Kripke's counterfactual position as too problematic may alternatively conceive worlds as metalinguistic entities, namely as *interpretation functions* of the object language (this was Kripke's early view in 1959, whereas in 1963a he introduced W as a set of primitive objects). Even more scrupulous, Carnap (1947: 9) had identified worlds with object-language entities – his so-called *state descriptions*. Carnap's concept was generalized by Hintikka's (1961: 57–9) to so-called 'extended state descriptions' which in the terminology of the section below entitled "Axiomatic Systems: Correctness, Completeness, and Correspondence" are nothing but worlds of canonical models. *In the upshot:* possible world semantics does not force one into a particular metaphysical position.

A logically decisive but rarely discussed question is the determination of the set W of possible worlds. There are two options:

C-SEMANTICS We identify W with the *fixed* set $W^{\mathcal{L}}$ containing *all* worlds, or interpretations, which are logically possible in the given language $\mathcal{L}$. Then $\Box$ is a logically *constant* symbol with a *logically fixed* interpetation – that of *logical* necessity. A (modal or non-modal) sentence is then defined as *logically* true iff it is true in all worlds of $W^{\mathcal{L}}$.

K-SEMANTICS Alternatively, we consider W as a *varying* set of possible worlds, or interpretations, which need not comprise *all* logically possible worlds. Then $\Box$, though formally a *logical* symbol, has an implicitly varying interpretation (similar to $\forall$ in first-order logic because of the varying domain; cf. Schurz 1999). For example, if W contains all logically possible worlds, then $\Box$ means 'logically possible,' while if W contains only all physically possible worlds, then $\Box$ means 'physically possible.' In this setting, we count a sentence as logically true only if its truth does not depend on such special choices of W; thus we consider a (non-modal or modal) sentence as *logically* true iff it is true in all worlds $w \in W$ *for all sets of possible worlds* W.

C-semantics is the semantics of Carnap (1946: 34; 1947: 9f, 174f). It leads to a modal logic which is called **C** in Schurz (2000). **C** is stronger than **S5** and exhibits non-classical features such as *failure of closure under substitution* (in **C**, $\Diamond p$ is a logically true for every propositional variable p, but $\Diamond(p \wedge \neg p)$ is logically false), or axiomatization by *non-monotonic* rules (if A is *not* a **C**-theorem, then $\Diamond\neg A$ is a **C**-theorem; cf. Schurz 2000 and Gottlob 1999). These are largely ignored facts, due to certain confusing historical peculiarities, for example that Carnap (1946) himself had announced to have obtained Lewis' system **S5**. But this result was based an *ad hoc* deviation: in his modal propositional logic, Carnap restricts the logical truths of **C** to the subclass of those formulas which are closed under substitution (1946: 40, D4-1) and shows that the so restricted class of theorems is equivalent to Lewis' **S5**. For further details see Hendry and Pokriefka (1985) and Schurz (2000), who defends **C** in spite of its non-classical features.

K-semantics has been introduced by Kripke (1959, 1963a, 1963b), whose papers have opened the highway to the modern modal logicians' industry. As proved by Kripke (1959), K-semantics leads exactly to the system **S5** (similar results were obtained by Hintikka (1961) and Kanger (1957a), not to forget Prior (1957), the founder of tense logic). K-semantics leads to modal logics which enjoy all classical properties of logics; on the cost that standard modal logics do not contain non-trivial possibility theorems (cf. Schurz 2000, theorems 3 + 4). Apart from this insufficiency, the theorems of **S5** are rather strong: for example for every *purely modal* sentence A (each atomic subformula of A occurs in the scope of a modal operator), $\Box A \vee \Box\neg A$ is **S5**-valid. The crucial step which utilized K-semantics for weaker systems and added an almost unlimited semantical flexibility to K-semantics was the introduction of the so-called *relation R of accessibility*, or 'relative possibility,' between possible worlds, independently by Kanger (1957a), Hintikka (1961) and Kripke (1963a). Thus, w_1Rw_2 means that world w_2 is accessible from (or possible with respect to) w_1, and the refined modal truth clause goes as follows:

(2) $\Box$A is true in a given $w \in W$ iff A is true in *all* w^* such that wRw^*.
$\Diamond$A is true in a given $w \in W$ iff A is true in *some* w^* such that wRw^*.

By varying structural conditions on the relation R (e.g. reflexive, transitive, symmetric) one gets different modal logics, among them the standard systems **T**, **S4**, and **S5**. A multitude of similar results were produced in the following decades, with outstanding modal logicians of the '2nd generation' such as Lemmon and Scott, Segerberg or Fine, to name just a few. While C-semantics was almost completeley neglected, K-semantics dominated the development of modal logic, whence the remaining sections focus on K-semantics. Due to K-semantical flexibility, various new philosophical interpretations of the modal operator have been discovered. For example, in systems weaker than **T**, the modal operator may be interpreted as 'it is obligatory that . . . ,' which leads to Kripkean semantics for so-called *deontic* logics, or as 'it is believed that. . . . ,' which brings us into *epistemic* logic, etc. (see Gabbay and Guenthner 1984, and ch. XIII of this volume). This development led to a broader understanding of 'modal logic' as the *logic of intensional propositional operators*, while the narrow meaning of modal logic as the logic of necessity and possibility is expressed in the specification '*alethic*' modal logic.

So far we have discussed only modal propositional logic – from now on: *MPL*. Many more difficulties are involved in *modal quantificational* (or predicate) *logic* – from now on: *MQL*. Here we have, besides W and R, a domain D of *individuals* (i.e. objects). Here we have two major choices.

CHOICE 1 Should we assume that singular terms denote the same object in all possible worlds (*rigid designators*), or that their reference object varies from world to world (*non-rigid designators*)?

CHOICE 2 Should we suppose that every object in D exists necessarily, that is exists in all possible worlds (*constant domain*), or should we better admit that some individuals may exist in one world without existing in another world (*varying domains*)?

Until today the difficulties connected with these choices have not been completely solved.

Quine's famous attack on the reasonableness of 'de re' modalities in 1943 started the well-documented debate on these choices (see Linsky 1971). A formula is called modally *de re* (in the 'strong' sense) iff an individual constant or variable in A occurs free in the scope of '□'; otherwise it is called *de dicto* (Fine 1978: 78, 135, 143; Forbes 1985: 48f). For example, □Fa and ∃x□Fx are *de re*, while □∃xFx is *de dicto*. The crucial semantical property of *de re* formulas is that their semantic evaluation requires an identification or correlation of objects across different worlds. For example, ∃x□Fx says that in our world there exists an individual which in all possible worlds has property F – in other words, F is an *essential* (i.e. necessary) properties of this individuals. Hence we assume that an object of our world – or at least some identifiable correlate of it – exists in all other possible worlds. In contrast, □∃xFx merely asserts that in all possible worlds *some* individual exists which has property F; this does not presuppose any correlation between individuals in different worlds. Thus, the semantical question of fixed versus varying domains and rigid versus nonrigid designators does not concern *de dicto* but only *de re* sentences.

Quine (1943) has argued that the reference of singular terms depends on *contingent* facts, whence modal contexts are opaque: substitution of identicals fails in them. In his famous example, both '□(9 > 7)' and '9 = the number of planets' are true, but '□(the number of planets > 7)' is obviously false. Quine concludes that modal *de re* statements lack clear meaning. Ruth Barcan-Marcus, who developed Lewis-style MQLs in 1946, gave a profound defence of MQL against Quine's attack. In 1960 Barcan-Marcus emphasized that the failure of substitution of identicals (a = b) ⊃ (A[a] ≡ A[b])) in MQLs does not deprive *de re* sentences from clear meaning. She also shows that substitution of necessary identicals (□(a = b) ⊃ (A[a] ≡ A[b])) still holds. In (1963), she argued that the reference of proper names – in contrast to definite descriptions – should indeed be regarded as the same across all possible worlds. In a modified form, this thesis was defended by Kripke (1972); he suggested the name 'rigid designator' and made it prominent, especially the connected thesis of *necessities aposteriori* (a = b ⊃ □(a = b); cf. 1972: 35–8).

Apart from Carnap's early work, the first semantically interpreted MQL-**S5** system was developed by Kripke (1959), who assumes rigid designators and constant domain.

Rigid designators are axiomatically reflected in the two theorems $(\neg)x = y \supset \Box(\neg)(x = y)$. Constant domains gets axiomatically cashed out in the Barcan formula BF: $\forall x\Box A \supset \Box\forall xA$ (introduced by Ruth Barcan 1946). Neither BF nor its converse cBF: $\Box\forall xA \supset \forall x\Box A$ are especially plausible; but varying domains cause drastic difficulties (see below, "Nonrigid designators, counterpart theory, and worldline semantics"). A comparatively simple system based on varying domains and rigid designators was developed by Kripke (1963b), on the cost of restricting necessitation rule. Hintikka (1961: 63f), argues in favor of varying domains and nonrigid designators (see also Hughes and Cresswell 1968: 190). Later, Lewis (1968) argued that individuals at different worlds can never be identical, but can merely be so-called *counterparts* of each other. His important philosophical point is that in order to avoid Quine's de re skepticism, it is not necessary to assume rigid designators; it suffices to assume the existence of a counterpart relation. To says that F is an *essential property* of a ($\Box$Fa) means in Lewis' theory that all counterparts of a in all possible worlds have property F (Lewis 1968: 118). Thus, although Kripke (1972) critizices Lewis, both agree in their *essentialism*, that is in their optimistic view about *de re* modalities.

The metaphysically significant alternative to both Kripke and Lewis is *de re skepticism*. The *de re* skeptics doubt that identifications or correlations of objects across possible world are an intelligible concept. Von Wright (1951: 26–8) suggested that in a satisfying modal logic all *de re* modalities should be eliminable in favor of *de dicto* modalities (see Hughes and Cresswell 1968: 184ff). This position was reconstructed as the position of 'anti-Haecceitism' by Fine (1978). According to its basic idea, the naming of individuals in possible worlds rests on purely conventional grounds. Thus, in an 'anti-Haecceitist' possible world model the accessible worlds should be closed under *local isomorphisms* w.r.t. their domains of individuals; Fine calls such possible world models *homogeneous* (1978: 283). A singular necessity statement $\Box$Fa is true in a world w of a homogeneous model only if its universal closure $\forall x\Box$Fx is true, too. Fine (1978: 281) proves that the quantificational system **S5** + **H** is complete for the class of homogeneous possible world models. This system is obtained from **S5** by adding all $\forall$-$\Box$-closures of axiom H: $(\mathrm{Dif}(x_1, \ldots, x_n) \wedge \Box A) \supset \Box\forall x_1 \ldots \forall x_n(\mathrm{Dif}(x_1, \ldots, x_n) \supset A)$, where $\mathrm{Dif}(x_1, \ldots, x_n) =_{df} \wedge \{x_i = x_j\colon 1 \leq i < j \leq n\}$ and A's free variables are among $x_1, \ldots, x_n$.

2 Modal Propositional Logics (MPLs)

Language

In what follows, capital Latin A, B . . . will vary over formulas of the object language and capital Greek Γ, Δ, . . . over sets of them. F will always denote a frame and M a model, **F** a set of frames and **M** a set of models, W a possible world set, R the accessibility relation, and V a valuation function. The letters w, u, v will range over possible worlds (all symbols may also be used in an indexed way). We use all standard symbols of informal first-order logic and informal set theory, which forms our *metalanguage* (see van Dalen et al. 1978); in particular $\Rightarrow$ is the implication sign of the metalanguage. Our object language is $\mathcal{L}$, the language of MPL. It contains as nonlogical symbols a

denumerably infinite set of propositional variables $\mathcal{P}$, and as primitive logical symbols the truth-functional connectives $\neg$ (negation), $\vee$ (disjunction) and the necessity operator $\Box$. The other truth-functional connectives $\wedge$ (conjunction), $\supset$ (material implication), $\equiv$ (material equivalence), T (Verum), $\bot$ (Falsum) and the possibility operator $\Diamond$ are defined as usual ($A \wedge B =_{df} \neg(\neg A \vee \neg B)$, $A \supset B =_{df} \neg A \vee B$, $A \equiv B =_{df} (A \supset B) \wedge (B \supset A)$, $\mathsf{T} =_{df} p \vee \neg p$, $\bot =_{df} p \wedge \neg p$, $\Diamond A =_{df} \neg\Box\neg A$). $\mathcal{L}$ is identified the set of its (well-formed) formulas, that is *sentences*, which are recursively defined as follows: (1) $p \in \mathcal{P} \Rightarrow p \in \mathcal{L}$, (2) $A \in \mathcal{L} \Rightarrow \neg A \in \mathcal{L}$, (3) $A, B \in \mathcal{L} \Rightarrow (A \vee B) \in \mathcal{L}$, (4) $A \in \mathcal{L} \Rightarrow \Box A \in \mathcal{L}$ (nothing else). $\mathcal{P}(A)$ = the set of propositional variables in A.

Possible Worlds Semantics

A *frame* is a pair $F = \langle W,R \rangle$ where $W \neq \varnothing$ (a nonempty set of 'possible worlds') and $R \subseteq W \times W$ (the accessibility relation; uRv abbreviates $\langle u,v \rangle \in R$). A *model* for $\mathcal{L}$ is triple $M = \langle W,R,V \rangle$ where $\langle W,R \rangle$ is a frame (we say that M is based on this frame) and V: $\mathcal{P} \rightarrow$ Pow(W) is a valuation function which assigns to each propositional variable $p \in \mathcal{P}$ the set of worlds $V(p) \subseteq W$ *at* which p is true ('Pow' for 'power set'). We also write W^F, R^F to indicate that W and R belong to F; and likewise for W^M, R^M and V^M. The assertion 'formula A is *true at* world w *in* model M' (where $w \in W^M$) is abbreviated as $(M,w) \models A$ and recursively defined as follows: (1) $(M,w) \models p$ iff $w \in V(p)$; (2) $(M,w) \models \neg A$ iff not $M \models A$, and $M \models A \vee B$ iff $M \models A$ or $M \models B$; finally (3) $(M,w) \models \Box A$ iff for all $u \in W$ such that wRu, $(M,u) \models A$. Sentence $A \in \mathcal{L}$ is defined as *valid in* model M, in short: $M \models A$, iff A is true at all worlds of M. The set of worlds verifying A in model M is also written as $\|A\|^M$ and considered as the *proposition* expressed by the sentence A in model M. Formula A is *valid on* a frame F, in short $F \models A$, iff A is valid in all models based on F. Formula A is valid w.r.t. (with respect to) a class of models $\mathbf{M}$, in short $\mathbf{M} \models A$, or w.r.t. a class of frames $\mathbf{F}$, in short $\mathbf{F} \models A$, iff A is valid in all $M \in \mathbf{M}$, or on all $F \in \mathbf{F}$, respectively. Analogously, a formula set Γ is valid in a model M, $M \models \Gamma$, iff all formulas in Γ are valid in M; analogously for validity of Γ on F, w.r.t. $\mathbf{M}$, and w.r.t. $\mathbf{F}$. A formula set $\Gamma \subseteq \mathcal{L}$ is said to be (simultaneously) *satisfiable* in a model M (or: w.r.t. a model-class $\mathbf{M}$) iff all formulas in Γ are true at some world in M (or: at some world in some $M \in \mathbf{M}$, respectively), and $\Gamma \subseteq \mathcal{L}$ is (simultaneously) *satisfiable* on a frame F (or: w.r.t. a frame-class $\mathbf{F}$) iff Γ is satisfiable on some model based on F (or: in some model based on some $F \in \mathbf{F}$).

Logics can be defined in a semantical way (this section) and in an axiomatic-syntactical way (next section). Let $\mathbf{M}(\mathbf{F})$ denote the class of all models based on some frame in frame-class $\mathbf{F}$, and call a model class $\mathbf{M}$ *frame-based* iff $\mathbf{M} = \mathbf{M}(\mathbf{F})$ for some $\mathbf{F}$. Frame classes are defined by purely structural conditions on R and allow all possible valuation functions. In contrast, not-frame-based model classes are defined by restrictions on the valuation function. A *logic*, however, should admit all possible valuations of its nonlogical symbols (see Schurz 1999). Therefore, frame-classes and frame-based model classes are the philosophically more important means to characterize modal logics, as compared to not-frame-based model-classes (such as the 'general frames' of cf. 'More metalogical results on PMLs' below). Semantically, a MPL $\mathbf{L}$ can be defined as the set of formulas which are valid w.r.t. a given class $\mathbf{F}$ of frames: $\mathbf{L} = \mathbf{L}(\mathbf{F}) = \{A: \mathbf{F} \models A\}$; the so-defined $\mathbf{L}$ is a 'normal' MPL. Formula A is said to be a valid consequence of

Γ w.r.t. frame class **F**, in short $\Gamma \models_{\mathbf{F}} A$, iff for all worlds w in all models M based on some frame in **F**, (M,w) $\models \Gamma$ implies (M,w) $\models$ A. If $\Gamma \models_{\mathbf{F}} A$, we also say that the *rule* Γ/A (read: 'Γ, therefore: A') is *valid* w.r.t. frame class **F**. Validity of a rule means *truth-preservation*. It is important to distinguish this from the *admissibility* of a rule, which means *validity*-preservation (Schurz 1994). In first-order logic, for example, Modus Ponens MP: A, A $\supset$ B/B is valid (truth-preserving) while Universal Generalization UG: A/$\forall$xA is merely admissible (validity preserving). We call a *rule* Γ/A (semantically) *admissible* w.r.t. frame class **F**, in short **F**-admissible, iff **F** $\models \Gamma$ implies **F** $\models$. A rule is called *frame-admissible* iff it preserves validity in every frame, and it is called *model-admissible* iff it preserves validity in every model. The reason for our definition of valid consequence (also called *local* consequence by van Benthem 1983: 37f) is that it implies the *Deduction Theorem:* Γ, A $\Vdash_{\mathbf{F}}$ B $\Rightarrow$ $\Gamma \Vdash_{\mathbf{F}}$ A $\supset$ B. This theorem does not hold for merely frame-admissible consequences (which correspond to van Benthem's 'global' consequence).

With **FK** for the class of all (Kripke) frames, the following implication relation holds: Γ/A is **FK**-*valid* $\Rightarrow$ Γ/A is model-admissible $\Rightarrow$ Γ/A is frame-admissible $\Rightarrow$ Γ/A is **FK**-admissible. We first consider the logic **K** (for Kripke) which is semantically defined as the set of modal formulas which **FK**-valid, **K** = **L**(**FK**). Some terminology: A[B/C] denotes the result of replacing *some* occurrences of subformula B in A by C (so, strictly speaking, 'A[B/C]' varies over several formulas). A substitution function s:$\mathcal{P} \to \mathcal{L}$ substitutes arbitrary formulas s(p) for propositional letters p. The substitution instance s(A) results from A by replacing every p $\in \mathcal{P}$(A) in A by s(p).

***FK**-valid theorems*

Taut: Every tautology

K:	$\Box(A \supset B) \supset (\Box A \supset \Box B)$	K$\Diamond$:	$(\neg\Diamond A \wedge \Diamond B) \supset \Diamond(\neg A \wedge B)$
T:	$\Box$T	T$\Diamond\Box$:	$\neg\Diamond\bot$
C:	$(\Box A \wedge \Box B) \supset \Box(A \wedge B)$	C$\Diamond$:	$\Diamond(A \vee B) \supset (\Diamond A \vee \Diamond B)$
M:	$\Box(A \wedge B) \supset \Box A \wedge \Box B$	M$\Diamond$:	$\Diamond A \vee \Diamond B \supset (\Diamond A \vee \Diamond B)$

Further theorems

1. $\Box A \vee \Box B \supset \Box(A \vee B)$,
2. $\Diamond(A \wedge B) \supset \Diamond A \wedge \Diamond B$,
3. $\Box(A \supset B) \supset (\Diamond A \supset \Diamond B)$,
4. $\Box A \wedge \Diamond B \supset \Diamond(A \wedge B)$,
5. $\Box(A \vee B) \supset \Diamond A \vee \Box B$,
6. $(\Diamond A \supset \Box B) \supset \Box(A \supset B)$,
7. $\Diamond(A \supset B) \equiv (\Box A \supset \Diamond B)$.

***FK**-valid rules*

TautR – all tautological rules in particular MP: A, A $\supset$ B/B.

Model-admissible rules

N: A/$\Box$B

E: A $\equiv$ B / $\Box$A $\equiv$ $\Box$B

RE: B $\equiv$ C / A $\equiv$ A[B/C]

Further
All rules resulting from valid theorems by applying deduction theorem

A frame-admissible rule
Subst: A/s(A) for every $s: \mathcal{P} \to \mathcal{L}$

PROOFS *Exercise* (see proof examples below). *Hints:* The tautological theorems and rules hold because the clauses for truth-functional connectives are the same as in non-modal logic. In other words, (classical) modal logics *contain* truth-functional logic. Rule RE ('replacing equivalents') is a consequence of E ('equivalence') and proved by induction on complexity of formulas. Rule N ('necessitation') and the principle K (Kripke) are characteristic for *normal* logics validated by Kripke frames, while rule E and principles M, C, and T are used to axiomatize the 'weaker' classical logics. Every $\Box$-theorem has a $\Diamond$-*dual* which obtained by replacing '$\Box$' by '$\neg\Diamond\neg$' and applying tautological transformations. Note that $\Box$ distributes over $\wedge$ in both directions (M, C), but $\Box$ distributes over $\vee$ only in one direction (i); thus $\Box$ behaves like an implicit universal quantifier. The same relations hold, dually, between $\Diamond$ and $\vee$, $\wedge$; so $\Diamond$ behaves like an implicit existential quantifier.

PROOF OF VALIDITY OF (K) We prove $\Vdash_{\mathbf{Rk}} \Box(A \supset B) \supset (\Box A \supset \Box B)$ by assuming, for an arbitrary model M and world w in W^M, that (a): $(M,w) \models \Box(A \supset B)$, and (b): $(M,w) \models \Box A$, and by proving that (a) and (b) implies (c): $(M,w) \models \Box B$. By (a) and truth clauses, $(M,u) \models A$ implies $(M,u) \Vdash B$, for all u with wRu. By (b), $(M,u) \Vdash A$ holds for all u with wRu. Therefore, $(M,u) \models B$ holds for all u with wRu, which gives us (c). Q.E.D.

PROOF THAT N IS MODEL- (AND HENCE FRAME-) ADMISSIBLE By contraposition. Assume (for arbitrary M) that $M \not\models \Box A$. Then there exists $w \in W^M$ such that $(M,w) \not\models \Box A$ and hence $u \in W^M$ with wRu such that $(M,u) \not\models A$. So $M \not\models A$. Q.E.D.

Syntactical substitutions are semantically mirrored by corresponding variations of the valuation function. This is the content of the following *substitution lemma*: Define, for arbitrary substitution function s and valuation function V, $V_s(p) = V(s(p))$, for all $p \in \mathcal{P}$; and for given $M = \langle W,R,V\rangle$, let $M_s = \langle W,R,V_s\rangle$; thus M and M_s are based on the same frame. Then: For every $A \in \mathcal{L}$, M and $w \in W^M$: $(M,w) \models s(A)$ iff $(M_s,w) \models A$

PROOF *Exercise*: By induction on formula complexity; (see, for example, van Benthem 1983: 27, Lemma 2.5).

PROOF THAT (SUBST) IS FRAME-ADMISSIBLE By contraposition. Assume, for arbitrary $F = \langle W,R\rangle$ and s, that $F \not\models s(A)$. Thus there exists $M = \langle W,R,V\rangle$ based on F and $w \in W^M$ such that $(M,w) \not\models s(A)$. By *substitution lemma*, $(M_s,w) \not\models A$, where M_s is based on F. Thus $F \not\models A$. Q.E.D.

Closure under substitution is an important condition on logics. Expressing theorems as *schemata* (with schematic letters A, B, ... ranging over arbitrary formulas) is a simple means of asserting that the theorems of a logic are closed under substitution.

For example, the set of all schematic instances of the formula schema $\Box A \supset A$ equals the set of substitution instances of the formula $\Box p \supset p$. The preservation properties of rules are summarized as follows (Schurz 1997: 52):

Rule	preserves: *truth at a world*	*model-validity*	*frame-validity*
TautR	+	+	+
N, E, RE	–	+	+
Subst	–	–	+

It is also important to prove that certain formulas are not theorems of a logic. This is usually done by giving *semantic counterexamples.* For example, the logic **K** does not contain the theorem T: $\Box A \supset A$, which says that whatever is necessarily true is also true. T seems intuitively to be an indispensable meaning postulate for 'necessity.' A *countermodel* for T is, for example, the Kripke frame F with $W = \{u,v\}$, $R = \{\langle u,v\rangle\}$ (graphically displayed as $u \rightarrow v$), with a valuation function $V(p) = \{v\}$ (it suffices to define V for the variables of the evaluated formula; this is often expressed as a lemma, cf. van Benthem 1983: 26, 2.4). We have $(M,v) \models p$ and thus $(M,u) \models \Box p$, but $(M,u) \not\models p$, and so, $M \not\models T$, whence $F \not\models T$ and thus $T \notin \mathbf{K}$.

EXERCISE Give countermodels for $\Box(A \vee B) \supset \Box A \vee \Box B$ (the converse of i) and for $\Diamond A \wedge \Diamond B \supset \Diamond(A \wedge B)$ (converse of ii).

We obtain stronger logics than **K** by imposing structural conditions on frames. The logic $\mathbf{T} = \mathbf{K} + T$ is semantically obtained by requiring frames to be *reflexive*, that is to satisfy the frame condition *Ref*: $\forall w: wRw$. We make this more precise by assuming a (first or higher order) *quantiflcational* language $\mathcal{L}(R)$ which contains the accessibility relation $\underline{R}$ as its only nonlogical predicate and has models of the form $\langle W,R\rangle$. $\models_R$ denotes the standard notion of verification for $\mathcal{L}(R)$-formulas (note that in $\mathcal{L}(R)$-contexts, 'uRv' abbreviates '$\underline{R}xy$'). Then we obtain:

T-CORRESPONDENCE THEOREM For every frame F: $F \models T$ iff $F \models_R$ *Ref.*

PROOF *Right-to-left:* We show that if F is reflexive, then $\Box A \supset A$ is true on every world $w \in W^M$ in every model M based on F. Assume $(M,w) \models \Box A$. Hence $\forall u: wRu \Rightarrow (M,u) \models A$. Since F is reflexive, wRw, so $(M,w) \models A$. Hence $(M,w) \models \Box A \supset A$. *Left-to-right:* We show, by contraposition, that if a given $F = \langle W,R\rangle$ is not reflexive, then we can construct a countermodel on F refuting the T-instance $\Box p \supset p$. So assume $w \in W^F$ is an irreflexive point, that is $\neg wRw$. Let p be true at all u with wRu but false at w ($\{u: wRu\} \subseteq v(p)$ and $w \notin v(p)$). Call the resulting model M. Now, $(M,w) \models \Box p$, but $(M,w) \not\models p$; so $(M,w) \not\models \Box p \supset p$. Hence $F \not\models T$. Q.E.D.

This is an example of a *correspondence* result. It tells us that the frame condition *Ref* can be *deflned* by (or *translated* into) the modal formula T, and *vice versa*. Generally, we say that a modal formula or formula schema $X \in \mathcal{L}$ *corresponds* to a (first or higher order) frame-condition $C_X \in \mathcal{L}(R)$ iff $\forall F \in \mathbf{KF}$: $F \models X \Leftrightarrow F \models_R C_X$. In this case, the frames of the modal logic $\mathbf{K} + X$ (obtained from **K** by adding the axiom schema X) are exactly all

frames satisfying C_X. A modal formula (schema) X is called *elementary*, or *first-order definable*, iff X corresponds to a first-order condition C_X. *Correspondence theory* is the field which explores the possibilities of intertranslations between modal and quantificational logic (van Benthem 1983, 1984; our notion of 'correspondence' corresponds to van Benthem's 'global equivalence'; 1983: 48f).

Correspondence results for the five standard principles of alethic PMLs

Modal Principle:	*Frame Condition:*
D: $\neg\Box\bot$	*Ser:* $\forall u\exists v$: uRv (R serial)
T: $\Box A \supset A$	*Ref:* $\forall w$:wRw (R reflexive)
4: $\Box A \supset \Box\Box A$	*Trans:* $\forall u,v,w$: uRv $\wedge$ vRw $\supset$ uRw (R transitive)
B: $\Diamond\Box A \supset A$	*Sym:* $\forall u,v$: uRv $\supset$ vRu (R symmetric)
5: $\Diamond A \supset \Box\Diamond A$	*Euc:* $\forall u,v,w$: wRu $\wedge$ wRv $\supset$ uRv (R euclidean)

PROOF *Exercise* (Chellas 1980: ch. 3.2).

Correspondence results hold only w.r.t. frames; they do *not* say that every *model* validating axiom X satisfies the corresponding frame-condition C_X. It is easy to construct models for a logic **L** which are *not* based on a frame for **L**. We call such models *non-standard* **L**-models. For example, a *irreflexive* ***T*-model** can be constructed by taking an irreflexive two world frame $u \leftrightarrow v$ where both worlds access each other, and by defining a valuation function V which agrees on both worlds, that is for all $p \in \mathcal{P}$, $u \in V(p)$ iff $v \in V(p)$. It is easily proved for the so obtained model M, by induction on formula-complexity, that $(M,u) \models A$ iff $(M,v) \models A$ holds for all $A \in \mathcal{L}$. Thus, all instances of T are verified on both worlds of M.

The above correspondence results imply that if a PML contains several modal principles, its frames will satisfy all of the corresponding frame conditions. According to the *Lemmon-code* (Lemmon and Scott 1966; see Bull and Segerberg 1984: 20f), we denote normal PMLs as '**KX** . . . ," where **X** is a set of additional axiom schemata for these logics (except for special names for logics like **T**, **S4**, or **S5**). Principle D has been suggested for deontic logics by von Wright. T was suggested by Feys and von Wright for the alethic logic **KT** (von Wright (1951) calls it M). B refers to the 'Browersche system' **KTB** and 4 to Lewis **S4 = KT4** (both B and 4 have been suggested by Becker), and finally 5 refers to **S5 = KT5**. Observe the following implication relations between frame-conditions and corresponding logics: (i) *Ref* $\Rightarrow$ *Ser*, thus **KT = KDT**; (ii) Sym $\Rightarrow$ (Trans $\Leftrightarrow$ Euc), thus **KB5 = KB4 = KB45**; (iii) *Trans* $\Rightarrow$ (*Euc* $\Leftrightarrow$ (*Sym* $\wedge$ *Trans*)), thus **KT5 = KTB4 = KDTB45 = S5**; (iv) *Ser* $\wedge$ *Sym* $\Rightarrow$ *Ref*, thus **KDB = KDTB**; (v) *Ser* $\wedge$ *Sym* $\Rightarrow$ (Trans $\Leftrightarrow$ Euc), thus **KDB4 = KDB5 = KDB45 = KDTB45 = S5**.

PROOF *Exercise* (Chellas 1980: 164).

The possible combinations of these five principles produce 15 *mutually nonequivalent* standard systems of PML (Chellas 1980: 132). Various theorems of PML's stronger than **K** are found in Hughes and Cresswell (1968: ch. 2–4) and Chellas (1980: 131ff); here are some of them.

EXERCISE – PROVE SEMANTICALLY (i) $A \supset \Diamond A$, $\Box^n A \supset A$, $A \supset \Diamond^n A \in$ **KT** ($\Box^n =_{df} \Box \ldots \Box$ n times iterated); (ii) $\Box(\Diamond A \supset B) \supset (A \supset \Box B) \in$ **KB**; (iii) $\Box A \equiv \Box\Box A$, $\Diamond A \equiv \Diamond\Diamond A$, $\Box\Diamond\Box\Diamond A \equiv \Box\Diamond A$, $\Diamond\Box \equiv \Diamond\Box\Diamond\Box A \in$ **S4**; (iv) $\Diamond\Box A \equiv \Box A$, $\Box\Diamond A \equiv \Diamond A \in$ **S5**. – A *modality* m is a possibly empty sequence of $\Box$s and/or $\Diamond$s. Two modalities m_1 and m_2 are **L**-equivalent iff $m_1 p \equiv m_2 p \in$ **L**. The stronger a PML, the more modalities collapse, that is become equivalent. In **S5**, all iterations of modal operators are either equivalent with '$\Diamond$' or with '$\Box$', thus **S5** has only three modalities, namely $\Box - \emptyset - \Diamond$. For modalities in other systems cf. Chellas (1980: 147ff).

Important semantical operations which *preserve* truth and validity of formulas are the formation of *generated submodels* (*subframes*) and *disjoint unions* of models (frames). This follows from the fact that the truth of $\Box$-formulas in a world w depends only on that part M_w of the given model M which is reachable from w by an R-chain. M_w is called the *w-generated submodel* of $M = \langle W,R,V\rangle$ and is defined as $\langle W_w, R_w, V_w\rangle$, with $W_w = \{u \in W: wRu\}$, where $\underline{R}$ is the transitive closure of R, $R_w = R \cap (W_w \times W_w)$, and $V_w(p) = V(p) \cap W_w$ (for all $p \in \mathcal{P}$). The *w-generated subframe* of F is defined accordingly as $F_w = \langle W_w, R_w\rangle$. If **M** is a class of models with pairwise disjoint world sets, then the *disjoint sum* of the models in **M** is defined as $DS(\mathbf{M}) = \langle \cup\{W^M: M \in \mathbf{M}\}, \cup\{R^M: M \in \mathbf{M}\}, \cup\{V^M: M \in \mathbf{M}\}\rangle$; and likewise for DS(**F**). It is straightforward to prove for all formulas A, models M and $w \in W^M$, (i) $(M,w) \models A$ iff $(M_w,w) \models A$, and hence $M \models A$ iff $M_w \models A$, and $F \models A$ iff $F_w \models A$, and (ii) for all model-classes **M** (or fames-classes **F**) with pairwise disjoint world sets, $\mathbf{M} \models A$ iff $DS(\mathbf{M}) \models A$ (or, $\mathbf{F} \models A$ iff $DS(\mathbf{F}) \models A$). A third truth- and validity-preserving operation is the formation of *p-morphic copies* of models and frames. It generalizes the notion of *isomorphic* copy and was introduced by Segerberg (1971: 37; also called 'contraction' by Rautenberg, 'zigzag morphism' by van Benthem and 'reduction' by Chagrov and Zakharyaschev 1997).

PROOF *Exercise* (Hughes and Cresswell 1986: 72f, 80; Chagrov and Zakharyaschev 1997: ch. 2.3).

A model (or frame) which validates the formula set Γ is called a model (or frame) *for* Γ. $\mathbf{M}(\Gamma)$, $\mathbf{F}(\Gamma)$ denote the set of models, or frames respectively, for Γ. The above results tell us that, for every Γ, the sets $\mathbf{M}(\Gamma)$ and $\mathbf{F}(\Gamma)$ are closed under the formation of generated submodels (subframes), disjoint unions of frames, and p-morphic models (frames). Preservation results of this sort have various important consequences. A simple example are **S5**-frames. Their accessibility relation is reflexive, symmetric, and transitive and, hence, an *equivalence* relation: it imposes a *partition* onto the world set W into mutually disjoint and exhaustive 'cells' (subsets) $W_1, \ldots, W_n$ (i.e. $W_i \cap W_j = \emptyset$, $\cup_i W_i = W$), such that all worlds in the same cell are mutually accessible, and are inaccessible to worlds in different cells. Hence, each **S5**-frame is the disjoint sum of *universal* frames $\langle W_i, R_i = W_i \times W_i\rangle$. They correspond to Carnap's and Kripke's original **S5**-frames without a relation R. It follows from the generated subframe theorem that all universal frames are in $\mathbf{F}(\mathbf{S5})$; that is **S5** is valid in all universal frames.

A final word on philosophical plausibility. Assume we understand possible worlds as variations of the real world which are possible relative to some *background theory*. If this background theory is logic, then $\Box A$ means that A's truth is determined by princi-

ples of logic alone. In this interpretation, all principles of **S5** seem to be valid, in particular all modal iteration principles. For if it is determined by logic that A is true, then it is also determined by logic – namely by *metalogic* – that A's truth is determined by logic: hence $\Box A \supset \Box\Box A$ holds. Likewise, if it is not determined by logic that A is false, then this fact is itself determined by logic: so $\Diamond A \supset \Box\Diamond A$ holds. The same reasoning applies if the background theory contains logic + laws of physics. Then $\Box A$ means that A's truth is determined by logic + laws of physics alone – so, also in this interpretation of $\Box A$, **S5** seems to be the right choice. To avoid confusions: of course, physical necessity is *stronger* than logical necessity, but modal *logics* contain only those principles which are closed under substitution and, hence, are independent from the *content* of a nonlogical symbol. *Proper* physical necessity statements such as 'it is necessary that everything is composed of matter' are content-specific and thus not part of a modal *logic*.

In the above understanding of '$\Box$' we must assume, in order to interpret *iterated* modalities, that either the language of our background theory is *closed* (i.e. it can speak about the truth of its own sentences; cf. ch. VIII of this volume), or that it contains a potentially infinite hierarchy of metalanguages. If we do not make these strong assumptions, then our interpretation will validate only the weaker logic **T**. If we are even more scrupulous and interpret truth-determination as syntactical *derivability* from our background theory, then even **T** will be too strong, as soon as our background theory contains *arithmetic*; the adequate logic for this interpretation is **G**, discussed below. If, on the other hand, '$\Box$' is interpreted in the sense of *tense logic* as 'being true in all future times', then only **S4** but not **S5** is the philosophically adequate logic. These remarks show that even in the narrow realm of *alethic* modal logics, different PMLs are needed for different interpretational purposes.

Axiomatic systems: correctness, completeness, and correspondence

The standard axiomatization of the minimal normal MPL, **K**, is its definition as the smallest set of $\mathcal{L}$-formulas which contains all instances of the axiom schemata Taut and K and is closed under the rules MP and N. The stronger alethic logics **KX**, with $\mathbf{X} \subseteq \{D,T,B,4,5\}$, are axiomatized by adding **X** as the set of so-called *additional* axiom schemata. A formula A is provable in logic **L**, in short $\vdash_{\mathbf{L}} A$, iff A has an **L**-proof, which is a sequence $\langle B_1, \ldots, B_n \rangle$ of formulas such that $B_n = A$, and every B_i $(1 \leq i \leq n)$ is either instance of an axiom schema of **L** or follows from previous members of the sequence by one of the rules of **L**. If $\vdash_{\mathbf{L}} A$, we also call A a theorem of **L**, and identify **L** with the set of its theorems: $\mathbf{L} = \{A\colon \vdash_{\mathbf{L}} A\}$. A formula A is said to be deducible from formula set Γ, in short $\Gamma \vdash_{\mathbf{L}} A$, iff $\vdash_{\mathbf{L}} \wedge \Gamma_f \supset A$ for some finite subset $\Gamma_f \subseteq \Gamma$. In particular, $\vdash_{\mathbf{L}} A$ iff $\varnothing \vdash_{\mathbf{L}} A$. Finally, Γ is called **L**-consistent iff $\Gamma \nvdash_{\mathbf{L}} p \wedge \neg p$, and **L** is called consistent iff $p \wedge \neg p \notin \mathbf{L}$.

The above axiomatization of PML's is a *Hilbert-style* axiomatization of their modal part together with *an unspecified* (syntactic) determination of tautology-hood. It is rather common for PMLs. If we additionally allow the application of tautological rules TautR, we obtain a *comfortable* way of proving theorems (see exercise below). Of course, various alternative but equivalent axiomatizations are possible. To highlight the relation of **K** to weaker PML's, **K** may equivalently be axiomatized by rules MP + E and

axiom schemata Taut + M + C + T (exercise below). $\Gamma \vdash_{\mathbf{L}} A$ may be equivalently defined by the existence of a proof of A from axioms in $\mathbf{L} \cup \Gamma$ with modal rules restricted to formulas which do *not* depend on Γ (see Schurz 1997: 53). There exist also various non-Hilbert-style axiomatizations for PMLs, such as sequent or tableau calculi (see, for example, Fitting 1983; Wansing 1996).

EXERCISE (1.) Prove the theorems of **K**, **T**, **S4** and **S5** listed in § 2.2. Prove that axioms Taut + M + C + T and rules TautR + E are an equivalent axiomatization of **K** (see, for example, Chellas 1980: ch. 4, ch. 8; see also example below).

Example: proof of K from Taut + M + C + TautR + E:

1. $(\Box(A \supset B) \wedge \Box A) \supset \Box((A \supset B) \wedge A)$	C-instance
2. $((A \supset B) \wedge A) \equiv (A \wedge B)$	Taut
3. $\Box((A \supset B) \wedge A) \equiv \Box(A \wedge B)$	E from 2
4. $(\Box(A \supset B) \wedge \Box A) \supset \Box(A \wedge B)$	TautR from 1,3
5. $\Box(A \wedge B) \supset \Box B$	M-instance
6. $\Box(A \supset B) \supset (\Box A \supset \Box B)$	TautR from 4,5

PROOF OF N FROM TAUT + M + C + $\mathcal{T}$ + TAUTR + E Assume $\vdash_{\mathbf{L}} A$. Hence $\vdash_{\mathbf{L}} A \equiv \mathsf{T}$, by TautR. So $\vdash_{\mathbf{L}} \Box A \equiv \Box\mathsf{T}$, by E. Because $\Box\mathsf{T}$ is an axiom, we get $\vdash_{\mathbf{L}} \Box A$ by TautR. Q.E.D.

In general, a *normal* PML (i.e. a normal extension of **K**) is defined (after Lemmon and Scott 1966) as any subset $\mathbf{L} \subseteq \mathcal{L}$ which contains **K** and is closed under the rules MP, N, and Subst. Clearly, every normal PML **L** is *representable* as $\mathbf{L} = \mathbf{KX}$, where **X** is *some* set of additional axiom schemata **X** (Schurz 1997: 50, lemma 4). If **X** is recursively enumerable, then **KX** is called (recursively) *axiomatizable* (Chagrov and Zakharyaschev 1997: 495f); if **X** is finite, **KX** is finitely axiomatizable. The class of a normal PMLs forms the infinite lattice Π with **K** as its bottom and $\mathcal{L}$ = the inconsistent logic as its top (for various results on this lattice cf. Chagrov and Zakharyaschev 1997). Not all $\mathbf{L} \in \Pi$ are axiomatizable.

The major properties of axiomatized logics (i.e. axiomatic systems) are their *correctness* and *completeness*. Generally, an axiomatic system L is correct w.r.t. an underlying semantics S iff everything what is L-provable is S-valid, and L is complete w.r.t. S iff everything what is S-valid is L-provable. In modal logics, these notions can be defined w.r.t. models as well as w.r.t. frames, as follows:

1.	**L** is *correct* w.r.t. **F**	iff	$\vdash_{\mathbf{L}} A \Rightarrow \models_{\mathbf{F}} A$ (for all A)
	L is *correct* w.r.t. **M**	iff	$\vdash_{\mathbf{L}} A \Rightarrow \models_{\mathbf{M}} A$ (for all A)
2.	**L** is *w.(eakly) complete* w.r.t. **F**	iff	$\models_{\mathbf{F}} A \Rightarrow \vdash_{\mathbf{L}} A$ (for all A)
	L is *w. complete* w.r.t. **M**	iff	$\models_{\mathbf{M}} A \Rightarrow \vdash_{\mathbf{L}} A$ (for all A)
3.	**L** is *s.(trongly) complete* w.r.t. **F**	iff	$\Gamma \models_{\mathbf{F}} A \Rightarrow \Gamma \vdash_{\mathbf{L}} A$ (for all Γ, A)
	L is *s. complete* w.r.t. **M**	iff	$\Gamma \models_{\mathbf{M}} A \Rightarrow \Gamma \vdash_{\mathbf{L}} A$ (for all Γ, A)
4.	**L** is w./s. *frame-complete* (simpliciter)	iff	**L** is w./s. complete w.r.t. **F**(**L**).
	L is w./s. *model-complete* (simpliciter)	iff	**L** is w./s. complete w.r.t. **M**(**L**).

Completeness *simpliciter* is defined w.r.t. the class of *all* frames or models for a logic. Correctness is the converse property of *weak* completeness. A separate notion of

'strong' correctness ($\Gamma \vdash_{\mathbf{L}} A \Rightarrow \Gamma \models_{\mathbf{M/F}} A$) is not needed: weak correctness implies strong correctness because '$\vdash_{\mathbf{L}}$' is by definition *finitary* ($\Gamma \vdash_{\mathbf{L}} A$ iff $\vdash_{\mathbf{L}} \wedge \Gamma_f \supset A$). Correctness of an axiomatic system **L** is standardly proved as follows: one demonstrates (1) that every axiom of **L** is valid, and (2) that every rule of **L** preserves validity, and concludes, by *induction on the length of the* ***L****-proof of A*, that A is valid w.r.t. the given **M** or **F**. Claim 2 has been established for all normal PMLs, and claim 1 for all PMLs **KX** with $\mathbf{X} \subseteq$ {D,T,B,4,5}. Hence all the latter PMLs are correct w.r.t the corresponding classes of models and frames. Moreover, all PMLs representable as **KX** are correct w.r.t. **M**(**X**) and **F**(**X**).

The standard technique to prove completeness rests on the following *consistency-formulation* of completeness which is classically equivalent:

CONSISTENCY-LEMMA **L** is w. complete w.r.t. **M** [or **F**] iff every **L**-consistent formula A is satisfiable in **M** [or **F**, resp.]; and **L** is s. complete w.r.t. **M** [or **F**] iff every **L**-consistent formula *set* Γ is satisfiable w.r.t. **M** [or **F**, resp.]. *Proof: Exercise.*

Strong completeness is stronger than weak completeness because semantical consequence is not by definition finitary. Strong frame- (or model-) completeness of **L** implies frame- (or model-) *compactness* of **L** in the sense that a formula set Γ is satisfiable on an **L**-frame [or in an **L**-model, resp.] whenever every finite subset of Γ is satisfiable on this **L**-frame [or in that **L**-model, resp.]. Weak completeness plus compactness imply strong completeness. If an axiomatic system **L** is correct and w./s. complete w.r.t. a given class **F** or **M**, then it is said to be w./s. *characterized* by **F** or **M**. This means that the syntactic definition of **L** (and of $\vdash_{\mathbf{L}}$ in the case of s. completeness) coincides with the semantic one.

The canonical technique of proving model-completeness of a normal PML has been introduced by Lemmon and Scott (1966) and Makinson (1966). It is an adaptation of the 'Lindenbaum–Gödel–Henkin' technique to modal logics. It consists in the construction of the so-called *canonical model* $M_c(\mathbf{L})$ of the given logic **L**, which contains maximally consistent formula sets, that is maximal state descriptions, as its worlds (cf. Hughes and Cresswell 1984: 22f; Chellas 1980: 173, def. 5.9):

DEFINITION OF THE CANONICAL MODEL (1) A formula set Γ is maximally **L**-consistent iff Γ is **L**-consistent and no proper extension Δ of Γ is **L**-consistent. (2) The canonical model $M_c(\mathbf{L})$ of **L** (in the given denumerably infinite language $\mathcal{L}$) is defined as $\langle W_c, R_c, V_c \rangle$ where (2.1) W_c is the class of all maximally **L**-consistent formula sets, (2.2) R_c is defined by $\forall u,v \in W_c$: uR_cv iff $\{A: \Box A \in v\} \subseteq u$, and (2.3) for all $p \in \mathcal{P}$, $V_c(p)$ is defined by $\forall w \in W_c$: $w \in V_c(p)$ iff $p \in w$.

It is well-known from truth-functional logic that maximally **L**-consistent formula sets enjoy the following *maximality* properties:

MAXIMALITY-LEMMA For all maximally **L**-consistent sets Δ and formulas A, B: (1) $\Delta \vdash_{\mathbf{L}} A$ implies $A \in \Gamma$ (deductive closure), (2) either $A \in \Delta$, or $\neg A \in \Delta$ (completeness), and (3) $(A \vee B) \in \Delta$ iff ($A \in \Delta$ or $B \in \Delta$) (primeness). Analogous properties exist for $\wedge$ and $\supset$. *Proof: Exercise* (cf. Hughes and Cresswell 1984: 18f).

LINDENBAUM-LEMMA Every **L**-consistent formula set Γ can be extended to a maximally **L**-consistent formula set $\Delta \supseteq \Gamma$ (*proof* cf. Hughes and Cresswell 1984: 19f).

The central idea of the canonical model construction is to prove the following:

TRUTH LEMMA For all $A \in \mathcal{L}$ and $w \in W_c$: $(M_c,w) \models A$ iff $A \in w$.

The three lemmata imply strong completeness of **L** as follows. By maximality lemma 1, **L** is a subset of every world of W_c. So by truth lemma, M_c is an **L**-model. By Lindenbaum-lemma, every given **L**-consistent Γ is subset of some world in W_c. So by truth lemma, Γ satisfied in the **L**-model M_c; Q.E.D. To prove the truth lemma we need the following lemma which guarantees that R_c is well-behaved in the sense that whenever $\Diamond A \in w \in W_c$, then $\exists u$ with wR_cu and $A \in u$:

CANONICAL MODEL LEMMA (1) If $\neg\Box B \in \Gamma$ and Γ is **L**-consistent, then $\{A: \Box A \in \Gamma\} \cup \{\neg B\}$ is **L**-consistent, too. (2) $\forall u \in W_c$: $\Box A \in u$ iff $\forall v$: $uR_cv \Rightarrow A \in v$.

PROOF *Exercise* (Hughes and Cresswell 1984: 21f; Chellas 1980: 172).

PROOF OF THE TRUTH-LEMMA We prove the claim by induction on the complexity of $\mathcal{L}$-formulas. (1) For $A = p \in \mathcal{P}$: $(M_c,w) \models p$ iff $p \in w$ holds for all $w \in W_c$ by definition of V_c. (2) For $A = \neg B$: $(M_c,w) \models \neg B$ iff $(M_c,w) \not\models B$ iff $B \notin w$ by induction hypothesis, iff $\neg B \in w$ by maximality lemma, 2. (3) For $A = B \vee C$: $(M_c,w) \models B \vee C$, iff $(M_c,w) \models B$ or $(M_c,w) \models C$, iff $B \in w$ or $C \in w$ by induction hypothesis and propositional logic, iff $(B \vee C) \in w$ by maximality lemma, 3. (4) For $A = \Box B$: $(M_c,w) \models \Box A$ iff $\forall u$: $wR_cu \Rightarrow (M_c,u) \models A$, iff $\forall u$: $wR_cu \Rightarrow A \in u$ by induction hypothesis and first-order logic, iff $\Box A \in w$ by canonical model lemma, 2. Q.E.D.

The foregoing proofs hold for all normal PMLs and thus establish:

PML-MODEL-CHARACTERIZATION-THEOREM Every normal PML **L** is strongly model-complete, and is strongly characterized by $\mathbf{M}(\mathbf{L})$.

Frame-completeness is stronger than model-completeness: it implies not only that every **L**-consistent formula (set) is satisfiable in *some* **L**-model, but that it is satisfiable in a *standard* **L**-model. So, to prove that **L** is s. frame-complete requires to prove something additional, namely: that the frame of $M_c(\mathbf{L})$ is a frame for **L**. Following Fine (1975a), we call normal PMLs satisfying this condition *canonical* (in general, canonicity is relativized to the cardinality of $\mathcal{P}$; but we always assume that $\mathcal{P}$ is denumerably infinite). Canonicity implies strong frame-completeness; whether the reverse direction holds is an open question. Clearly, **K** is canonical because $M_c(\mathbf{K})$ is based on a frame. For stronger systems, canonicity has to be proved for each additional axiom schema separately. Axiom schema X is called canonical iff the frame of $M_c(\mathbf{L})$ is a frame for **L** whenever **L** contains X. If $X_1, \ldots, X_n$ are canonical, then every $\mathbf{KX}$ with $\mathbf{X} \subseteq \{X_1, \ldots, X_n\}$ will be canonical, too.

CANONICITY-THEOREM D, T, B, 4 and 5 are canonical.

PROOF *Exercise* (Hughes and Cresswell 1984: ch. 2; Chellas 1980: ch. 5.4).

EXAMPLE *Proof of canonicity of 4:* Assume $4 \in \mathbf{L}$. To show that for $\forall u,v,w$ in $M_s(\mathbf{L})$, $uR_sv \wedge vR_sw$ implies uR_sw, we assume (a) $\{A: \Box A \in u\} \subseteq v$, (b) $\{A: \Box A \in v\} \subseteq w$, and prove thereof that (c) $\{A: \Box A \in u\} \subseteq w$. Take any $\Box B \in u$. By deductive closure of canonical worlds, $\Box B \supset \Box\Box B \in u$, and thus $\Box\Box B \in u$. So $\Box B \in v$ by (a) and $B \in w$ by (b).Thus for all $\Box B \in u$, $B \in w$, which is exactly (c). Q.E.D.

In general, if a normal PML **L** is correct w.r.t. **F**, then it is also correct with respect to every subclass $\mathbf{F}' \subset \mathbf{F}$; and if it is w./s. complete w.r.t. **F**, then it is also w./s. complete w.r.t. every superclass $\mathbf{F}' \supset \mathbf{F}$ (and likewise for models). It often happens that a normal MPL **L** which is characterized by $\mathbf{F}(\mathbf{L})$ is *also* characterized by an interesting subclass $\mathbf{F}' \subset \mathbf{F}(\mathbf{L})$. For example, every canonical $\mathbf{L} \in \Pi$ is strongly characterized by a single frame, namely the frame of the canonical model $M_s(\mathbf{L})$ (this follows direct from the completeness proof). Or, every $\mathbf{L} \in \Pi$ is w./s. characterized by the class of its generated subframes (which follows from the generated subframe lemma). Another way of producing characteristic subclasses of $\mathbf{F}(\mathbf{L})$ is based on the fact that the following first-order conditions on frames cannot be expressed by modal formulas: *Irr* $\neg wRw$ (irreflexivity), *Asym* $uRv \supset \neg vRu$ (asymmetry), *Antisym* $uRv \wedge u \neq v \supset \neg vRu$ (antisymmetry), *Intrans* $uRv \wedge vRw \supset \neg uRw$, and *Anticon* $\forall u,v,w: u \neq v \neq w \wedge uRw \supset \neg vRw$ (anticonvergence). In other words, *correspondence* fails for these conditions in the right-to-left direction. For **K**, this can be proved by the technique of *unraveling*, which is a validity-preserving transformation of arbitrary models into irreflexive, asymmetric, and intransitive models (due to Dummett and Lemmon 1959; see Bull and Segerberg 1984: 45). Thus, **K** is also strongly characterized by all irreflexive, asymmetric, and intransitive frames.

Characterization by subclasses is important for the PML's of *ordering relations*. The technique of *bulldozing* introduced by Segerberg (1971: 78ff) transforms every reflexive and transitive model M into a validity-preserving partially ordered model M; if M is merely transitive then the ordering of M's frame is strict. It follows from this that **K4** is strongly characterized by the class of strict partially ordered frames, and **S4** by the class of partially ordered frames. Finally, *ramification* transforms arbitrary [reflexive, transitive] models into validity-preserving [reflexive, transitive, resp.] models based on *tree-frames* (Chagrov and Zakharyaschev 1997: 32–5). Tree models represent branchings of possible future states in time and are important for the logic of causality and agentship (cf. Kutschera 1993, Prendinger and Schurz 1996).

A brief remark on *classical* modal logics concludes this section. They are weaker than **K** and are mainly used for nonalethic (e.g. epistemic, deontic) interpretations of the modal operator (for details see Segerberg 1971: ch. 1 and Chellas 1980: part III). The minimal classical modal logic, **E**, is axiomatized by Taut, MP, and the rule E. **E** allows it to regard the intensional operator $\Box$ as applying to *propositions*; this requires truth-preservation of $\Box$ under replacements of logically equivalent sentences. $\mathbf{M} = \mathbf{E} + M$ is the minimal *monotonic* and $\mathbf{C} = \mathbf{E} + M + C$ the minimal *regular* modal logic. If we finally add **T** we get an alternative axiomatization of **K**. Semantically, classical modal logics

are characterized by so-called *neighborhood* frames, which are pairs $\langle W,N\rangle$ with $W \neq \phi$ a possible world set and $N:W \rightarrow Pow(Pow(W))$ a function assigning to each world $w \in W$ a neighborhood $N(w)$ which contains exactly those 'propositions' (i.e. W-subsets) which are necessarily true at w. Completeness proofs proceed via canonical neighborhood models: **E** is characterized by the class of all neighborhood frames, and the semantic conditions corresponding to **M**, **C**, and **T** are closure of neighborhoods under supersets, finite intersections, and containment of W. **K**-neighborhood-frames with all three properties can be transformed into point-wise equivalent Kripke frames, and **C**-neighborhood frames can be transformed into Kripke-frames with an additional set of so-called 'queer' worlds (Segerberg 1971: 23ff).

Decidability and finite model property

A logic $\mathbf{L} \subseteq \mathcal{L}$ is called *decidable* iff for every $A \in \mathcal{L}$ it can be decided after a finite number of primitive computation steps whether or not $A \in \mathbf{L}$. Of course, the mere axiomatizability (i.e. recursive enumerability) of a logic does *not* imply its decidability. It is a famous fact that a logic **L** is decidable iff the theorems of **L** as well as the non-theorems of **L** (i.e. the elements of $\mathcal{L} - \mathbf{L}$) are recursively enumerable (Chagrov and Zakharyaschev 1997: 492). A logic **L** is said to have the *finite model property* (f.m.p.) iff every **L**-consistent formula A is satisfiable on a *finite* **L**-model. (Likewise for the 'finite frame-property.') A standard way of proving the decidability of an axiomatizable logic is by proving that it has the f.m.p. For, we can effectively enumerate all finite models of a given formula A and test whether they are A-countermodels. So if an axiomatizable logic **L** has the f.m.p., then after a finite number of steps either the enumeration of **L**-theorems will output A, or the enumeration of $\mathcal{L}$'s finite models will produce an A-countermodel (Chagrov and Zakharyaschev 1997: 492). Note, however, that there are also decidable logics which do *not* have the f.m.p. (Gabbay 1976: 258–65).

Note the following fundamental *f.m.p.-theorem:* For all $\mathbf{L} \in \Pi$: **L** has the f.m.p. $\Leftrightarrow$ **L** has the finite frame property $\Leftrightarrow$ **L** is w. complete w.r.t. **L**'s finite frames. The second equivalence is an immediate consequence of the first, which has been proved by Segerberg (1971: 29ff) as follows: (1) for every model there exists an elementary equivalent *distinguishable* model (where no two worlds verify the same formulas); and (2) if a finite distinguishable models validates **L**, then its frame is an **L**-frame.

A standard technique to produce finite models for a given formula or finite formula set is *filtration*. Assume Γ is a set of formulas closed under subformulas (i.e. if $A \in \Gamma$, and B is a subformula of A, then $B \in \Gamma$). Given $M = \langle W,R,V\rangle$, two worlds $u, v \in W^M$ are called Γ-equivalent, in short $u \equiv_\Gamma v$, iff they verify the same formulas in Γ (i.e. iff $\forall A \in \Gamma$: $(M,u) \models A \Leftrightarrow (M,v) \models A$). For $w \in W^M$, $[w]_\Gamma =_{df} \{u \in W^M: w \equiv_\Gamma u\}$ denotes the Γ-equivalence class of w. Then, a model $M_\Gamma = \langle W_\Gamma, R_\Gamma, V_\Gamma\rangle$ is called a *Γ-filtration of M*, iff $M_\Gamma = \{[w]_\Gamma: w \in W^M\}$, $V_\Gamma(p) = \{[w]_\Gamma: w \in v(p)\}$ for all $p \in \mathcal{P}$, and R satisfies two conditions ($u, v \in W^M$): (F1): If uRv, then $[u]_\Gamma R_\Gamma [v]_\Gamma$, and (F2): If $[u]_\Gamma R_\Gamma [v]_\Gamma$, then $\forall A \in \Gamma$: $(M,u) \models \Box A \Rightarrow (M,v) \models A$. Note that there exist several Γ-filtrations of a given model M. The frame $\langle W_\Gamma, R_\Gamma\rangle$ is the corresponding Γ-filtration of $\langle W,R\rangle$.

FILTRATION THEOREM If M_Γ is a Γ-filtration of M, then $\forall A \in \Gamma\ \forall w \in A^M$: $(M,w) \models A$ iff $(M_\Gamma, [w]_\Gamma) \models A$. *Proof: Exercise* (Hughes and Cresswell 1984: 139).

M_Γ is a finite model whenever Γ is finite. Thus, by filtering a model M for an **L**-consistent formula A through the (finite) set *subf(A)* of A's subformulas, we obtain a finite model $M_{subf(A)}$ verifying A. To prove by this method that **L** has the f.m.p. requires in addition to prove that the filtered model is indeed an **L**-model. A simple way to do this is to show that the filtered frame $\langle W_{subf(A)}, R_{subf(A)}\rangle$ is a frame for **L**. This is easy for the logics **K**, **KD**, and **KT**, since one can prove that every filtration of a frame preserves seriality or reflexivity (Chellas 1980: 105). For other standard systems such as **KB**, **K4**, **S4**, etc., special filtrations are necessary to demonstrate preservation of the corresponding frame-properties (Chellas 1980: 106ff). Segerberg has proved that all normal extensions of **K45** have the f.m.p. and, thus, can be classified as the logics of certain simple frame classes (Segerberg 1971: 123ff).

We finally remark that, though f.m.p. proves decidability for axiomatizable logics, it does not produce a practically feasible decision method. Practically feasible decision methods for standard systems are, for example, *tableau methods* (Hughes and Cresswell 1968: ch. 5–6; Chagrov and Zakharyaschev 1997: ch. 3.4).

More metalogical results on PMLs

Further examples of axiom schemata which are both first-order definable and canonical are:

Axiom schema:	*Corresponding first-order condition:*
$(G^{k,l,m,n})$ $\Diamond^k\Box^l A \supset \Box^m\Diamond^n A$	R is k, l, m, n-incestual: $\forall u,v,w,w'$: $uR^kv \wedge vR^mw \supset \exists w'(vR^lw' \wedge wR^nw')$
0.3: $\Box(\Box A \supset B) \vee \Box(\Box B \supset A)$	R is locally strongly connected: $\forall u,v$: $\exists w(wRu \wedge wRv) \supset (uRv \vee vRu)$
0.3*: $\Box(A \wedge \Box A \supset B) \vee \Box(B \wedge \Box B \supset A)$	R is locally connected: $\forall u,v$: $\exists w(wRu \wedge wRv) \wedge u \neq v \supset (uRv \vee vRu)$
0.2: $\Diamond\Box A \supset \Box\Diamond A$	R is locally strongly convergent: $\forall u,v$: $\exists w(wRu \wedge wRv) \supset \exists w'(uRw' \wedge vRw')$
0.2*: $\Diamond(A \wedge \Box B) \supset \Box(A \vee \Diamond B)$	R is locally convergent: $\forall u,v$: $\exists w(wRu \wedge wRv) \wedge u \neq v \supset \exists w'(uRw' \wedge vRw')$
Dense: $\Box\Box A \supset \Box A$	R is dense: $\forall u,v$: $Ruv \supset \exists w(uRw \wedge wRv)$
Triv: $\Box A \equiv A$	Every world reaches only itself: $\forall u,v$: $uRv \supset u = v$
Ver: $\Box\bot$	Every world is a dead end: $\forall u,v$: $\neg uRv$
Alt_n: $\Box A_1 \vee \Box(A_1 \supset A_2) \vee \ldots \ldots \vee \Box(A_1 \wedge \ldots \wedge A_n \supset A_{n+1})$	Every world reaches at most n distinct worlds: $\forall u, v_1, \ldots, v_{n+1}$: $\wedge \{uRv_i: 1 \leq i \leq n\} \supset \vee \{v_i = v_j: 1 \leq i < j \leq n\}$

$G^{k,l,m,n}$ is a very general schema introduced by Lemmon and Scott (1966) (Hughes and Cresswell 1984: 42); note that D is **K**-equivalent with $G^{0,1,0,1} = \Box A \supset \Diamond A$, $T = G^{0,1,0,0}$, $B = G^{0,0,1,1}$, $4 = G^{0,1,2,0}$, $5 = G^{1,0,1,1}$, $0.2 = G^{1,1,1,1}$, $Dense = G^{0,2,1,0}$. **S4.2** = **S4** + 0.2. Sahlqvist (1975) has proved first-order definability and canonicity for a class of axiom schemata which is even more general than $G^{k,l,m,n}$ (cf. Chagrov and Zakharyaschev 1997: ch. 10.3). The schemata 0.3, 0.3* (introduced by Lemmon) and 0.2, 0.2* (introduced by Geach) are important for the modal logics of orderings; 0.3, 0.3* are equivalent for

reflexive frames, and 0.2, 0.2* for serial frames. **S4.3** = **S4** + 0.3 is the logic of linear orderings. Likewise, **K4.3** is the logic of strict linear orderings and **KD4.3** the logic of strict linear orderings without last element. By adding Dense one obtains the logics of corresponding dense orderings. Ver and Triv are famous because they are characterized by the two singleton frames $\langle\{w\}, \langle w,w\rangle\rangle$ and $\langle\{w\}, \varnothing\rangle$, respectively, and every consistent $\mathbf{L} \in \Pi\mathbf{0}$ is either contained in Triv or in Ver (Makinson's theorem). The logics $\mathbf{S5}(\mathbf{Alt_n})$ are the only consistent extensions of **S5** (Scroggs' theorem; $\mathbf{S5Alt_1} = \mathbf{KTriv}$). For more details and canonical axioms see, for example, Segerberg (1971); Hughes and Cresswell (1984); or Chagrov and Zakharyaschev (1997).

Let us turn to examples where completeness and/or correspondence fails. We call an axiom schema X (and the corresponding logic **KX**) *non-compact* iff it is weakly but not strongly frame-complete, and *frame-incomplete* iff it is not even weakly frame-complete. Examples of axiom schemata which are both non-compact and not elementary (not first-order definable) are Löb's axiom G (also called W) and McKinsey's axiom 0.1, along with their corresponding frame-condition:

(G)	$\Box(\Box A \supset A) \supset \Box A$	R is transitive and terminal, that is there are no infinite R-chains $w_1Rw_2 \ldots w_nRw_{n+1} \ldots$
(0.1)	$\Box\Diamond A \supset \Diamond\Box A$	For no $w \in W$ there exist disjoint nonempty W-subsets U, V, such that all $w \in \{w^*{:}wRw^*\}$ have R-successors in U and V

G-frames are irreflexive, since a reflexive w implies the infinite chain wRwRw . . . As a result, **KG** contains 4, but neither T nor D (Hughes and Cresswell 1984: 101). That the G-corresponding condition C_G at the right side (proof see van Benthem 1984: 195f) is genuinely second-order is seen as follows. Consider the infinite formula set $\Delta = \{C_G\} \cup \{x_iRx_{i+1}: i \in \omega\}$. Every finite subset of Δ is satisfiable in a G-frame. Since first-order logic is compact, it follows that if C_G were first-order, then Δ would be satisfiable in a G-frame. But it is not, since by asymmetry of R this would imply an infinite ascending R-chain. So C_G is not first-order (Chagrov and Zakharyaschev 1997: 166). That **KG** is not canonical can be proved by showing that the frame of $M_c(\mathbf{KG})$ contains reflexive worlds and, thus, is not a **KG**-frame: this follows from the fact that the so-called Solovay's logic **S** = **KG** + T is consistent, and hence, produces reflexive worlds in $M_c(\mathbf{KG})$ (Chagrov and Zakharyaschev 1997: 165). *Weak* completeness of **KG** is proved by a suitable filtration of $M_c(\mathbf{KG})$ (Hughes and Cresswell 1986: 47ff).

Correspondence for the second-order condition corresponding to McKinsey's axiom 0.1 is established in Fine (1975b). **K0.1** has the f.m.p. and thus is w. frame-complete (Fine 1975a), but it is not canonical (Goldblatt 1991). The logic **KG** ('G' for 'Gödel') has become famous because it allows a translation of Gödel's incompleteness proof for first-order arithmetics into modal logic. If one translates $\Box A$ into the arithmetical language with Gödel-numbering g and provability predicate Pr(x), by the translation function $t(\Box A) = Pr(g(t(A)))$, then **KG** contain all modal theorems which are valid in this arithmetical interpretation, and Gödel's incompleteness results have a direct translation into the modal language (for details see Smorynski 1984). Also McKinsey's axiom is remarkable, for two reasons. First, 0.1 becomes canonical if it is added to **K4** or **S4**:

K4.1 = **K4** + 0.1 and **S4.1** = **S4** + **0.1** are first-order definable and canonical (proved in Lemmon and Scott 1966: 75). **K4.1**-frames are defined by the transitivity of R and the condition that every world reaches a 'dead end' ($\forall u \exists v$: $uRv \wedge \forall w(vRw \supset v = w)$); and **S4.1**-frames are additionally reflexive (van Benthem 1984: 202). Second, **S4.1** is then a simple example of a canonical PML with a frame-incomplete quantificational counterpart (§3.1 below).

All **K45**-extensions and all **S4.3**-extensions are weakly frame-complete. Lemmon conjectured in 1966 that all normal PMLs are w. frame-complete. In 1974, Fine and Thomason gave first examples of *frame-incomplete* PMLs. A standard way of proving the *frame-incompleteness* of $\mathbf{L} \in \Pi$ is the following: prove (i) that $\mathbf{F}(\mathbf{L})$ validates a certain formula schema X, and (ii) that X is not derivable in **L**, by specifying a class **M** of *non-standard* models of **L** which falsifies X. (Note that **M** cannot be standard because then **M** would also verify X.) By *correctness* w.r.t. models, this implies that $\nvdash_{\mathbf{L}} X$, and hence, that **L** is frame-incomplete.

Model-classes for a given **L**, even if nonstandard, must preserve validity under the rule of substitution. Such model-classes have been introduced as so-called *general frames* by Thomason (1972) (for details see Chagrov and Zakharyaschev 1997: ch. 8). A *general frame* G is defined as a pair $G = \langle F, \mathrm{Prop} \rangle$ where F is an ordinary frame and $\mathrm{Prop} \subseteq \mathrm{Pow}(W)$ is a set of 'valuation-admissible' subsets of W which is closed under intersection, relative complement, and under the operation '$\Box$:$W \to W$' defined by $\Box X = \{w: \forall u(wRu \supset u \in X\}$. The class $\mathbf{M}(G)$ of G-models is the class of all models based on F with valuation function $V: \mathcal{P} \to \mathrm{Prop}$ taking values in Prop. By definition, $G \models A$ iff $\mathbf{M}(G) \models A$. This definition entails, by the closure conditions on Prop, that whenever a general frame G validates A, then G validates every substitution instance s(A) of A. Moreover, to every model $M = \langle W,R,V \rangle$ there corresponds a *minimal* general frame G_M defined as $\langle \langle W,R \rangle, \mathrm{Prop}_M \rangle$ with Prop_M = the set of W-subsets which are the value of some $\mathcal{L}$-formula under V (Chagrov and Zakharyaschev 1997: 237). It follows that every substitution-closed formula set and in particular every logic $\mathbf{L} \in \Pi$ which is valid in M must also be valid in G_M. As a result, model-completeness of a logic implies its completeness w.r.t. general frames. (For details on general frames and their connection to *modal algebras* see Chagrov and Zakharyaschev 1997.)

A simple example of a frame-incomplete PML is van Benthem's logic **KVB**, where $VB = \Diamond\Box\bot \vee \Box(\Box(\Box B \supset B) \supset B)$. It is easily checked that every frame for VB satisfies the first-order condition that every world is a dead end or reaches a dead end, and hence, validates the axiom $\Diamond\mathrm{Ver} = \Diamond\Box\bot \vee \Box\bot$. Van Benthem constructs a countably infinite general frame with allowable values based on finite and cofinite W-subsets which validates VB but falsifies $\Diamond$Ver (Hughes and Cresswell 1984: 57ff). Van Benthem's examples shows also that first-order definability does not imply frame-completeness of a logic (which was an earlier conjecture). That also the reverse implication relation does not hold was demonstrated by an example of a canonical logic which is not first-order definable, given by Fine (1975a), namely the logic $\mathbf{KF} = \mathbf{K} + F =_{df} \Diamond\Box A \supset \Diamond\Box(A \wedge B) \vee \Diamond\Box(A \wedge \neg B)$.

Both examples show that there is no simple relationship between completeness and correspondence. First of all, correspondence has *two sides*. Modal formulas may or may not have corresponding frame-conditions (correspondence I), and frame-conditions

may or may not have corresponding modal formulas (correspondence II) (van Benthem 1984: 192, 211). Concerning correspondence I, van Benthem (1984: 169ff) shows that every modal formula has a corresponding second-order frame condition; so the only interesting question is whether a modal formula is elementary. Concerning correspondence II, various examples of modally undefinable first-order conditions have been given in §2.3. General theorems about correspondence I and II are found in van Benthem (1983, 1984). The following connections between frame-completeness and correspondence I have been proved in Fine (1975a): (1) If $\mathbf{L} \in \Pi$ is first-order definable and w. frame-complete, then $\mathbf{L}$ is canonical, and (2) If $\mathbf{L} \in \Pi$ is natural, then $\mathbf{L}$ is first-order definable. Naturality, as defined by Fine, is a stronger property than canonicity (a generalization of Fine's theorem for general frames in terms of 'D-persistence' is given by van Benthem (1983); see Chagrov and Zakharyaschev 1997: 341–4).

For many purposes, one needs PMLs with several different modal operators, for example an alethic-deontic PML for the is–ought problem (Schurz 1997). A *multimodal language* $\mathcal{L}_I$ contains a set $\{\Box_i\colon i \in I\}$ of modal operators (I an index set). The simplest kind of a normal PML in $\mathcal{L}_I$ is a so-called *combination* (or join) of normal monomodal $\Box_i$-logics $\{\mathbf{L}_i\colon i \in I\}$, denoted by $\otimes\{\mathbf{L}_i\colon i \in I\}$, and defined as the smallest normal PML in $\mathcal{L}_I$ containing every $\mathbf{L}_i$. Frames for these logics have the form $\langle W, \{R_i\colon i \in I\}\rangle$. Syntactically, $\otimes\{\mathbf{L}_i\colon i \in I\}$ is obtained from the $\mathbf{L}_i$ ($i \in I$) by joining their representative axiom sets $\mathbf{X}_i$ and under substitution in the combined language $\mathcal{L}_I$. Hence, $\otimes\{\mathbf{L}_i\colon i \in I\}$ is representable as $\mathbf{K}_I\mathbf{X}_I$ with $\mathbf{X}_I = \cup_i\mathbf{X}_i$. Instead of proving metalogical properties like completeness, etc. for all possible multimodal combinations separately, it is more desirable to prove general *transfer theorems* in the following sense: whenever all $\mathbf{L}_i$ have a certain property, then $\otimes\{\mathbf{L}_i\colon i \in I\}$ has it, too. The following general transfer theorem holds for combined multimodal logics: (1) Weak and strong frame-completeness, canonicity and f.m.p. transfer from the $\mathbf{L}_i$ ($i \in I$) to $\otimes\{\mathbf{L}_i\colon i \in I\}$; and (2) decidability, interpolation, and Halldén-completeness transfer under presupposition of weak completeness of the $\mathbf{L}_i$ ($i \in I$). The theorem was independently proven by Kracht and Wolter (1991) and Fine and Schurz (1996) (the latter paper was written up in 1990 but its publication was delayed). If a multimodal logic contains in addition *interactive* axioms which relate distinct modalities (e.g. $\Box_1\Box_2 A \supset \Box_2\Box_1 A$), then transfer theorems are possible only in special cases (Fine and Schurz 1996: 210ff, for such examples). The investigation of combined logics and transfer has become a topic of increasing interest in modal logics; for a survey see Kracht and Wolter (1997).

3 Modal Quantificational Logics (QMLs)

Fixed domain and rigid designators: Q1MLs

$\mathcal{L}Q1$ is the object language modal quantificational logics of 'type 1,' in short: Q1MLs. It contains a set $\mathcal{V}$ of individual variables (x, y, . . .), in short: variables, a set C of individual constants (a, b, . . .), in short: constants, and for each $n \geq 0$, a set $\mathcal{R}^n$ of n-ary relation symbols (F, G, Q, . . .). All these sets are denumerably infinite (0-ary relation symbols are propositional variables). For reasons of simplicity we omit function symbols; thus singular terms, denoted by t, t_1, t_2, . . . , are constants or variables; J(the

set of all terms) $=_{df} \mathcal{V} \cup C$. The new primitive logical symbols are the universal quantifier $\forall$ ($\forall$x = 'for all x:'), and the identity symbol, =. The existential quantifier $\exists$ ($\exists$x = 'there exists an x:') is defined as usual by $\exists xA =_{df} \neg\forall x\neg A$. *ℒQ1* is again identified with the set of its (well-formed) formulas, which are recursively defined by the following clauses: (1) $t_1, \ldots, t_n \in J$, $Q \in \mathcal{R}^n \Rightarrow Qt_1 \ldots t_n \in$ *ℒQ1*; (2) A, B $\in$ *ℒQ1* $\Rightarrow \neg A$, A $\vee$ B, $\Box$A $\in$ *ℒQ1*; and (3) x $\in \mathcal{V}$, A $\in$ *ℒQ* $\Rightarrow \forall xA \in$ *ℒQ1*.

We assume acquaintance with the notions of bound and free *occurrences* of variables. Variable x is called free in A iff A contains at least one free x-occurrence. $\mathcal{V}_f(A)$ = the set of free variables in A; likewise for $\mathcal{V}(A)$, $C(A)$, $\mathcal{R}^n(A)$. A* is an *alphabetic variant* of formula A iff A* results from A by replacing every bound occurrence of some variables $x_1, \ldots x_n$ in A by variables $y_1, \ldots y_n$, respectively, provided (for each $1 \leq i \leq n$) that no x_i-occurrence in A lies in the scope of an $\forall y_i$-quantifier and no free y_i-occurrence in A lies in the scope of an $\forall x_i$-quantifier. A[t/x] denotes the result of the correct substitution of term t for variable x in A and is defined as the result of the replacement of every free occurrence of x in A* by t; where A* is the first alphabetic variant of A (according to a given formula enumeration) in which x does not occur in the scope of a quantifier binding t. $A[t_{1-n}/x_{1-n}]$ denotes the result of the correct (simultaneous) substitution of t_i for x_i in A (for all pairwise distinct x_i).

The notion of a *frame* remains the same for all kinds of QML-semantics. The simplest way of extending Kripke models to modal quantificational languages are *Q1-models*. They contain one fixed domain D of objects, which is the same for all worlds in W, and assume that singular terms are rigid – so only the interpretation of the relation symbols is world-relative. More precisely, a Q1-model is a quadruple $M = \langle W,R,D,V\rangle$ where $\langle W,R\rangle$ is a frame, $D \neq \emptyset$ is a nonempty domain of individuals, and the valuation function V is defined as follows: (1) $V: J \rightarrow D$, hence $\forall t \in J$. $V(t) \in D$, and (2) for all $n \geq 0$, $V: W \times \mathcal{R}^n \rightarrow D^n$, hence $V(w,R) =_{df} V_w(R) \subseteq D^n$. The value $V_w(R)$ is also called the 'extension' of R at world w, and the partial function $V(R): W \rightarrow D^n$ is called R's *intension* (this view of intension' goes back to Carnap). The restriction of V to constants and relation symbols is often called an *interpretation* of *ℒQ1*, and the restriction of v to variables an *assignment* for variables. Because we treat free variables and constants semantically on a par, we don't need to distinguish between closed and open formulas. This setting is close to Machover (1996: 151f). Of course, variations are possible. For example, one may drop constants and let free variables play their role, as in Hughes and Cresswell (1984); or one may give free variables the closure-interpretation, as in Fine (1978).

$M[x:d]$ denotes a model which is like M except that v^M assigns $d \in D$ to x; and similar for $M[x_{1-n}/d_{1-n}]$. The definition of '$(M,w) \models A$' is as follows: for atomic formulas, $(M,w) \models Rt_1 \ldots t_n$ iff $\langle V(t_1), \ldots, V(t_n)\rangle \in v_w(R)$ and $(M,w) \models t_1 = t_2$ iff $V(t_1) = V(t_2)$; for A = $\neg$B, B $\vee$ C, $\Box$B as in the propositional case; and for quantified formulas: $(M,w) \models \forall xA$ iff $\forall d \in D$, $(M[x:d],w) \models A$. The other semantical notions are as in the propositional case. The *coincidence lemma* tells us that $(M,w) \models A[t/x]$ iff $(M[x:v^M(t)],w) \models A$ (*Proof: exercise* (Hughes and Cresswell 1984: 168).

DEFINITION OF NORMAL Q1MLS Given a PML **KX**, its *Q1-counterpart* is denoted as **Q1KX** and is defined as the smallest set of *ℒQ1*-formulas which contains all *ℒQ1*-instances of the axiom schemata of **KX** plus the following axiom schemata for quantification (UI, $\forall$, UG, BF) and identity (I, rISub, rI$\neg$) (for all x $\in \mathcal{V}$, t $\in J$):

UI: $\forall xA \supset A[t/x]$ ('universal instantiation')
BF: $\forall x\Box A \supset \Box\forall xA$
$\forall 1$: $\forall x(A \supset B) \supset (\forall xA \supset \forall xB)$
$\forall 2$: $A \supset \forall xA$, provided x is not free in A.
I: $t = t$
rISub: $t_1 = t_2 \supset (A[t_1/x] \supset A[t_2/x])$ ('rigid identity-substitution')
rI¬: $\neg t_1 = t_2 \rightarrow \Box\neg t_1 = t_2$ ('rigid identity w.r.t. ¬')

and which is closed under the rules of **KX** (TautR, MP, N) and under the rule:

UG: $A/\forall xA$ ('rule of universal generalization').

Provability $\vdash_L A$ and deducibility $\Gamma \vdash_L A$ is defined as for PMLs.

RECAPITULATION OF NONMODAL QL Prove the dual axiom of 'existential instantiation' EI: $A[t/x] \supset \exists xA$, and the dual rule of 'existential generalization' EG: $A \supset B/\exists xA \supset B$, provided $x \notin \mathcal{V}_f(B)$. Prove the equivalence of UG with UGt: $A[x/t]/\forall xA$, provided $t \notin J(A)$. Prove that our axiomatization UI + $\forall 1$ + $\forall 2$ + UG (also used by Fine 1978) is equivalent with UI + UG*: '$A \supset B/A \supset \forall xB$, provided $x \notin \mathcal{V}_f(A)$' (used, e.g. in Hughes and Cresswell 1984: 166).

SYNTACTICAL THEOREMS ABOUT Q1MLS (1) The rule UG is neither valid nor model-admissible, but merely frame-admissible. (2) BF is valid in every Q1-model. (3) The converse Barcan formula, cBF: $\Box\forall xA \supset \forall x\Box A$, is a **Q1K**-theorem. (4) The rigidity principle rI: $t_1 = t_2 \supset \Box(t_1 = t_2)$ is a **Q1K**-theorem, and the rigidity axiom rI¬ is a **Q1B**-theorem. (5) The formula $\Box\exists xA \supset \exists x\Box A$ is invalid.

PROOF *Exercise* (see examples below; for 3 see Hughes and Cresswell 1968: 143; for 4 see Schurz 1997: fn.s 108, 109). The counterintuitivity of 5 was illustrated by Quine (1953, p. 148) as follows: it necessary that one player will win, but for no one of the players is it necessary that just he will win (cf. Hughes and Cresswell 1968, p. 197).

DISPROOF OF MODEL-ADMISSIBILITY OF UG By the following countermodel $M = \langle \{w\}, \emptyset, \{d_1, d_2\}, V\rangle$ with $V_w(F) = \{d_1\}$. It yields $M \models Fa$ but $M \not\models \forall xFx$.

PROOF OF FRAME-VALIDITY OF UG BY CONTRAPOSITION Assume $\langle W,R\rangle \not\models \forall xFx$. So there exist D, V, w such that for $M = \langle W,R,D,V\rangle$ and $w \in W^M$, $(M,w) \not\models \forall xFx$. Hence $\exists d \in D^M$: $(M[x:d],w) \not\models Fx$. Since the model M[x:d] is based on $\langle W,R\rangle$, this implies that $\langle W,R\rangle \not\models Fx$. Q.E.D.

A general definition of normal Q1MLs requires a suitable formulation of the *rule of substitution for predicates*. This rule was first described by Kleene (1971: 155–62) and is explained as follows. A substitution instance of formula A w.r.t. an n-ary predicate Q in 'name form variables' $z_1 \ldots z_n$ is a formula A* which results from the simultaneous replacement of every occurrence of a term-instance $Qt_1 \ldots t_n$ in A** by a corresponding term-instance $B[t_{1-n}/z_{1-n}]$, for a given formula ('complex predicate') B; where A** is the first alphabetic variant of A in which no free variable of B other than

$z_1, \ldots, z_n$ gets bound (for details see Schurz 1995: 45–52). Important in our context is the following *QML-substitution-theorem:* frame-validity of Q1ML-formulas is preserved under substitution for predicates (proof see Schurz 1997: 46–8). This theorem guarantees that for every frame class **F**, **L**(**F**) will be closed under substitution and, hence, will be a *normal Q1-logic*. Moreover, our notion of substitution allows us to define *a normal* **Q1**-logic as any formula set $\mathbf{L} \subseteq \mathcal{LQ}1$ which contains **Q1K** and is closed under the rules of **Q1K** and under substitution for predicates. **Q1**Π denotes the lattice of normal Q1-logics.

As in the propositional case, every $\mathbf{L} \in$ **Q1**Π is *representable* (but not necessarily axiomatizable) as **Q1KX**. $\mathbf{L} \in$ **Q1**Π is called *propositionally representable* iff $\mathbf{L} =$ **Q1KX** for some **X** consisting solely of *propositional* axiom schemata – in other words, iff **L** is the Q1-counterpart of the PML **KX**. Propositionally representable Q1-logics are the standard case. However, **X** may also contain additional quantificational (or identity) axiom schemata – on two reasons: First, some frame-complete PMLs have frame incomplete Q1-counterparts, which need additional $\mathcal{LQ}1$-axioms to become frame-complete (cf. **Q1S4.1** below). Second, there exist interesting cases of additional schemata which are only characterizable by nonstandard model-classes, such as Fine's anti-Haecceitistic axiom **H**.

Correspondence, correctness, and w./s. completeness w.r.t. models or frames (and related notions) are defined as in the propositional case. Of course, Q1-logics do neither have the f.m.p. w.r.t the domain, nor are they decidable, because nonmodal first-order logic lacks these properties. Correctness of Q1-logics is proved, as usual, by showing that all **Q1L**-axioms are valid on all frames, and that all **Q1K**-rules preserve frame-validity; this was done above. The correctness-proof also establishes that every propositionally representable Q1-logic corresponds to the same class of frames as its propositional counterpart. This gives us a following *frame transfer theorem* from PMLs to their Q1-counterparts: For every PML **KX**: **F**(**KX**) = **F**(**Q1KX**). Hence, if a frame-condition C_X corresponds to **KX**, then it corresponds also to **Q1KX**.

As in non-modal QL, the domain of the *canonical model* of a Q1ML is constructed from the =-equivalence classes of terms. On this reason, the canonical worlds need not only be maximally **L**-consistent formula sets, they also have to be 'ω-complete'. The canonical model $\mathbf{M}_c(\mathbf{L},\Delta)$ of a Q1-logic **L** is explicitly relativized to a saturated formula set Δ which extends the given **L**-consistent formula set Γ and *fixes the rigid term identities*. Implicitly, the notion of ω-completeness and the canonical model is also relativized to the *term set* $J(\mathcal{LQ}1)$ of the given denumerably infinite language $\mathcal{LQ}1$.

DEFINITION OF CANONICAL Q1-MODELS (1) A formula set $\Gamma \subseteq \mathcal{LQ}1$ is *ω-complete* (w.r.t. $\mathcal{LQ}1$) iff for every $A \in \mathcal{LQ}1$: $\Gamma \vdash_{\mathbf{L}} \forall x A$ iff $\Gamma \vdash_{\mathbf{L}} A[t/x]$ for every $t \in J(\mathcal{LQ}1)$; Γ is called *L-saturated* iff it is both maximally **L**-consistent and ω–complete. (2) The canonical model $\mathbf{M}_c(\mathbf{L},\Delta) = \langle W_c, R_c, D_c, V_c\rangle$ of $\mathbf{L} \in$ **Q1**Π for the **L**-saturated formula set Δ in given language $\mathcal{LQ}1$ is defined as follows: (2.1) W is the set of all **L**-saturated $\mathcal{LQ}1$-formula sets w which preserve the Δ-identities; that is for all t_1, t_2: $t_1 = t_2 \in w$ iff $t_1 = t_2 \in \Delta$ (this ensures constant domain and rigid designators); (2.2) R_c is as in the propositional case; (2.3) for all $t \in J$, $V_c(t) = \{t^*: t = t^* \in \Delta\}$, and for all $Q \in \mathcal{R}^n$ and $w \in W_c$, $V_w(Q) = \{\langle V_c(t_1), \ldots, V_c(t_n)\rangle: Qt_1 \ldots t_n \in w\}$; (2.4) $D = \{V_c(t): t \in J\}$.

The proof of strong model-completeness proceeds in the following three steps (this technique was suggested by Thomason (1970)):

STEP 1: LINDENBAUM–HENKIN-SATURATION-LEMMA Every **L**-consistent formula set Γ in language $\mathcal{LQ}1$ can be extended to an **L**-saturated formula set Δ in a language $\mathcal{LQ}1^*$ which differs from $\mathcal{LQ}1$ only in that it contains an additional denumerably infinite set C^* of *new* constants (i.e. $C^* \cap C(\mathcal{LQ}1) = \emptyset$, $C(\mathcal{LQ}1^*) = C \cup C^*$). Given an enumeration of all formulas $A_0, A_1 \ldots$ in $\mathcal{LQ}1^*$ and of all constants in C^*, one defines:

$$\Delta_0 =_{df} \Gamma,$$

$$\Delta_{n+1} =_{df} \begin{cases} \Delta_n \cup \{A_n, \neg B[a/x]\} \text{ (where a is the first constant in } C^* - C(\Delta_n, A_n)), \\ \qquad \text{if } \Gamma_n \cup \{A_n\} \text{ is consistent and } A_n \text{ is of the form } \neg\forall xB \\ \Gamma_n \cup \{A_n\}, \text{ if } \Gamma_n \cup \{A_n\} \text{ consistent and } A_n \text{ is not of the form } \neg\forall xB \\ \Gamma_n \qquad \text{if } \Gamma_n \cup \{A_n\} \text{ is inconsistent} \end{cases}$$

$$\Delta =_{df} \bigcup \{\Delta_n : n \in \omega\}$$

For each n, there are infinitely many new constants remaining in $C^* - C(\Delta_n, A_n)$; thus the required new constant always exists. As in the non-modal case it is proved that Δ is **L**-saturated (Garson 1984: 271).

STEP 2 New in the quantificational case is the proof of the canonical model lemma. This lemma now assures the *existence* of a formula set which is not only maximally **L**-consistent, but also ω-complete *w.r.t. the same* language $\mathcal{LQ}1^*$ of Δ. For Q1-logics, this is proved by exploiting the Barcan formula.

CANONICAL Q1-MODEL LEMMA (1) If $\Gamma, \Sigma \subseteq \mathcal{LQ}1$, Γ is ω-complete, and ς is finite, then $\Gamma \cup \Sigma$ is ω-complete. (2) Every **L**-consistent and ω-complete formula set Γ can be extended to an **L**-saturated set Δ written in the *same* language (i.e. the language w.r.t which Γ was ω-complete). (3) If $\neg\Box B \in w \in W_0(\mathbf{L},\Delta)$, and w is **L**-consistent, then $\{A: \Box A \in w\} \cup \{\neg B\}$ is (3.1) **L**-consistent, (3.2) ω-complete, (thus) (3.3) has an **L**-saturated extension u written in the same language, such that (3.4) for all t_1, t_2: $t_1 = t_2 \in u$ iff $t_1 = t_2 \in \Delta$. (4) $\forall u \in W_0(\mathbf{L},\Delta)$: $\Box A \in u$ iff $\forall v \in W_0(uR_0v \Rightarrow A \in v)$.

PROOF *Exercise. Hints:* For 1 see Garson (1984: 274) (his lemma 1). For 2 see Garson (1984: 274f) (his lemma 2). The proof constructs Δ as above except that it shows that for each A_n of the form $\neg\forall xB$, the required constant a exists in the old language, because Δ_n is already ω-complete by 1 of our lemma. The proof of our lemma 1 + 2 rests solely on classical quantifier principles. 3.1 is proved as in the propositional case. For 3.2, see Garson (1984: 275) (his lemma 3) – this proof depends on the Barcan formula. 3.3 follows from 3.1 + 2 by 2. For 3.4, see Garson (1984: 277f) – this proof uses the rigidity axiom (rI–) and the theorem (rI). 4: this follows from 3 as in the propositional case.

STEP 3 It is now straightforward to prove the *Q1ML-Truth Lemma:* For every $A \in \mathcal{LQ}1^*$ and $w \in W_0$: $(M_0(\mathbf{L},\Delta), w) \models A$ iff $A \in w$. *Proof* by induction on formula complexity (Garson 1984: 275f; Hughes and Cresswell 1984: 84, 176). The atomic case holds by

definition, the steps for propositional operators are as before; the only new item are the the following steps for the identity of formulas and quantifiers: Step 3.1: $t(M_c,w) \models t_1 = t_2$ iff $v_c(t_1) = v_c(t_2)$ iff $t_1 = t_2 \in \Delta$ (by definition of $v_c(t)$) iff $t_1 = t_2 \in w$ (by def. of W_c). Step 3.2: $(M_c,w) \models \forall xA$ iff $\forall d \in D_c$: $(M_c[x:d],w) \models A$, iff $\forall t \in J(\mathcal{LQ}1^*)$: $(M_c[x:v_c(t)],w) \models A$ (by def. of D_c), iff $\forall t \in J(\mathcal{LQ}1^*)$: $(M_c,w) \models A[t/x]$ (by coincidence lemma), iff $\forall t \in J(\mathcal{LQ}1^*)$: $A[t/x] \in w$ (by induction hypothesis), iff $\forall xA \in w$ (by ω-completeness of w). Q.E.D.

Lindenbaum–Henkin and Truth Lemma establish as in the propositional case that:

Q1ML-MODEL-COMPLETENESS (1) Every normal Q1ML is strongly model-complete, and is strongly characterized by the class of its models. (2) **Q1K** is canonical.

As in the propositional case, to prove that a Q1-logic **L** stronger than **Q1K** is canonical requires to show that the frame of **L**'s canonical model is an **L**-frame. It is natural to *conjecture* that canonicity transfers from all *propositionally representable* Q1-logics to their Q1-counterpart. This conjecture was stated as an open problem in Hughes and Cresswell (1984: 183f) and was (wrongly) positively answered by Garson (1984: 276), But quite astonishingly, general canonicity transfer *fails*. An example of a canonical **L** $\in \Pi$ with a frame-incomplete Q1-counterpart is **S4.1**:

Q1S4.1-THEOREM (1) **S4.1** is canonical. (2) **Q1S4.1** is frame-incomplete. (3) **Q1S4.1** + $(\Diamond\Box\exists xA \supset \Diamond\exists x\Box A)$ is canonical.

PROOF For 1 see earlier. 2 is proved by showing that $\Diamond\Box\exists xA \supset \Diamond\exists x\Box A$ is valid on all **S4.1**-frames, but invalid in a certain nonstandard **Q1S4.1**-model; see Schurz (1997: 292f), the proof is due to Kit Fine. A proof of 3 is found in Schurz (1997: 293–5).

The reason why the proof of canonicity works for **S4.1** but *not* for **Q1S4.1** is that the first-order frame condition corresponding to **S4.1** contains an *existential* quantifier. This means in the propositional case that it has to be shown that a certain formula set has a maximally consistent extension, while in the predicate logical case it has to be shown that this formula set has a maximally consistent *and* ω-*complete* extension; but this is only possible if the additional axiom schemata $(\Diamond\Box\exists xA \rightarrow \Diamond\exists x\Box A)$ is available. However, the following restricted transfer theorem holds:

RESTRICTED Q1-CANONICITY-TRANSFER THEOREM (1) If a normal PML **L** = **KX** has the *subframe property* (which means that **L**'s frames are closed under subframes), then canonicity transfers from **KX** to **Q1KX**. (2) If **L**'s frames are definable by a *purely universal* first-order formula, then **L** has the subframe property.

The proof of 1 is based on the fact that the frame of M_c(**Q1KX**) is isomorphic with a subframe of M_c(**KX**) (see Schurz 1997: 295: for a similar result for intermediate logics see Shimura 1993: 36). The proof of 2 is straightforward. The theorem covers the axiom schemata D, T, 5, Alt_n, Ver, Triv, 0.3, because they correspond to universal first-order formulas; moreover it covers all subframe logics in the sense of Fine (1985: 624;

see Chagrov and Zakharyaschev 1997: 380ff) which include, among others, **KG** and **KGrz**. It is an open problem whether canonicity-transfer holds for larger classes of normal Q1MLs. In lack of stronger transfer results, canonicity has to be proved for each QML separately (see Gabbay (1976) and Bowen (1979) for various special canonicity results).

The transfer problem from monomodal to multimodal logics exists also in the quantificational case, but the propositional proof technique (§2.5) does not generalize to the quantified case. So far, only *canonicity transfer* from monomodal Q1-logics to their multimodal combination has been proved in Schurz (1997: 67).

Varying domains, rigid designators and free quantification: Q2-logics

The constant domain assumption implies that whatever exists in the actual world, exists necessarily, that is in all possible worlds. Moreover, for every $t \in J$, 't exists' ($\exists x(x = t)$) is a theorem of nonmodal QL. Hence, $\Box\exists x(x = t)$ ('t necessarily exists') is a **Q1K**-theorem for every $t \in J$. This idealization is inadequate when worlds are interpreted as possible states of the real world, because individuals do not have 'eternal' life. So there is a need to develop semantics with varying domains.

In models with world-relative domains, every world w has its own domain D_w of individuals – those objects which exist in world w. D_w is the range of the quantifier at w: $\forall xFx$ is true at w iff every $d \in D_w$ has property F. The Barcan formula $\forall x\Box A \supset \Box\forall xA$ is now invalid: it might well be that all individuals in D_w have the property F at all worlds v accessible from w, but some world v accessible from w has an individual in its domain D_v which is not in D_w and does not have property F at v (i.e. *not* $\forall v(Rwv \Rightarrow \forall d \in D_v((M[x:d],v) \models Fx)$). But also, the classical quantifier principles become problematic. Recall that the converse Barcan formula cBF $\Box\forall xA \supset \forall x\Box A$ is a theorem of every normal modal logic with classical quantifier principles. But in models with world-relative domains, cBF can only be valid if the condition of *nested domains* is satisfied: $uRv \Rightarrow D_u \subseteq D_v$. In order to keep classical quantifier principles, Hughes and Cresswell (1968: 171ff), Gabbay (1976: 44ff) and Bowen (1979: 8ff) adopt this condition.

The nested domain condition is rather restrictive. For symmetric R it even implies a *constant* domain for every generated model (recall syntactic Q1ML-theorem no. 5 in the previous subsection); so the difference to Q1-logics would vanish for all **QKB**-extensions with nested domains. But even *if* this condition is accepted, the classical quantifier principles are problematic, at least if designators are rigid. Assume $a \notin D_w$; for example, a = Pegasus and w = the real world. What truth value should be given to the sentence Fa, for example 'Pegasus has wings,' at world w? Since designators are rigid, the so-called requirement of *local* predicates, which says that only objects which exist at world w may be elements of predicate extensions at w, cannot avoid a conflict with classical quantifier principles. For, if $\forall xFx$ is true at worlds w, but $V(a) \notin D_w$ and $V(a) \notin V_w(F)$, then $\neg Fa$ and hence (by classical quantifier principles) $\exists x\neg Fx$ is true at w, contradicting the truth of $\forall xFx$ at w. A way out is to give sentences about nonexistents at w *no* truth-value at w. This leads to a semantics with *truth value gaps*, which has been developed by Hughes and Cresswell (1968: 170–3) and Gabbay (1976: 44ff).

If truth-value gaps should be avoided, we must allow that individuals may have properties at a world without being existent at world w. For example, we must allow that

'Pegasus has wings' is regarded as true at our world although Pegasus does not exist in our world. Classical quantifier principles can then no longer be valid, for $Fa \rightarrow \exists xFx$ comes out false at our world w. Hence, we must adopt the principles of *free logic*. I agree with Garson (1984: 261) that free QMLs are the most adequate choice for models with varying domains. In free logic, the classical UI-axiom is replaced by its *free* logic variant fUI: $\forall xA \supset (Et \supset A[t/x])$; in words: 'if all objects have property A, and t exists, then t has property A.' 'Et' is the *existence predicate*, defined as '$\exists x(x = t)$.' A like change is made for the rule UG. A first system of this kind has been suggested by Kripke (1963b); but Kripke only mentions this possibility (1936b: 70) and prefers to an axiomatization which avoids formulas with constants or free variables. Fine (1978) has given an elaboration of this kind of free modal QL for **S5**, and several further systems are discussed in Garson (1984: 257, 285). We call these logics **Q2**-logics and define their basic concepts as follows.

DEFINITIONS The Q2-language $\mathcal{LQ}2$ is syntactically like $\mathcal{LQ}1$, but it is *interpreted* in different way. The *existence predicate* E in Q2-languages is *defined* by $Et =_{df} \exists x(x = t)$. An *Q2-model* (based on frame $\langle W,R\rangle$) is a quintuple $M = \langle W,R,U,Df,V\rangle$, where $U \neq \emptyset$ is the *total domain* of possible individuals and Df: $W \rightarrow Pow(U)$ is the *domain function* assigning to each world $w \in W$ its domain $D_w \subseteq U$. D_w is called the *inner* domain of w (the existing objects of w) and $U–D_w$ the *outer* domain of w (the nonexisting objects of w). (One could add the requirement $U = \cup_{w \in W} D_w$; but this would not bring new theorems; see Schurz (1997: 198).) The valuation function for terms and predicates and the truth clauses for atomic formulas, identity formulas, and propositionally compound formulas are as in Q1-logics. The only new clauses concern quantification: $(M, w) \models \forall xA$ iff for every $d \in D_w$, $(M[x:d],w) \models A$. This yields for the existence predicate: $(M,w) \models Et$ iff $v(t) \in D_w$.

The minimal normal Q2-logic, **Q2K**, is axiomatically defined like **K1** *except* that (1) the axiom schema BF is *dropped*, (2) the axiom UI is replaced by its *free* version fUI: $\forall xA \rightarrow (Ey \rightarrow A[t/x])$, and (3) the rule UG is replaced by its free version fUG: $Ex \rightarrow A/\forall xA$ (Garson 1984: 252; Fine 1978: 131f, suggests an equivalent axiomatization which keeps UR and adds '$\forall xEx$'). *Exercise:* Prove the duals fEI: $(Et \wedge A[t/x]) \supset \exists xA$, and fEG: $A \wedge Ex \supset B/\exists xA \supset B$, provided $x \notin \mathcal{V}_f(B)$. **L** is a normal Q2-logics, ($\mathbf{L} \in \mathbf{Q2\Pi}$) iff **L** extends **Q2K** and is closed under the rules of **Q2K** and under substitution for nonlogical predicates. As before, every $\mathbf{L} \in \mathbf{Q2\Pi}$ is representable as **Q2KX**. The strategy of proving *model-completeness* which was used for Q1-logics *fails* for Q2-logics, because the Barcan formula is missing which allowed us to construct saturated sets in the same language. Fine (1978: 131–5) gives a proof of canonicity for **Q2S5** based on so-called *nice diagrams* (these are saturated sets of formula-world pairs). As far as I can see, this technique generalizes to all Q2-logics containing **Q2B**, but not to all Q2-logics. A general proof of model-completeness via a canonical model construction is possible by replacing the rule (fUG) by the following *stronger* rule. A *G-function* is a function G: $\mathcal{LQ}2 \rightarrow \mathcal{LQ}2$ which assigns to each $A \in \mathcal{LQ}2$ a formula of the form $G(A) := B_0 \rightarrow \Box(B_1 \rightarrow \Box(B_2 \rightarrow \ldots \Box(B_n \rightarrow A)\ldots)$, for given $B_0, B_1, \ldots, B_n$ (n, 0) where B_0 may be missing. The stronger rule is:

GUG: $G(Ex \rightarrow Ax)/G(\forall xA)$, provided $x \notin \mathcal{V}_f(G(\forall xA)$

With minor simplifications I am following Garson (1984: 282ff; Garson also replaces fUI by GUI, but this replacement is redundant; see Schurz 1997: 199f). GUG preserves frame-validity and, thus, is correct w.r.t. the class of Q2-models. GUG also covers also rule (fUG*) (recall §3.1), and thus, it implies the axioms $\forall 1 + 2$. A Q2-logic where UG + $\forall 1 + 2$ are replaced by GUG is called a *QG2-logic*. Model-completeness of QG2-logics can be proved similar as for Q1-logics. The worlds of the canonical model $M_c(\mathbf{L},\Delta)$ (in given *ℒQ2**) are now all *G-saturated* formula sets; these are all maximally **L**-consistent formula sets Γ which are *G-complete*: $\Gamma \vdash_{\mathbf{L}} G(\forall xA)$ iff $\Gamma \vdash_{\mathbf{L}} G(Et \supset A[t/x])$ for every $t \in \mathcal{J}^*$. The proof proceeds through the same steps as before; due to the stronger G-rule it can be proved, without BF, that for every $w \in W_c$ with $\neg\Box B \in w$, $\{A: \Box A \subseteq w\} \cup \{\neg B\}$ can be extended to a G-saturated set in the same language. For terms and predicates, R_c and V_c are defined as before; $U_c = \{V_c(t): t \in \mathcal{J}^*\}$, $Df_c: W \to Pow(U_c)$ such that $D_c(w) = \{V_c(t): Et \in \Delta\}$. We thus obtain the *QG2ML-model-completeness-theorem:* Every QG2-logic **L** is correct and strongly complete w.r.t. the class of its Q2-models, and **Q2GK** is canonical.

Garson (1984: 284f) claims it as an open problem if and to which extent the rule GUG is indeed stronger than UG. Schurz (1997: 200) gives a partial answer, by proving the *GUG-Theorem:* In all Q2-logics which contain B, GUG is admissible. Hence, all normal extensions of **Q2B** are strongly model-complete. It is an open problem whether there exist model-incomplete Q2MLs that don't extend **Q2B**.

Concerning frame-completeness, the same restricted tranfer result as for Q1-logics can be proved for propositionally representable Q(G)2-logics. Schurz (1997: 201f) defines a translation function t: *ℒQ2* → *ℒQ1* which translates Q2-formulas into semantically equivalent Q1-formulas, and Q2-models into corresponding Q1-models. An inverse translation is impossible: the *ℒQ1*-quantifier figures like a *possibilistic* quantifier for translated *ℒQ2*-logics; thus *ℒQ1* has greater expressive power than *ℒQ2*. With the help of this translation function, various transfer theorems from Q1- to Q2-logics are established; in particular the following *frame-transfer:* for every $\mathbf{L} \in \mathbf{Q2\Pi}$: $\mathbf{F(L)} = \mathbf{F}(t(\mathbf{L}))$, where $t(\mathbf{L})$ is the Q1-translation of Q2-logic **L**. If **X** is propositional, then $t(\mathbf{X}) = \mathbf{X}$; hence propositionally representable Q2-logics have the same frame-classes as their Q1-counterparts. Whether transfer of frame-completeness from Q1MLs to Q2MLs is possible remains an open problem (Schurz 1997: 204).

Nonrigid designators, counterpart theory, and worldline semantics: Q3-logics

Rigid designators presuppose that the fixation of their reference does not depend on any contingent property of the individual to which they refer. This may be true for mathematical objects such as '7' or '9' (Kripke 1972 uses them often as illustrations), but is it possibly true for empirical objects? According to Putnam's famous account of meaning (1975), the fixation of rigid reference is based (1) on an indexical relation of direct acquaintance with the individual in the present (*hic et nunc*) state (the act of 'babtizing'), and (2) on a unique relation of causal successorship or predecessorship. Accordingly, there are two problems with that account. Concerning (2), nothing guarantees that the relation of predecessor- or successorship in past and future states is uniquely determined. Take Frege's old example of the morning and the evening star,

both of which are identical with the planet Venus: assume that in some future time, Venus splits into two planets, one appearing only at the morning and the other in the evening, then, to what objects will the names 'morning star' and 'evening star' refer in that distant future state? (A more realistic example is the process of cell division.) And concerning (1), the relation of 'acquaintance' in the act of 'baptizing' is never absolutely 'direct' but always mediated through contingent properties.

Hintikka (1961) and Kanger (1957b) have already made suggestions for QMLs with non-rigid designators, in short: *nonrigid* QMLs (also see Hughes and Cresswell 1968: 195). Syntactically, the axiom rI¬ has to be dropped for nonrigid QMLs, and the rigid principle of substitution of identicals rISub must be restricted to *nonmodal* formulas as follows:

ISub: $t_1 = t_2 \supset (A[t_1/x] \supset A[t_2/x])$, provided A does not contain '□'.

As a result, the identity theorems of nonrigid QMLs are no longer closed under the unrestricted rule of substitution for predicates. They are still closed under substitution of arbitrary nonmodal formulas for predicates (cf. Schurz 1997: 221).

Semantically, nonrigid designators require a world-relativization of the valuation functions for terms; v: $J \times W \rightarrow U$ where $v(t,w) =_{df} v_w(t)$ is the extension of term t at world w, and the partial function $v(t)$: $W \rightarrow U$ such that $v(t)(w) = v_w(t)$ is the *intension* of term t. The debate between Kripke and Lewis, whether individuals in different worlds are strictly identical (Kripke) or merely counterparts of each other (Lewis), is logically less decisive than one might think. Rigid designator axioms are also adequately characterized by the *unique counterpart view*, according to which every individual possesses a unique counterpart in every possible world (see also Forbes 1985: 60ff). We just have to assume that the valuation function v assigns to each term t and world w a pair $\langle d,w\rangle$, which stands for the world-relativized individual *d-in-w*, such that the domain component 'd' of this pair is the same in all worlds. Then, each world w has its own domain $D \times \{w\}$ and each world-relativized individual $\langle d,w\rangle$ has a unique counterpart $\langle d,u\rangle$ in each world $u \in W$. The resulting logic would be Q1ML (but the same modification can be made for Q2MLs). Hence, the important point of a semantics for nonrigid designators, which do not obey rigid identity axioms, is the assumption of a counterpart relation which is *not unique*.

The real problem of nonrigid designators is the semantical interpretation of *quantified de re formulas*. Take, for example, $(M,w) \models \exists x \Box Fx$. This means formally that there exists $d \in D_w$ such that for all w-accessible worlds v: $(M[x:d],v) \models Fx$. But how do we define the x-variation V[x:d] of V^M if designators are non-rigid? The most simple possibility would be to assume that V[x:d] assigns d to x in *all* worlds. In the effect, this means that variables are interpreted as *rigid* designators; only constants are nonrigid. This option is chosen by Thomason (1970). However, the free quantification axiom (fUI) becomes invalid in these systems: from the (fEG)-instance $\Box(t = t) \wedge Et \supset \exists x \Box(x = t)$ the formula $Et \supset \exists x \Box(x = t)$ is provable, though it is invalid, because it requires t to have the same extension at all accessible worlds, which need not be the case (Garson 1984: 262). Hintikka (1970) suggests to replace (fUI) by a complicated instantiation rule, which in case of Thomason's system **Q3-S4** reduces to $\forall xA \supset (\Box Et \supset A[x/t])$ (Garson 1984: 263); generalized completeness proofs for these kinds of systems have not been

found. A possibility of handling systems with rigid variables and nonrigid constants, elaborated by Bowen (1979), is to assume that terms are *local*, that is that their extensions at worlds always exist in that world; this locality option becomes available when terms are nonrigid. Bowen also accepts the nested domain condition, with the result that classical quantification principles are valid in his systems, and generalized proofs of model-completeness are possible.

All systems with rigid variables contain the theorem $\forall x \forall y(x = y \supset \Box x \equiv y)$, which a strict defender of nonrigidity wants to avoid. Another possibility of defining v[x:d] would be to allow V[x:d] to be any function from W into D, satisfying only the restriction that $V_w(x) = d$. In counterpart terminology, this means that anything may count as a counterpart of d in other worlds. This semantics corresponds to Garson's *conceptual* interpretation (1984: 266). Apart from the resulting incompleteness (Garson 1984: 266), this semantics validates the counterintuitive formula $\Box\exists xFx \supset \exists x\Box Fx$, which is a clear reason to reject it (Hughes and Cresswell 1968: 197f).

What one needs is a way to *restrict* the 'allowed' functions over which v[x:d] may range, and the natural way to do this is by way of a *counterpart relation* which specifies the counterparts of $d \in D_w$ in all w-accessible worlds. This account has been developed by Lewis (1968), though not within the framework of modal logic but within that of ordinary first-order logic, and by assuming universal frames. Specifically, Lewis introduces the predicates Ww for 'w is a world,' 'Ixw' for 'object x exists in world w,' and Cxy for 'y is a counterpart of x.' The counterpart relation need neither be symmetric nor transitive. Let us present Lewis' theory in the framework of modal logic, interpreting Cxy as 'y is a counterpart of x in a world w accessible from x's world.' Then Lewis' proposed semantical interpretation of *de re* sentences can be reformulated in this way (1968: 118):

$(M[x_{1-n}:d_{1-n}],u) \models \Box A$ iff for all w-accessible worlds v and *all* $d'_1, \ldots, d'_n$ such that each d'_i is a counterpart of d_i in v, $(M[x_{1-n}:d'_{1-n}],v) \models A$.

$(M[x_{1-n}:d_{1-n}],u) \models \Diamond A$ iff for some w-accessible worlds v and *some* $d'_1, \ldots, d'_n$ such that each d'_i is a counterpart of d_i in v, $(M[x_{1-n}:d'_{1-n}],v) \models A$.

For terms: $(M,u) \models$ $\mathbf{o}A[t_1/x_1, \ldots, t_n/x_n]$ iff $(M[x_{1-n}:v(t_{1-n})],u) \models \mathbf{o}A$, with $\mathbf{o} \in \{\Box,\Diamond\}$; i.e., iff for all/some w-accessible worlds v and all/some $d_1, \ldots, d_n$ such that each d_i is a counterpart of t_i's extension at w in v, $(M[x_{1-n}:d'_{1-n}],v) \models A$.

The problem is that Lewis' counterpart theory, if taken as a semantics for modal logic, is logically not well-behaved. It is not closed under substitution, even not substitution of atomic formulas for propositional variables. For example, $\Box p \supset \Box\Box p$ will be valid on a transitive frame, yet $\Box Ft \supset \Box\Box Ft$ might be invalid in a Lewis model imposed on that frame, because the counterpart relation need not be transitive. More drastically, Wollaston (1994) shows that Lewis' semantics invalidates the modal principles K and M, and even the nonmodal principle UI. Ghilardi (1991) has developed a semantics for nonrigid QML which adopts the nested domain condition and models counterpart relations as functions c: $D_u \to D_v$ for uRv. His systems are logically well-behaved, but he obtains drastic incompleteness results; for example the QML $\mathbf{Alt}_n$ is incomplete in his semantics, though canonical in our Q1-, Q2- and Q3-semantics (cf. corollary 7.5 of

Ghilardi). More recently, Skvortsov and Shehtman (1993) have introduced a new kind of frame semantics, so-called *metaframe semantics*, which is a generalization of Ghilardi's functor semantics. They are able to show that completeness w.r.t. metaframes generally transfers from a PML to its quantificational counterpart. Technically, this is a great success. However, metaframe semantics is *not* based on domains of individuals, but on domains of 'abstract' n-tuples which are not reducible to the nth Cartesian product of an ordinary domain. So far, no one has given a philosophically transparent interpretation of metaframe semantics.

The philosophically more transparent alternative is the *substantial* interpretation of quantifiers, where quantifiers do not range over objects (term extensions), but over functions from worlds into objects (term intensions). This suggestion has been introduced by Hughes and Cresswell (1968: 198ff) and is extensively elaborated in Garson (1984: 267ff); specifically in his system **QS**. One assumes here, for each world, a set of term intensions, that is functions from W to D, which are the 'substances' which exist at that world; quantifiers range over these term intensions. Schurz (1997) shows that with some modifications, Garson's semantics can be reinterpreted from the *objectual* view as a certain kind of counterpart semantics, so-called *world-line semantics*. We call the logics based on it Q3-logics, and explain it as follows.

A 'term-intension,' that is a function l:W → U is called a *worldline* (in analogy to wordlines in Minkowski's space–time diagrams). A worldline l *lands* object d at world w if l(w) = d. The important component of Q3-models is a set L of worldlines ('substances') which specifies the possible term intensions. L determines a four-placed counterpart relation 'object d_1 in u has d_2 as a counterpart in v' defined as follows: there exists a worldline in L which lands d_1 at u and d_2 at v. For each w ∈ W, U_w is the set of objects landed by some wordline at w, that is, the set of all w-counterparts of *possible* objects. Predicate extensions at w are taken from U_w. To obtain a *free* logic version of this semantics we also need a domain function Df which assigns to each world w a subset $D_w \subseteq U_w$ of objects *existing* in w. The world-specific sets of worldlines L_w over which quantifiers range are given as the set of worldlines which land some object in D_w.

Q3ML-DEFINITIONS The Q3-language *ℒQ3* is syntactically like an *ℒQ1*-language; the existence predicate E is defined as in *ℒQ2*. A Q3-model based on a frame ⟨W,R⟩ is a 6-tupe ⟨W,R,L,U,Df,V⟩, with $\emptyset \neq L \subseteq \{l{:}W \to U\}$ a nonempty set of possible worldlines, where $U \neq \emptyset$ is a nonempty set of possible objects; Df: W → U such that $Df(w) =_{df} D_w \subseteq U_w$ is the domain function, where $U_w =_{df} \{d \in U{:}\ \exists l \in L(l(w) = d)\}$ is the set of term-extensions at w. We define $L_w =_{df} \{l \in L{:}\ \exists d \in D_w(l(w) = d)\}$. Concerning V: for each $t \in J$, $V(t) \in L$; and for each n-ary $Q \in \mathcal{R}^n$, $V_w(Q) \subseteq U_w{}^n$. M[x:l] denotes a model which is like M except that it assigns the wordline l to x. The truth clauses are as follows: (i) (M,w) ⊨ $Qt_1 \ldots t_n$ iff $\langle V_w(t_1), \ldots, V_w(t_n)\rangle \in V_w(Q)$; (M,w) ⊨ $t_1 = t_2$ iff $V_w(t_1) = V_w(t_2)$; for propositional operators as before; and for the quantifier: (M,w) ⊨ ∀xA iff for all $l \in L_w$, (M[x:l],w) ⊨ A; this yields (M,w) ⊨ Et iff $V_w(t) \in D_w$ for the existence predicate.

Worldline semantics is fully compatible with the objectual view. Identity and existence statements depend only on the extensions of terms. The truth clauses for quantifiers may be rephrased in Lewis' counterpart style where quantifiers range over objects as follows: (M,w) ⊨ ∀xA [∃xA] iff for all [some, resp.] $d \in D_w$ and $l \in L$ such that

$V_w(l) = d$: (M[x:l],w)$\models$ A. For each particular formula, the quantification over worldlines is eliminable, for example: (M,u) $\models \forall x\Box Fx$ [$\exists x\Diamond Fx$] iff for all [some, resp.] $d \in D_w$, w-accessible world(s) v, and v-counterpart(s) d′ of d: d′ is in $V_v(F)$.

The essential difference to Lewis' counterpart semantics is threefold. First, the counterpart relation defined by worldlines is symmetric and obeys further structural properties which are not satisfied by Lewis' counterpart relation. Second, quantification over counterparts is in worldline semantics governed by the quantifier, but in Lewis' semantics governed by the modal operator. The *de re* formulas $\forall x\Box Fx$ and $\exists x\Diamond Fx$ are evaluated in the same way, but the *de re* formulas $\forall x\Diamond Fx$ and $\exists x\Box Fx$ are evaluated differently: in Lewis semantics, (M,w) $\models \forall x\Diamond Fx$ iff for every $d \in D_w$ there exists *some* w-accessible world u such that *some* counterpart d′ of d in u is in $V_u(F)$, while in worldline semantics, (M,w) $\models \forall x\Diamond Fx$ iff for every $d \in D_w$ there exists *some* w-accessible world u such that *every* counterpart d′ of d in u is in $V_u(F)$. Likewise for the formula $\forall x\Box Fx$. Third, Lewis' semantics does not assign wordlines (term intensions) to terms t, but quantifies over the counterparts of term extensions $V_w(t)$ in de re scopes, while wordline semantics determines the counterparts of $V_w(t)$ by their wordlines l(t). For example, assume Dudu is the name of an amoeba a at world w which in all accessible worlds u splits up into two, namely b and c, where b keeps alive and c is dying in u, and we decide that Dudu should name b (but not c) at all w-accessible u. Then the sentence 'necessarily Dudu is alive' is true in worldline semantics, but false in Lewis style counterpart semantics. Note finally that the language of worldline semantics has a greater expressive power than that of Lewis' counterpart semantics. Lewis' modal operators ($\Box_L$, $\Diamond_L$) are *definable* within worldline semantics as follows (Schurz 1997: 222):

$\Box_L A[t_{1-n}/x_{1-n}] =_{df} \forall y_{1-n}(\wedge\{t_i = y_i\colon 1 \leq i \leq n\} \supset \Box A[y_{1-n}/x_{1-n}])$, and
$\Diamond A[t_{1-n}/x_{1-n}] := \exists y_{1-n}(\wedge\{t_i = y_i\colon 1 \leq i \leq n\} \wedge \Diamond A[y_{1-n}/x_{1-n}])$, where $x_1, \ldots, x_n = \mathcal{T}(A)$.

The essential difference of worldline semantics as compared to Garson's substantial semantics is that Garson does not define the world-relative sets of worldlines ('substances') L_w by the extension of an ordinary existence predicate, as we did, but he introduces them directly, without such a predicate, and the truth clause of his existence predicate is: (M,w) $\models$ Et iff $v(t) \in L_w$ (Garson 1984: 279). This turns his existence predicate into an 'intensional' one which contains as its world-specific extension a set of term intensions. As a result, substitution of E in the identity axiom (ISub) is not allowed in Garson's system (1984: 268); though it is allowed in our system. Besides this greater simplicity, it seems to be philosophically more intuitive *not* to assume world-specifics set L_w as a *primitive* notion, for the existence of worldlines ('substances') is not a contingent matter; only the existence of objects is contingent.

The logic **Q3K** is defined like **Q2K** except that the rigid identity axiom rI$\neg$ is dropped and rISub is replaced by ISub above. ISub is only closed under restricted substitution for predicates, while the other axiom schemata are closed under general substitution. On this reason, normal Q3-logics cannot be defined as before. We rather have to define a normal Q3-logic as a subset $\mathbf{L} \subseteq \mathcal{LQ}3$ which is representable as **Q3KX**, that is it contains all axioms (not merely the schemata) of **Q3K**, is closed under the rules of **Q3KX** (TautR, fUG, N) and contains all (unrestricted substitution) instances of the additional set of axiom schemata **X**, except for additional identity axioms in **X** to which only non-

modal substitution applies (Schurz 1997: 221). The canonical model $M_c(\mathbf{L})$ of a Q3ML **L** need no longer be relativized to an initial saturated set Δ which determines rigid identities. It is defined as $\langle W_c, R_c, L_c, U_c, Df_c, V_c\rangle$, where W_c is now the set of all G-saturated formula sets, R_c is an s usual, and $V_c(t)$: $W_c \to U_c$ is such that $V_{c,w}(t) = \{t' \in J^*: t = t' \in w\}$, $U_c = \{V_{c,w}(t): w \in W, t \in J^*\}$, $L_c = \{V_c(t): t \in J^*\}$, $Df_c: W_c \to Pow(U_c)$ such that $Df(w) = \{V_c(t): t \in J^*, Et \in w\}$. Model-completeness is proved with help of the stronger G-rule GUG in the same way as for Q2-logics (Garson 1984: 282ff). We thus arrive at the *QG3ML-model-completeness-theorem*: all normal QG3MLs are adequately characterized by the class of their models, and **Q3K** is canonical. Q(G)3-logics behave similar as Q2-logics: we can show that GUG is admissible in all normal extensions of **Q3B**, that restricted canonicity-transfer holds, and that the frames of Q3-logics are the same as their Q1-counterparts (Schurz 1997: ch. 10.8–10). A different technique proves model-completeness for Q3-logics by introducing for each canonical world a new set of constants. This proof avoids the stronger G-rules, but it is not completely general: certain properties of the canonical frame cannot be proved in the standard way because canonical worlds don't share the same language (Garson 1984: 276–81).

References

Barcan (Marcus), R. (1946) A functional calculus of first order based on strict implication. *Journal of Symbolic Logic*, 11, 1–16.

Barcan-Marcus, R. (1960) Extensionality. *Mind*, 69, 55–62. Reprinted in Linsky (1971), pp. 44–51.

Barcan-Marcus, R. (1963) Modalities and intensional languages. In *Boston Studies in the Philosophy of Science* 1. Reprinted in Barcan-Marcus, R. (1993), *Modalities* (ch. 1). Oxford: Oxford University Press.

Bowen, K. (1979) *Model Theory for Modal Logic. Kripke Models for Modal Predicate Calculi*. Dordrecht: Reidel.

Bull, R. and Segerberg, K. (1984) Basic modal logic. In D. Gabbay and F. Guenthner (eds.) (1984), pp. 1–88.

Carnap, R. (1946) Modality and quantification. *Journal of Symbolic Logic*, 11, 33–64.

Carnap, R. (1947) *Meaning and Necessity*. Chicago: University of Chicago Press.

Chagrov, A. and Zakharyaschev, M. (1997) *Modal Logic*. Oxford: Oxford University Press.

Chellas, B. F. (1980) *Modal Logic*. Cambridge: Cambridge University Press.

Chellas, B. and Segerberg, K. (1996) Modal logics in the vicinity of S1. *Notre Dame Journal of Formal Logic*, 37/1, 1–34.

Dummett, M. and Lemmon, E. J. (1959) Modal logics between S4 and S5. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 3, 250–64.

Fine, K. (1975a) Some connections between elementary and modal logic. In S. Kanger (ed.), *Proceedings of the Third Scandinavian Logic Symposium* (pp. 15–19). Amsterdam: North Holland.

Fine, K. (1975b) Normal forms in modal logic. *Notre Dame Journal of Formal Logic*, 16, 229–34.

Fine, K. (1978) Model theory for modal logic. Part I + II. *Journal of Philosophical Logic*, 7, 125–56 (Part I: The *de re/de dicto* distinction), 277–306 (Part II: The elimination of de re modality).

Fine, K. (1985) Logics containing *K4*. Part II. *Journal of Symbolic Logic*, 50, 619–51.

Fine, K. and Schurz, G. (1996) Transfer theorems for multimodal logics. In J. Copeland (ed.), *Logic and Reality: Essays in Pure and Applied Logic* (pp. 169–213). Oxford: Oxford University Press.

Fitting, M. (1983) *Proof Methods for Modal and Intuitionistic Logics*. Dordrecht: Reidel.

Forbes, G. (1985) *The Metaphysics of Modality*. Oxford: Clarendon Press.
Gabbay, D. (1976) *Investigations in Modal and Tense Logics*. Dordrecht: Reidel.
Gabbay, D. and Guenthner, F. (eds.) (1984) *Handbook of Philosophical Logic*, vol. II: *Extensions of Classical Logic*. Dordrecht: Reidel.
Garson, J. W. (1984) Quantification in modal logic. In Gabbay and Guenthner (1984), pp. 249–308.
Ghilardi, S. (1991) Incompleteness results in Kripke semantics. *Journal of Symbolic Logic*, 56, 517–38.
Goldblatt, R. (1991) The McKinsey axiom is not canonical. *Journal of Symbolic Logic*, 56, 554–62.
Gottlob, G. (1999) Remarks on a Carnapian extension of S5. In J. Woleński and E. Köhler (eds), *Alfred Tarski and the Vienna Circle* (pp. 243–59). Dordrecht: Kluwer.
Hendry, H. E. and Pokriefka, M. L. (1985) Carnapian extensions of S5. *Journal of Philosophical Logic*, 14, 111–28.
Hintikka, J. (1961) Modality and quantification. *Theoria 27*. Reprinted in J. Hintikka. *Models for Modalities* (pp. 57–70). Dordrecht: Reidel.
Hintikka, J. (1970) Existential and uniqueness presuppositions. In K. Lambert (ed.), *Philosophical Problems in Logic*. Dordrecht: Reidel.
Hughes, G. E. and Cresswell, M. J. (1968) *An Introduction to Modal Logic*. London: Methuen.
Hughes, G. E. and Cresswell, M. J. (1984) *A Companion to Modal Logic*. London and New York: Methuen.
Hughes, G. E. and Cresswell, M. J. (1986) A companion to modal logic – some corrections. *Logique et Analyse*, 113, 41–51.
Kanger, S. (1957a) *Provability in Logic*. Stockholm: Almqvist & Wiksell.
Kanger, S. (1957b) The morning star paradox. *Theoria*, 23, 1–11.
Kleene, S. C. (1971) *Introduction to Metamathematics*. Groningen: Wolters-Noordhoff.
Kracht, M. and Wolter, F. (1991) Properties of independently axiomatizable bimodal logics. *Journal of Symbolic Logic*, 56, 1469–85.
Kracht, M. and Wolter, F. (1997) Simulation and transfer results in modal logic – a survey. *Studia Logica*, 59/2, 149–77.
Kripke, S. (1959) A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24/1, 1–14.
Kripke, S. A. (1963a) Semantical analysis of modal logic I: Normal modal propositional calculi. *Zeitschrift für mathemathische Logik und Grundlagen der Mathematik*, 9, 67–96.
Kripke, S. A. (1963b) Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16, 83–94.
Kripke, S. A. (1972) *Naming and Necessity*. Oxford: Basil Blackwell.
Kutschera, F. (1993) Causation. *Journal of Philosophical Logic*, 22, 563–88.
Lemmon, E. J. and Scott, D. (1966) *Intensional Logic. Preliminary Draft of Initial Chapters by E. J. Lemmon*. In K. Segerberg (ed.), *The "Lemmon Notes": An Introduction into Modal Logic*. Oxford: Blackwell.
Lewis, C. I. and Langford, C. H. (1932) *Symbolic Logic*. New York: Dover Publications.
Lewis, D. (1968) Counterpart theory and quantified modal logic. *Journal of Philosophy*, 65, 113–26.
Lewis, D. (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.
Linsky, L. (ed.) (1971) *Reference and Modality*. Oxford: Oxford University Press.
Machover, M. (1996) *Set Theory, Logic, and their Limitations*. Cambridge: Cambridge University Press.
Makinson, F. (1966) On some completeness theorems in modal logic. *Zeitschrift für mathemathische Logik und Grundlagen der Mathematik*, 14, 202–12.
Prendinger, H. and Schurz, G. (1996) Reasoning about action and change: a dynamic logic approach. *Journal of Logic, Language and Information*, 5, 209–45.

Prior, A. (1957) *Time and Modality*. Oxford: Oxford University Press.
Quine, W. V. O. (1943) Notes on existence and necessity. *Journal of Philosophy*, 40, 113–27.
Quine, W. V. O. (1953) Reference and modality. In W. V. O. Quine, *From a Logical Point of View*. Reprinted in Linsky (ed.) (1971), pp. 17–34.
Putnam, H. (1975) The meaning of "meaning." In K. Gunderson (ed.), *Language, Mind and Knowledge*. Minnesota: University of Minnesota Press.
Rautenberg, W. (1979) *Klassische und Nichtklassische Aussagenlogik*. Brauschweig: Friedrich Vieweg & Sohn.
Sahlqvist, H. (1975) Completeness and correspondence in the first and second order semantics for modal logic. In S. Kanger (ed.), *Proceedings of the Third Scandinavian Logic Symposium*, 110–43. Amsterdam: North-Holland.
Schurz, G. (1994) Admissible versus valid rules. A case study of the modal fallacy. *The Monist*, 77/3, 376–88.
Schurz, G. (1995) Most general first order theorems are not recursively enumerable. *Theoretical Computer Science*, 148, 149–63.
Schurz, G. (1997) *The Is–Ought Problem. A Study in Philosophical Logic*. (Trends in Logic, vol. 1 – Studia Logica Library.) Dordrecht: Kluwer.
Schurz, G. (1999) Tarski and Carnap on logical truth – or: what is genuine logic? In J. Wolenski and E. Köhler (eds.), *Alfred Tarski and the Vienna Circle* (pp. 77–94). Kluwer: Dordrecht.
Schurz, G. (2000) Carnap's modal logic. In W. Stelzner and M. Stöckler (eds.), *Nichtklassische logische Ansätze im Übergang von traditioneller zu moderner Logik*. Paderborn: Mentis Verlag.
Segerberg, K. (1971) *An Essay in Classical Modal Logic*. Uppsala: Filosofiska Studier.
Shimura, T. (1993) Kripke completeness of some intermediate predicate logics with the axiom of constant domain and a variant of canonical formulas. *Studia Logica* 52, 23–40.
Skvortsov, D. P. and Shehtman, V. B. (1993) Maximal Kripke-type semantics for modal and super-intuitionistic predicate logics. *Annals of Pure and Applied Logic*, 63, 69–101.
Smorynski, C. (1984) Modal logic and self-reference. In Gabbay and Guenthner (eds.) (1984), pp. 441–96.
Thomason, R. S. (1970) Some completeness results for modal predicate calculi. In K. Lambert (ed.) (1970) *Philosophical Problems in Logic* (pp. 56–76). Dordrecht: Reidel.
Thomason, S. K. (1972) Semantic analysis of tense logics. *Journal of Symbolic Logic*, 37, 150–8.
Van Benthem, J. F. (1983) *Modal Logic and Classical Logic*. Napoli: Bibliopolis.
Van Benthem, J. F. (1984) Correspondence theory. In Gabbay and Guenthner (1984), pp. 167–248.
Van Dalen, D., Doets, H. C. and De Swart, H. (1978) *Sets: Naive, Axiomatic and Applied*. Oxford: Pergamon Press.
Von Wright (1951) *An Essay in Modal Logic*. Amsterdam: North Holland.
Wansing, H. (ed.) (1996) *Proof Theory of Modal Logic*. Dordrecht: Kluwer.
Wollaston, L. (1994) Counterpart theory as a semantics for modal logic. *Logique et Analyse*, 147–8, 255–63.

30

Epistemic Logic

NICHOLAS RESCHER

1 Accessible Knowledge

The antecedents of epistemic logic – the logical theory of propositions regarding belief, knowledge and, by extension, also assertion, assumption, and presupposition, go back to the Middle Ages – especially to William of Ockham (see Boh 1993). However, as a significant branch of philosophical logic, epistemic logic is an innovation of the period 1945–75, the first generation after World War II. At its center lies the relational operator *Kxp* for '*x* knows that *p*,' where *Kx* can be thought of as a parametized modality characterizing the person-relative epistemic status of a proposition. For such an operator to stand coordinate with something worthy of being called a 'logic' it is requisite to begin with a detailed analysis of the sort of 'knowledge' that is to be at issue. Construed in this way, with a focus upon knowledge (*epistêmê*) as such, epistemic logic is part of a broader project that addresses also the logic of belief, supposition, conjecture, etc. – that is, a logic of cognitive processes in general.

The conception of 'knowledge' represents clearly a flexible and internally diversified idea. In general terms, it relates to the way in which persons can be said to have access to information. This can, of course, occur in rather different ways:

- *Occurrent knowledge* This is a matter of actively paying heed or attention to accepted information. A person can say: 'I am (at this very moment) considering or attending to or otherwise taking note of the fact that hydrogen is the lightest element.' The present evidence of our senses – 'I see a cat on the mat' – is an example of this sort of thing.
- *Dispositional knowledge* This is a matter of what people would say or think if the occasion arose – of what, for example, they would say if asked. Even when *X* is reading Hamlet or, for that matter, sleeping, we would say that this individual knows (in the presently relevant dispositional manner) that Tokyo is the capital of Japan.
- *Accessible knowledge* This is a matter not of what a person *would* say if asked (= dispositional knowledge) but of what one *could* say if he is sufficiently clever about using the information that is at one's disposal occurrently or dispositionally. In other words it is what is *implied by* or *inferable from* the facts he already knows in any of these senses.

As we propose to understand it here, knowledge will be construed recursively in that third sense of what is *inferentially accessible* from one's own information. For reasons that will become increasingly clear below, our focus is upon *available* rather than *occurrent* knowledge. Accordingly, a person knows something (1) if this is known to him occurrently, or (2) if this is known to him dispositionally, or (3) if this can be derived by logical deduction or by other secure inferential means from information that is (already) known to him. It is this recursive conception of knowledge that will concern us here, the relationship Kxp, for 'x knows that p,' being understood in the specified manner. We thus immediately secure such relationships as the *Inferential Accessibility Principle*:

$$[Kxp \,\&\, (p \rightarrow q)] \rightarrow Kxq$$

as well as the Knowledge Compilation (or Conjunctivity) Principle:

$$(Kxp \,\&\, Kxq) \rightarrow Kx(p \,\&\, q)$$

(*Note*: In all such formulas, apparently free variables are to be thought of as bound by initial universal quantifiers.)

Our knowers can systematically draw appropriate conclusions and 'put two and two together.' The controlling consideration here is not that they are 'logically omniscient,' but rather that the availability-oriented sense of knowing that is at issue here provides for such inferential projection. Admittedly, to construe knowing in terms of these capabilities is to interpret the idea in a particularly generous sense. However, this approach is amply justified by the aims of the enterprise in so far as one's special interest is in the *limits* of knowledge.

Given this inclusive and generous sense of knowing, it should be noted that if p is a thesis demonstrable on logico-conceptual grounds alone, then p will be universally available since it can be deductively derived from any thesis whatsoever so that Kxp can be held to obtain for any knower x. Accordingly, given the inferential accessibility reading of K, we have

$$\mathbf{N}p \rightarrow (\forall x)Kxp$$

subject to the convention that $\mathbf{N}$ and $\mathbf{P}$ will represent logico-conceptual necessity and possibility, respectively, within the setting of modal system S5 of C. I. Lewis. This is simply an aspect of our governing supposition that logico-conceptual matters are universally accessible.

Inferential prowess notwithstanding, the 'knowers' at issue in this discussion are finite knowers. Thus while we have it that *Every knower knows something* (i.e. some truths, and specifically all necessary ones), and thus $(\forall x)(\exists p)Kxp$, we also have it that: *Every knower is ignorant of some truth*: $(\forall x)(\exists p)(p \,\&\, \sim Kxp)$. Moreover, any truth is a candidate for being known: whenever p is true, then $\mathbf{P}(\exists x)Kxp$.

As these remarks indicate, the present discussion will move beyond quantified modal logic (QML) to articulate principles of a qualified modal *epistemic* logic (QMEL).

2 Actual vs. Putative Knowledge

The distinction between actual and merely putative knowledge is critical for present purposes: we can and must operate with the distinction between 'our truth' and 'the truth.' Nevertheless, while we realize full well that some of the claims that we regard as true will turn out to be false, we of course cannot particularize here: 'Give me an example of a proposition that you accept as true but that really isn't' represents an absurd request. It lies in the nature of things that we see *our* truth as *the* truth in the realm of specifics.

Rational people are committed to seeing their knowledge as real knowledge – and therefore as subject to those principles which hold for genuine knowledge in general. Now in taking 'our own putative knowledge' to be true – that is, viewing it as *actual* knowledge – we accept the principle:

$Kip \rightarrow p$ (where *i* should be construed as 'I myself' and/or 'we ourselves').

And since we standardly credit others with the same privileges and liabilities that we claim for ourselves we can generalize the preceding principle to:

$Kxp \rightarrow p$

When claiming Kxp we take the stance that p is something that x really and truly knows to be so. That means that we take ourselves to know p to be true. On this basis, $Kxp \,\&\, \sim p$ – 'x knows that p but it isn't so' – is to all intents and purposes a self-contradiction. We would not say that someone *knows* something if we thought that this were not so but would instead have to say something like 'he merely *thinks* he knows that p.' For this reason $Kxp \,\&\, \sim Kip$ is also a comparable self-contradiction. To attribute knowledge of a *particular* fact to another is also to claim it for oneself. On the other hand, the generic $(\exists p)(Kxp \,\&\, \sim Kip)$ – that is 'x knows something I don't' – is a perfectly plausible proposition. It is just that one cannot concretize it to the level of specifics: particularizing existential instantiation becomes impracticable here.

The thesis $Kxp \rightarrow p$ also means that no knower ever knows that he is mistaken about something concrete that he takes himself to know. This was a commonplace among medieval logicians, who held that *Nihil scire potest nisi verum* (see Boh 1993: 48). The thesis $Kxp \,\&\, Kx(\sim Kxp)$ is thereby self-contradictory since its second conjunct entails the denial of what the first conjunct affirms. The idea at issue here is not new but was also a commonplace among medieval logicians. Thus Albert of Saxony (ca. 1325–90) argued in his treatise on *Insolubilia* that "Socrates knows that he is mistaken in believing A" is a self-contradictory contention. (See Kretzmann and Stump 1988: 363–4.)

3 Levels of Acceptance and Rejection

In articulating epistemological principles we must come to terms with the fact that one can distinguish three different levels or bases of assertability on which such principles can be affirmed:

1. *Conceptual truth* A thesis that holds good on logico-conceptual goals of meaning and usage alone; its denial involves one in saying things which, while perhaps understandable, are acceptable only subject to elaborate explanations and qualification and in their absence are effectively paradoxical.
2. *Contingent truth* A thesis whose acceptability cannot be substantiated by any amount of merely conceptual or verbal elucidation but whose validity roots in the cognitively discernible contingent features of the real world.
3. *Plausible truth-candidates* A thesis not clearly spoken for by the available facts but for whose substantiation cogent considerations of plausibility can be adduced and which therefore merits at least qualified endorsement and provisional acceptance.

Each of these defines a level of tenability or assertability which may be characterized as levels 1, 2, and 3, respectively. (The lower the tenability level of a principle, the more unproblematic and probatively secure it will be.)

Let ⊢Z indicate (as usual) that Z is an assertion of the system we are engaged in formulating. Then with respect to level-one principles we have:

If $\vdash_1$ Z then **N**Z – and therefore also, as we have seen, **N**$(\forall x)Kx$Z

Since all the principles of our system are to be seen as matters of logico-conceptual necessity, the unqualified prefix ⊢ is to be construed as $\vdash_1$. In being matters of logico-conceptual necessity, all level-one principles are accordingly universally available in the inferential-accessibility mode of knowledge.

With respect to level two principles by contrast we merely have:

If $\vdash_2$ Z then *Ki*Z and thus also (but merely) $(\exists x)Kx$Z (whence also Z)

Since *i* = we ourselves, it is unavoidable that Z be seen as representing something that we really know to be true.

Finally, with respect to level-three principles we merely have

If $\vdash_3$ Z then Z

Here we do indeed regard the thesis in question as being true but without claiming actual *knowledge* of the matter. For in general, the propositions we ourselves see as eminently plausible are accepted by us as true. (In theory something viewed as highly plausible can in fact be false, but we are of course incapable of giving a current first-hand example: 'I see *p* as deserving of acceptance, but it is false' comes close to being a contradiction in terms. Illustrations from the past or those involving others are, of course, another matter.) Here at level three we claim truth in a tentative and provisional way that falls short of actual knowledge. Accordingly, the inference

$\vdash_3$ Z then $(\exists x)Kx$Z

is inappropriate – and thus *a fortiori* also the inference to *Ki*Z. On the contrary, we have it that if $\vdash_3$ Z then $\sim(\exists x)Kx$Z. Nobody *knows* a level-three principle (ourselves included!):

every assertion at level-three has to be seen as a truth that is not actually *known*. Such theses may be surmised or presumed, but even at best they are plausible truths that nobody *knows* to be such – such as the thesis 'There are mountains on the far side of the moon' is the cognitive state of the art of the nineteenth century.

The existence of the third level of assertion is a reminder that epistemology is broader than the theory of *knowledge*. For matters of presumption, conjecture, reasonable belief, and warranted assertability also clearly fall within its purview.

On this basis, then, all three of these modes of 'assertion' do indeed convey a commitment – an *assertion*. A claim that Z is the case obtains in every instance, but with different assertoric modalities, so to speak. For in this context we must deploy the distinction between what is known to be true and what is accepted or asserted (as true) on a weaker basis – conjuncture, plausible suppositions, or the like. The latter sort of thing is being claimed as true, alright, but in a substantially less firm and confident tone of voice. However, the tenability of level-three principles is at odds with acknowledging that someone knows the contrary. For note that when $(\forall x)\sim Kx\sim Z$ is false, so that $\sim(\forall x)\sim Kx\sim Z$ obtains, then of course we will have $(\exists x)Kx\sim Z$. This means that $\sim Z$ would have to obtain (at least at level 2), so that Z would not be a level-three assertion after all – contrary to our initial stipulation.

Despite the acceptability of $(\exists p)(p\ \&\ \sim(\exists x)Kxp)$, no *particular* proposition of the form $p_0\ \&\ \sim(\exists x)Kxp_0$ is ever assertable at levels one or two. For asserting this at level one would mean accepting $\mathbf{N}p_0$ which is at odds with $\sim(\exists x)Kxp_0$. And asserting it at level two would involve a commitment to Kip_0 which is also at odds with $\sim(\exists x)Kxp_0$.

One can, of course, use some epistemic principles to deduce others; here, as elsewhere, inference from givens is a cognitively viable project. And the *epistemic* level of a conclusion derived from premises *cannot be greater than the largest index-level of the premises required for its derivation*. In point of cognitive tenability or assertability, the status of a derived thesis cannot be weaker, so to speak, than the weakest link among the premises from which it derives.

Theses that entail the negation (denial) of an assertion must themselves be denied (at the appropriate level). We shall employ the symbol $\dashv$ to indicate denial/rejection. This should be subscripted to indicate the appropriate level, subject to the convention that $\dashv Z$ obtains at a level iff $\vdash \sim Z$ does so.

4 Level One Principles: Logico-Conceptual Truths

Let us consider some examples of cognitive principles at each assertion level category, beginning with the first, that of principles which inhere in the very nature of the logico-conceptual construction of 'knowledge' as *accessible* knowledge. The following seven basic principles obtain here:

K_1 *Knower capacity*
$(\forall x)(\exists p)Kxp$ and even more strongly $(\forall x)(\exists p)[Kxp\ \&\ \sim\mathbf{N}p]$

K_2 *Knower finitude*
$(\forall x)(\exists t)\sim Kxt$ or equivalently $\sim(\exists x)(\forall t)Kxt$, where t ranges specifically over truths.

K_3 *Knowledge authenticity*
$\sim(\exists t)(\exists x)Kx\sim t$ or equivalently $(\forall t)(\forall x)\sim Kx\sim t$.

K_4 *Inferential accessibility*
$(p \to q) \to (Kxp \to Kxq)$

K_5 *Conjunctivity*
$(Kxp \ \& \ Kxq) \to Kx(p \ \& \ q)$

K_6 *Reflexivity*
$Kxp \to KxKxp$

K_7 *Truth Availability*
$(\forall t)\mathbf{P}(\exists x)Kxt$

Here **N** and **P** represent logico-conceptual necessity and possibility, respectively, and $\to$ is a strong (logico-conceptual) implication such that $p \to q$ is equivalent with $\mathbf{N}(p \to q)$. Also, the variables *t*, *t'*, *t''*, etc. will serve to range over truths. And throughout, free variables are to be taken as tacitly bound to initial universal quantifiers.

Each of these principles merits a brief explanation.

K_1 $(\forall x)(\exists p)(Kxp \ \& \ \sim\mathbf{N}p)$ simply asserts: *Every knower knows something – and indeed some contingent (i.e. non-necessary) truth or other.* This obtains simply in virtue of the fact that we are supposed to be talking about knowers.

K_2 $(\forall x)(\exists t)\sim Kxt$ reflects the fact that we are dealing with finite knowers. In the present context of discussion, *no knower is omniscient*; none knows of all truths that they are true – not even on the present generously undemanding construal of knowledge. Since *t* ranges specifically over truths we have it that, for example, $(\exists t)Kxt$ comes to $(\exists p)(p \ \& \ Kxp)$.

K_3 $(\forall t)(\forall x)\sim Kx\sim t$ asserts: *Only true propositions can be known.* This thesis roots in the very nature of 'knowledge' as this concept is generally understood. For it makes no sense to say: '*x* knows that *p*, but *p* is not true.' Of course, someone may *think* or *believe* that he knows something that is false. But to say that he actually knows it is to acknowledge its truth.

Let us further adopt the abbreviation *Up* for $\sim(\exists x)Kxp$ or equivalently $(\forall x)\sim Kxp$ – that is, for '*p* is unknown.' Then the just-stated finding means that $(\forall t)U(\sim t)$. No one knows something that is false, that is: Nobody knows an *untruth* to be the case. (But of course one can know *that* it is an untruth.)

K_4 $(p \to q) \to (Kxp \to Kxq)$. *Knowers automatically know the things that follow from what they know.* This obtains because it is the tacit or implicit sense of 'knowledge' as inferentially accessible information that is at issue in our discussion.

Since in virtue of K_4 our knowers know all necessary propositions, we of course have it that every knower knows *that* any given *p* is true-or-false: $(\forall x)Kx(p \vee \sim p)$ or equiva-

lently $\sim(\exists x)\sim Kx(p \vee \sim p)$. But in view of K_2 there certainly can be knowers who do not know *whether p* is true or is false: $(\exists x)(\sim Kxp \,\&\, \sim Kx\sim p)$.

K_5 $(Kxp \,\&\, Kxq) \rightarrow Kx(p \,\&\, q)$. *Knowers know conjointly and collectively anything they know distributively.* This too obtains in virtue of the generous accessibility-oriented sense of 'knowledge' that concerns us here, which supposes that knowers 'can put two and two together.'

K_6 $Kxp \rightarrow KxKxp$. *When knowers know something this very fact is cognitively accessible to them.* This again follows from the presently operative accessibility-geared sense of knowledge. For clearly, when knowledge is construed as *available* knowledge – that is, in terms of what can be inferred on the basis of what is known – then Kxp will carry $KxKxp$ in its wake. When a certain fact is known to someone, they are in a position to infer that this is so. (Observe that K_6 yields $Kx\sim p \rightarrow Kx\sim Kx\sim p$, which is quite different from and emphatically does not imply $\sim Kxp \rightarrow Kx\sim Kxp$.)

K_7 $(\forall t)\mathbf{P}(\exists x)Kxt$. *Any actual truth is (in theory) knowable.* Such potential availability also inheres in our understanding of the relationship of knowers to knowledge. (Note that this principle is equivalent with $\sim(\exists t)\mathbf{N}U(t)$: no truths are *necessarily* unknown.

K_7 stipulates that any truth is a candidate for knowledge. This reflects our present understanding of **N** and **P** as logico-conceptual necessity/possibility rather than with physical necessity/possibility. It is certainly conceivable that some region of physical reality is such that its facts are inaccessible to intelligent creatures.

Could K_7 be strengthened to $(\forall t)\mathbf{P}(\forall x)Kxt$? This would preclude the prospect of 'blind spots' – bits of self-knowledge inherently unavailable to the subject himself. (On this theme see Sorensen 1988.) On this basis it seems unacceptable.

Note moreover that accepting $(\forall t)\mathbf{P}(\exists x)Kxt$ does *not* mean that any truth is knowable by some actual existent $(\forall t)(\exists x)\mathbf{P}Kxt$? The knowability at issue looks not to actual but to merely possible knowers.

5 Further Consequences

Given the principles K_1–K_7 formulated above, one can proceed to derive various further epistemic principles by purely logical means:

K_8 *Conjunctivity*
$Kx(p \,\&\, q) \rightarrow (Kxp \,\&\, Kxq)$
Knowledge of a conjunction is tantamount to knowledge of its conjuncts.
This follows from K_4 and K_5.

K_9 *Substitutivity*
$(p \rightarrow q) \rightarrow (Kxp \rightarrow Kxq)$
To know something is to know it in all of its logically equivalent guises.
This thesis pivots on K_3.

K_{10} *K-Consistency*
$Kxp \rightarrow \sim Kx\sim p$
This follows directly from K_3. (The prospect of ignorance – of having both $\sim Kxp$ and $\sim Kx\sim p$ obtain – means that the converse does not hold.)

K_{11} *Transmissibility*
$[Kxp \,\&\, Kx(p \rightarrow q)] \rightarrow Kxq$
This follows from K_4 and K_5.

K_{12} *Self-limitation*
$(\forall x)Kx(\exists t)\sim Kxt$
By K_2 we have $(\forall x)(\exists t)\sim Kxt$. And since this is a level 1 principle we will also have $(\forall y)Ky(\forall x)(\exists t)\sim Kxt$. This entails $(\forall y)Ky(\exists t)\sim Kyt$. Not only are individuals not omniscient, but they all know it.

In accepting that another knows a certain fact one is thereby effectively claiming that fact as part of one's own knowledge. And so, to know that another person knows some specific fact one must know this fact oneself. We thus have:

K_{13} *Knowledge cooptation*
$KxKyp \rightarrow Kxp$
To know that someone actually knows some fact to be so one must know this fact itself.

This principle can be derived from the preceding considerations by the following argument:

1.	$[Kxp \,\&\, (p \rightarrow q)] \rightarrow Kxq$	From K_3
2.	$(KxKyp \,\&\, (Kyp \rightarrow p)] \rightarrow Kxp$	From (1) by substituting Kyp/p and p/q
3.	$Kyp \rightarrow p$	From K_2
4.	$KxKyp \rightarrow Kxp$	From (2), (3)

This means that the specifically *acknowledged* knowledge of others is also knowledge. (Of course it will not be the case for unacknowledged knowledge. We certainly do *not* have: $Kxp \rightarrow Kyp$.)

Note further that we have the principle:

K_{14} *Necessity cognition*
$\mathbf{N}p \rightarrow (\forall x)Kxp$
Logico-conceptual truths are cognitively available to all.

This principle pivots on K_4 via the following proof:

1. For any x: Kxq for some suitable q, by K_1.
2. Whenever $\mathbf{N}p$, then $q \rightarrow p$, for any q, by mere logic.
3. Whenever $\mathbf{N}p$, then Kxp rom (1) and (2) by K_4.
4. $(\forall p)(\mathbf{N}p \rightarrow Kxp)$ from (1)–(3).
5. $(\forall p)(\mathbf{N}p \rightarrow (\forall x)Kxp)$ from (4) since x is a free variable. Q.E.D.

Note, however, that the converse of K_{14} does *not* hold: some merely contingent fact might well be known universally.

Via the substitution $\mathbf{N}p/p$ K_{14} yields:

K_{15} *Necessity recognition*
$\mathbf{N}p \rightarrow (\forall x)Kx\mathbf{N}p$
Knowledge of necessity is universal. This principle represents a salient feature of inferentially accessible knowledge.

K_1 has it that every knower knows some truth $(\forall x)(\exists t)Kxt$. In virtue of K_{15}, we have the stronger thesis that there are truths that everyone knows $(\exists t)(\forall x)Kxt$. For any necessary truth clearly fills the bill here, given the presently operative liberal construction of *K* as *available* knowledge.

6 Cognitive Limitations

Let us consider somewhat more closely the matter of ignorance and unknowing, recalling that $U(p)$ comes to: $\sim(\exists x)Kxp$ or equivalently $(\forall x)\sim Kxp$.

Under what conditions on *f* would we have it as a general principle that $f(p)$ entails $Uf(p)$? Note that this would mean that $f(p) \rightarrow \sim(\exists x)Kxf(p)$ or equivalently $(\exists x)Kxf(p) \rightarrow \sim f(p)$. Since K_3 has it that the antecedent yields $f(p)$, it follows that there can be no principle of the indicated format as long as $f(p)$ is self-consistent. No significant feature of *p* is automatically unknowable.

A further important epistemic principle is represented by the thesis:

K_{16} *Cognitive myopia*
$\sim(\exists p)(\exists x)Kx(\sim Kxp \,\&\, p)$ or equivalently $(\forall x)(\forall p)\sim Kx(p \,\&\, \sim Kxp)$
Nobody ever knows of a proposition that while they do not know it, it is nevertheless true.

PROOF

1.	$Kx\sim Kxp \rightarrow \sim Kxp$	From K_2
2.	$\sim(Kx\sim Kxp \,\&\, Kxp)$	From (1)
3	$\sim Kx(\sim Kxp \,\&\, p)$	From (2), K_{14}. Q.E.D.

It is important to observe that the thesis at issue here – or equivalently $(\forall x)\sim(\exists p)Kx(p \,\&\, \sim Kxp)$ – differs significantly from $(\forall x)\sim Kx(\exists p)(p \,\&\, \sim Kxp)$ or equivalently $(\forall x)\sim Kx(\exists t)\sim Kxt$ or $(\forall x)\sim Kx\sim(\forall t)Kxt$, that is, 'For aught that anyone knows they know it all.' This latter contention is emphatically unacceptable.

A pivotal fact of the cognitive domain is:

K_{17} *Knowledge limitation*
$\sim(\forall t)(\exists x)Kxt$ or equivalently $(\exists t)(\forall x)\sim Kxt$ or $(\exists t)\sim(\exists x)Kxt$ or $(\exists t)Ut$.
There are altogether unknown truths: it is not the case that all truths are known.

This is easily established on the basis of the prior stipulations. For since K_2 assures that we are, by hypothesis, dealing with finite knowers, it transpires that for each knower x_i there is some truth that this knower does not know. Now let t^* be the conjunction of *all* these truths t_i over our (obviously finite) collection of knowers. Then in virtue of K_{12} no knower knows t^*. It follows that $(\exists t)(\forall x)\sim Kxt$ or equivalently $(\exists t)U(t)$. Of course, any such unknown truth will have to be a non-necessary, and thus contingent, truth, given the presently operative inferential-accessibility sense of knowledge.

A different route to the same destination is that any level 3 thesis represents what must be regarded as an unknown truth. (This will be amplified below.)

In general one cannot, of course, make the transition from $(\forall x)(\exists t)f(x, t)$ to $(\exists t)(\forall x)f(x, t)$. (Thus 'For any integer there exists another that is greater' does *not* entail 'There exists an integer that is greater than any other integer.') But in the special case of $f(x, t) = \sim Kxt$ this inference is valid, as the preceding argumentation for K_{17} shows. And a community of finite knowers is thereby subject to substantial limitations.

One might be tempted to offer the following objection to the just-indicated implication thesis: $(\forall x)(\exists t)\sim Kxt \rightarrow (\exists t)(\forall x)\sim Kxt$: 'What if one divided the realm of truth T into two disjoint parts T_1 and T_2 such that x_1 knows all (but only) T_1 truths and x_2 knows all (but only) the T_2 truths. Then clearly $(\forall x)(\exists t)\sim Kxt$ but not $(\exists t)(\forall x)\sim Kxt$.' However, this objection is flawed. For the hypothesis that it projects cannot be realized in the circumstances of our discussion, where knowledge is inferentially transmissible in that $[Kxp \,\&\, (p \rightarrow q)] \rightarrow Kxq$. Thus consider a truth $t_1 \vee t_2$ where $t_1 \in T_1$ and $t_2 \in T_2$. Then by inferential transmission this must be a known commonality for x_1 and x_2, so that the disjointness condition cannot be met. The hypothesis of truth-division runs afoul of our implicit availability construction of knowledge.

To be sure, K_{17} only assures the existence of unknown truth. To this point, we have not claimed to provide an example of this. (This awaits the discussion of level three principles in Section 8.)

What follows regarding p from $(\forall x)\sim Kxp$ or equivalently $U(p)$? Certainly not not-p. For if we had $(\forall x)\sim Kxp \rightarrow \sim p$ then it would follow that $p \rightarrow (\exists x)Kxp$ which must of course be rejected. On the other hand, poss(not-p), that is $\mathbf{P}\sim p$, must indeed be held to follow. For consider $(\forall x)\sim Kxp \rightarrow \mathbf{P}\sim p$ which is equivalent with $\mathbf{N}p \rightarrow (\exists x)Kxp$. In view of K_{14} this must be accepted on the presently operative construction of knowledge.

7 Level Two Principles and the Consideration that Knowledge of Contingent Fact is itself Contingent

The epistemic principles to which we now turn reflect the contingent facts of life regarding the ways and means of our knowledge of things. We shall continue to use the variables τ, τ', τ'', etc. to range over the limited propositional subdomain of specifically *contingent* truths, with the variables t, t', t'', etc. ranging over truths in general.)

An elemental principle of this domain is

K_{18} $(\exists \tau)(\exists x)\sim Kx\tau$ or equivalently $\sim(\forall \tau)(\forall x)Kx\tau$.
There are contingent truths that not everyone knows.

This follows from K_2 in view of the fact that necessary truths are automatically known to all (by K_9). So here we still have a level 1 principle.

Another more positive principle is:

K_{19} $(\forall x)(\exists \tau)K x\tau$ or equivalently $\sim(\exists x)(\forall \tau)\sim K x\tau$.
No knower is an utter ignoramus: every knower knows some contingent truth or other.

This principle projects K_1 into the contingent domain and is a more or less natural supposition relative to the liberal construction of knowledge we have taken into view. However, this new principle will obtain at level two; it does not follow from anything that precedes.

The following important principle obtains:

K_{20} *Wherever τ is a contingent truth, then $Kx\tau$ is also contingent: contingent truth is by nature cognitively contingent.* That is to say we have both $(\forall \tau)\mathbf{P}(\exists x)\sim K x\tau$ or equivalently $\sim(\exists \tau)\mathbf{N}(\forall x)K x\tau$ and also $(\forall \tau)\mathbf{P}(\exists x)K x\tau$ or equivalently $\sim(\exists \tau)\mathbf{N}(\forall x)\sim K x\tau$.

The first of the two components holds because its denial $(\exists \tau)\mathbf{N}(\forall x)K x\tau$ falls foul of the fact that it is only for necessary truths t that $\mathbf{N}(\forall x)Kxt$, seeing that $\mathbf{N}(\forall x)Kxp \rightarrow \mathbf{N}p$ follows from $(\exists x)Kxp \rightarrow p$. And the second follows from $(\forall t)\mathbf{P}(\exists x)Kxt$ – the potential availability of truth stipulated by K_7. And so for any specifically contingent (i.e. non-necessary) truth τ we have it that $\mathbf{P}(\exists x)\pm K x\tau$. Equivalently for no τ do we have $\mathbf{N}(\forall x)\pm K x\tau$: neither $(\forall x)K x\tau$ nor $(\forall x)\sim K x\tau$ is ever necessary in the case of contingent truths. Indeed it can be shown that even $(\exists x)\pm K x\tau$ is always contingent for contingent τ. That someone does (or does not) know a given *contingent* fact is always itself contingent.

8 Level Three Principles: Plausible Truth-Candidates

It will be recalled from the discussion of Section 4 above that any level three principle instantiates the idea of an *unknown* truth, seeing that if actual knowledge were being claimed, then the assertion in question would have to be made at a lower (deeper) level. Accordingly, the epistemic theses that will now be at issue have a standing of mere plausibility in contrast to knowability as such.

Every knower knows something. And we can actually even lay claim to a rather stronger level three principle:

K_{26} $(\exists \tau)(\forall x)K x\tau$
There are (contingent) truths that everyone knows.

K_{26} means that we cannot accept it as a principle that only necessary truths are universally known. Thus while we have endorsed its converse (as per K_{15}), we must reject: $(\forall x)Kxp \rightarrow \mathbf{N}p$. From $(\forall x)Kxp$ – and indeed even from $(\exists x)Kxp$ – we can infer that p is true, but certainly not that it is necessary.

Let us investigate the prospect of principles of the format: If $f(p)$, then $(\forall x)Kxf(p)$. Note that

$$f(p) \rightarrow (\forall x)Kxf(p)$$

holds when $f(p) = \mathbf{N}p$ in virtue of K_{16}. On the other hand, $f(p) = (\forall x)Kxp$ leads to $(\forall x)Kxp \rightarrow (\forall y)Ky(\forall x)Kxp$. And this has some claim to plausibility. For when something is obvious enough to be known to everyone, this fact itself is presumably something about which people-in-general can secure knowledge. The principle in view thus holds at level three.

It is of interest to ask what sort of knowledge follows from ignorance. Consider a thesis of the format $\sim Kxp \rightarrow Kxf(p)$. Since $\sim Kxp$ always obtains when p is false, this would mean that $Kxf(p)$ will always obtain when p is false – as $f(p)$ must therefore also do. Thus nothing of any real interest *regarding someone's knowledge follows on general principles from his ignorance of a given fact.*

9 Knowledge of the Unknown?

Consider the contention 'I know that t_0 is an unknown truth,' symbolically $Ki(t_0 \,\&\, Ut_0)$. In view of K_8 his amounts to $Kit_0 \,\&\, KiU(t_0)$. But $KiU(t_0)$ comes to $Ki\sim(\exists x)Kxt_0$. This entails $\sim(\exists x)Kxt_0$ which in turn yields $\sim Kit_0$. And this produces a contradiction. There is an instructive lesson here: *We cannot concretize* $(\exists t)Ut$ *in the mode of knowledge*: That is, we cannot instantiate this thesis by advancing a particular truth t_0 which at once and the same time we claim to know to be true and also characterize as an unknown. It is perfectly true that 'There are truths I do not know' but I cannot possibly produce any concrete examples in the mode of categorical cognition. Accordingly, the reality of it is that we can only instantiate $(\exists t)Ut$ in the mode of conjecture, which is to say at the third level of assertion.

Clearly, whenever we assert a thesis Z at the third level, so that

$$\vdash_3 Z$$

we can indeed move on to the claim that Z is true (which, after all, is why we assert it), but must nevertheless acknowledge that no one actually *knows* this to be so and accordingly must ourselves refrain from claiming actual knowledge here. For if this were known, so that $(\exists x)KxZ$ then Z would obtain at the second level of assertion: $\vdash_2 Z$. And (by hypothesis) this is not the case.

'But how can you possibly maintain something that you do not actually know to be true?' The appropriate answer, clearly, is: *cautiously and tentatively*, in the decidedly guarded and hesitant tone of voice of mere conjecture. In other words, at level three.

10 Conclusion

This survey of principles of metaknowledge has not issued in one big culminating result but rather in a diversified mosaic of smaller ones. Yet in the aggregate this complex

provides a unified overall picture of the epistemic situation from which some significant overall lessons emerge.

Perhaps the most important of these lessons is that we must operate a two-tier epistemology – one that looks not just to knowledge alone but also to the lesser level of epistemic commitment represented by plausible conjecture or supposition. Another lesson is that a systematic rational account of the cognitive situation is possible with 'knowledge' understood as in the sense of inferentially accessible information. Last but not least, we have seen that even under this most liberal and generous of constructions, our 'knowledge' is such that we must recognize the existence of a whole spectrum of cognitive limitations. For even as 'knowledge' in the mode of inferential accessibility is logically self-ampliating in that we have

If Kxp and $p \rightarrow q$, then Kxq

so also is ignorance, since we analogously have:

If $\sim Kxp$ and $q \rightarrow p$, then $\sim Kxq$

Both of these principles are two sides of the same coin. And the price we pay for the knowledge-amplification assured by the former principle is that ignorance-proliferation assured by its equivalent counterpart. Just as knowledge is self-ampliating, so is its lack.

In a world of finite beings even the most generous construction of 'knowledge' leaves ample scope for ignorance. And one of the ironic aspects of this topic of metaknowledge is that the very fact that our knowledge is limited inhibits our capacity to be specific about the matter by going on to specify just exactly what those limits are. Among the most difficult sorts of knowledge to achieve is detailed information about the nature of our ignorance.

References

Boh, Ivan (1993) *Epistemic Logic in the Middle Ages*. London: Routledge.
Hintikka, Jaakko (1962) *Knowledge and Belief*. Ithaca, NY: Cornell University Press.
Kretzmann, Norman and Stump, Eleanore (1988) *The Cambridge Translations of Medieval Philosophical Texts*, vol. 1. Cambridge: Cambridge University Press.
Prior, A. N. (1955) *Formal Logic*. Oxford: Clarendon Press.
Rescher, Nicholas (1960) The logic of belief statements. *Philosophy of Science*, 27, 88–95.
Rescher, Nicholas (1968) *Topics in Philosophical Logic*. Dordrecht: Reidel.
Rescher, Nicholas and Arnold vander Nat (1973) On alternatives in epistemic logic. *Journal of Philosophical Logic*, 2, 119–35.
Sorensen, Roy A. (1988) *Blindspots*. Oxford: Oxford University Press.

31

Deontic, Epistemic, and Temporal Modal Logics

RISTO HILPINEN

1 Modal Concepts

Modal logic is the logic of modal concepts and modal statements. Modal concepts (modalities) include the concepts of necessity, possibility, and related concepts. Modalities can be interpreted in different ways: for example, the possibility of a proposition or a state of affairs can be taken to mean that it is not ruled out by what is known (an *epistemic* interpretation) or believed (a *doxastic* interpretation), or that it is not ruled out by the accepted legal or moral requirements (a *deontic* interpretation), or that it has not always been or will not always be false (a *temporal* interpretation). These interpretations are sometimes contrasted with *alethic* modalities, which are thought to express the ways ('modes') in which a proposition can be true or false. For example, logical possibility and physical (real or substantive) possibility are alethic modalities.

The basic modal concepts are represented in systems of modal logic as propositional operators; thus they are regarded as syntactically analogous to the concept of negation and other propositional connectives. The main difference between modal operators and other connectives is that the former are not truth-functional; the truth-value (truth or falsity) of a modal sentence is not determined by the truth-values of its subsentences. The concept of possibility ('it is possible that' or 'possibly') is usually symbolized by $\Diamond$ and the concept of necessity ('it is necessary that' or 'necessarily') by $\Box$; thus the modal formula $\Diamond p$ represents the sentence form 'it is possible that p' or 'possibly p,' and $\Box p$ should be read 'it is necessary that p.' Modal operators can be defined in terms of each other: 'it is possible that p' means the same as 'it is not necessary that not-p'; thus $\Diamond p$ can be regarded as an abbreviation of $\neg\Box\neg p$, where $\neg$ is the sign of negation, and $\Box p$ is logically equivalent to $\neg\Diamond\neg p$. Systems modal propositional logic or quantification theory (predicate logic) are obtained by adding the symbols $\Diamond$ and $\Box$ (and possibly other modal signs), together with appropriate rules of sentence formation (e.g. if A is a formula, $\Diamond A$ and $\Box A$ are formulas), to a system of (non-modal) propositional logic or quantification theory.

2 The Semantics of Modalities and Systems of Modal Logic

As was observed above, modal sentences are not truth-functional: the truth-value of a modal sentence is not determined by the truth-values of its constituent. Given a true proposition p, 'it is necessary that p' may be true or false, depending on what p states (the *content* of p), and if p is false, 'possibly p' may be true or false, depending on the content of p. Consequently the logical relationships among modal propositions cannot be explained solely by means of possible truth-value assignments to simple (atomic) sentences, as in non-modal (truth-functional) propositional logic. A more complex semantics is needed. Since antiquity, modal concepts have been regarded as analogous to the quantifiers 'some' and 'all,' and modal propositions have been regarded as involving quantification over possible cases or possibilities of some kind. 'It is necessary that p' can be taken to mean that p is true (or it is true that p) no matter how things turn out to be, and 'it is possible that p' can be interpreted as saying that things may turn out to be or might have turned out to be in such a way that p is true. If the ways in which things can turn out to be are called *possible scenarios*, *situations*, or *possible worlds*, this account can be formulated as the standard possible worlds interpretation of modalities:

(CTN1) $\Box$p is true if and only if p is true in all possible worlds (situations),

and

(CTM1) $\Diamond$p is true if and only if p is true in some possible world (situation).

The possible worlds analysis of modalities goes back (at least) to the fourteenth century; for example, it seems to have been the basis of Duns Scotus's (1265–1308) modal theory (Knuuttila 1993: 143–5). G. W. Leibniz's use of the concept of possible worlds in the seventeenth century suggests a similar analysis, even though Leibniz himself did not analyze the concepts of necessity and possibility in this way. In the formal semantics of modal logic, the truth of a sentence is truth at (or relative to) a possible world, and modal formulas (sentences) are interpreted by means of a valuation function which assigns a truth-value to each sentence at each possible world. Non-modal propositional logic can be regarded as a limiting case in which only one possible world (the actual world) is considered.

In many applications of modal logic, the modal status of a given proposition depends on the situation in which it is evaluated. Many modal statements are contingent: what is possible or necessary depends on the point of evaluation. For example, what is epistemically possible for an individual depends on what the individual in question knows, and this varies from situation to situation. Thus the interpretation of modal sentences should also depend on a relation of *relative possibility* among worlds. The worlds which are possible relative to a given world (or situation) u are called the *alternatives to* u or worlds *accessible from* u. Consequently conditions (CTN1) and CTM1) should be reformulated as follows:

(CTN2) $\Box q$ is true at a world u if and only if q is true in all alternatives to u,

and

(CTM2) $\Diamond q$ is true at a world u if and only if q is true in some alternative to u.

The alternativeness relation was introduced into modal semantics in the 1950s by Marcel Guillaume (1958), Jaakko Hintikka (1957a, 1957b), Stig Kanger (1957), Saul Kripke (1963), Richard Montague (1960), and others. According to (CTN2) and (CTM2), an interpretation or a *model* of a (propositional) modal language is a triple $M = \langle W, R, V\rangle$, where $W = \{u,v,w, \ldots\}$ is a set of possible worlds (also called the *points* of the model), R is a two-place alternativeness relation defined on W, and V is an interpretation function or a *valuation function* which assigns to each sentence A a truth-value (1 for truth and 0 for falsity) at each possible world u. The pair $\langle W, R\rangle$ is called the *frame* of the model; thus a model consists of its frame and its valuation function. '$V(A,u) = 1$' (the truth of A at u in M) is expressed '$M, u \models A$,' briefly '$u \models A$;' if A is not true at u, it is false at u. (I shall use below A, B, etc., as metalogical symbols which represent arbitrary formulas of a formal language of modal logic.) A sentence is called *valid* (logically true) if and only if it is true at every world $u \in W$ for any interpretation M, and A is valid in a model M if and only if it is true at every point of the model. A sentence B is a logical consequence of A if and only if there is no interpretation M and world u such that $M, u \models A$ and not $M, u \models B$. The valuation function is subject to the usual Boolean conditions which ensure that the truth-functional compounds of simple sentences receive appropriate truth-values at each possible world, in other words:

(C$\neg$) $u \models \neg A$ if-if (if and only if) not $u \models A$,
(C&) $u \models A \,\&\, B$ if-if both $u \models A$ and $u \models B$,
(C$\vee$) $u \models A \vee B$ if-if $u \models A$ or $u \models B$ or both, and
(C$\supset$) $u \models (A \supset B) = u \models \neg A$ or $u \models B$ or both.

The truth-conditions of simple modal sentences are expressed in terms of the alternativeness relation R as follows:

(CN) $u \models \Box A$ if and only if $v \models A$ for every $v \in W$ such that $R(u,v)$,

and

(CM) $u \models \Diamond A$ if and only if $v \models A$ for some $v \in W$ such that $R(u,v)$.

This semantics validates (for example) the following modal schemata:

(K) $\Box(A \supset B) \supset (\Box A \supset \Box B)$;
(2.1) $\Box(A \,\&\, B) \supset (\Box A \,\&\, \Box B)$; (The conjunctive distributivity of $\Box$.)
(2.2) $(\Box A \,\&\, \Box B) \supset \Box(A \,\&\, B)$; (The aggregation principle for $\Box$.)
(2.3) $\Box A \supset \Box(A \vee B)$;

(2.4) $\Box(A \supset B) \supset (\Diamond A \supset \Diamond B)$;
(2.5) $\Diamond A \supset \Diamond(A \vee B)$;
(2.6) $\Diamond(A \vee B) \supset (\Diamond A \vee \Diamond B)$; (The disjunctive distributivity of $\Diamond$.)
(2.7) $\Diamond (A \& B) \supset \Diamond A$.

This system, called the system K (from Kripke), can be characterized axiomatically by a set of axioms (or axiom schemata) for propositional logic, the Modus Ponens rule, the axiom schema (K) given above, the definition

(D$\Diamond$) $\Diamond A \equiv \neg\Box\neg A$,

and the modal 'rule of necessitation'

(RN) From A, to infer $\Box A$.

Rule (RN) means that if A is provable, so is $\Box A$.

The system K involves no assumptions about the structural properties of the alternativeness relation R. Different assumptions about the properties of R lead to different extensions of K, that is, systems of modal logic including K. For example, the assumption that R is a *serial* relation, that is satisfies the condition

(CD) For every $u \in U$, $R(u,v)$ for some $v \in U$,

validates the principle that whatever is necessary is possible:

(D) $\Box A \supset \Diamond A$.

It is clear that this principle holds for most 'standard' concepts of necessity and possibility. A counterexample to this principle would be a situation in which both a proposition and its negation are necessary ($\Box A \& \Box\neg A$); most interpretations of modal expressions clearly exclude this. Alethic and epistemic modalities should also satisfy the schema

(T) $\Box A \supset A$,

which is equivalent to

$A \supset \Diamond A$.

Whatever is necessary is true, and a proposition cannot be known (to be true) in the proper sense of the word unless it is in fact true. Principle (T) distinguishes alethic and epistemic modalities from deontic and doxastic interpretations. It is true in all frames in which R is a reflexive relation, that is,

(CRefl) For every $u \in W$, $R(u,u)$.

Moreover, if R is transitive,

(4S) $\Box A \supset \Box\Box A$

is valid, and the assumption that R is symmetrical validates the schema

(B) $\Diamond\Box A \supset A$.

The schema

(E) $\Diamond A \supset \Box\Diamond A$.

holds in all symmetrical and transitive frames. By making various assumptions about R it is thus possible to generate a great variety of modal systems. There is no single 'correct' system of a modal logic, but different systems are appropriate for different purposes and applications. Modal systems can be characterized semantically by the properties of the R-relation, and syntactically by their characteristic axioms (or axiom schemata), for example:

System KD (or briefly D): K + D;
System KT (briefly, T): K + T;
System KT4 (S4): KT + 4S; and
System KT5 (S5): KT + E or KT + 4S + B.

The expressions 'S4' and 'S5' are due to C. I. Lewis, who investigated in the 1910s the concept of strict (necessary) implication, and developed five alternative axiom systems for strict implication, S1–S5 (Lewis and Langford 1932). Lewis's system S4 can be characterized semantically by means of reflexive, and transitive frames, and the semantics of Lewis's S5 can explained by means of models in which R is an equivalence relation (a reflexive, transitive, and symmetric relation). (For different systems and interpretations of modal logic, see Chellas 1980: ch. 4; van Benthem 1988; Hughes and Cresswell 1996: 23–71.)

3 Modality and Quantification

The systems characterized above are systems of propositional logic. When modal operators are added to predicate logic (quantification theory), possible worlds can serve their interpretive function only if they are thought of as having a structure of individuals, properties, and relations. Thus the models of quantified modal logic provide, for each world w, the set D(w) of individuals existing in that world, and a valuation function which assigns an *extension* (an object, a set of objects, or a relation) to each nonlogical expression at each possible world. In other words, a valuation function assigns to each nonlogical expression a function from possible worlds to extensions. Such functions are called the *intensions* of individual terms, predicates, or relational expressions.

The truth conditions of the sentences of modal quantification theory can be interpreted and formulated in different ways. For quantifiers, perhaps the most natural

choice is to let them range over the world-relative domains (rather than over all possible individuals). According to this approach, the semantic rules for $\forall$ and $\exists$ can be formulated, in a simplified and self-explanatory notation, as follows:

(CM$\forall$) M, u $\models \forall$xA(x) if and only if for all individuals d $\in$ D(u), M, u $\models$ A(d),

and

(CM$\exists$) M, u $\models \exists$xA(x) if and only if for some individual d $\in$ D(u), M, u $\models$ A(d).

The semantic rules for modalities must also be revised, and as in the case of the quantifier rules, different revisions are possible here. Perhaps the most reasonable interpretation of the necessity operator is to regard a sentence of the form $\Box$A as true at u if and only if A is true in all alternatives to u whose domain contains the individuals denoted by the individual terms in A (including those assigned to individual variables) (van Benthem 1988: 16).

The validity of various principles involving modalities and quantifiers depends on the properties of the frames, in particular, on the relationships among the domains for different worlds. Of particular interest are in this context the following operator exchange principles:

(3.1) $\Box\forall$xAx $\supset \forall$x$\Box$Ax
(3.2) $\forall$x$\Box$Ax $\supset \Box\forall$xAx
(3.3) $\Box\exists$xAx $\supset \exists$x$\Box$Ax
(3.4) $\exists$x$\Box$Ax $\supset \Box\exists$xAx

If nothing is assumed about the domains of different possible worlds, only principle (3.1) is valid, and the rest of the formulas are invalid. However, formula (3.2) (called the Barcan formula, see Barcan (1946: 2)) is valid if the following inclusion principle holds for the domains D(u),

(3.5) If R(u,w), D(w) $\subseteq$ D(u),

and principle (3.4) is valid in all frames satisfying the condition

(3.6) If R(u, w), D(u) $\subseteq$ D(w).

The acceptability of (3.1)–(3.4) depends on the interpretation of the modal operators.

Above, the antecedents of (3.1) and (3.3) and the consequents of (3.2) and (3.4) are *de dicto* propositions, which means that the modal operator is attached to a complete proposition or *dictum*. The consequents (3.1) and (3.3) and the antecedents of (3.2) and (3.4) are called modal propositions *de re*: the modal operators are attached to expressions which contain a free individual term, thus the modality in question is ascribed to the object or thing (*res*) to which the term is regarded as being applicable. Sentences (3.1)–(3.4) describe possible relationships among *de dicto* and *de re* modalities.

4 Deontic, Epistemic, and Temporal Modalities

If modal propositions are understood in terms of the possible worlds semantics, their interpretation as deontic, epistemic, or temporal propositions depends on the interpretation of possible worlds and the alternativeness relation between possible worlds. It is often interesting to consider different (kinds of) modalities simultaneously; for example, a statement of the form

If it is true that p, it is possible to know that p,

or more briefly, 'if p is true, it is knowable,' contains an alethic concept of possibility and an epistemic modality ('to know'). An analysis of such sentences requires models which represent more than one concept of necessity and possibility, with a corresponding multitude of alternativess relations. In such situations different modalities (that is, different concepts of necessity and possibility) require special symbols. **O** and **P** are often used for deontic necessity (the concept of ought or obligation) and possibility (the concept of permissibility), the expressions $\mathbf{K}_i$ and $\mathbf{P}_i$ for the concept of (propositional) knowledge ('i knows that . . .') and the associated concept of epistemic possibility ('it is possible, for all that i knows, that . . .'), and $\mathbf{B}_i$ and $\mathbf{C}_i$ for the concepts of belief and doxastic possibility. (If only one person's knowledge or beliefs are being considered, the subscript can be omitted.) It is possible to define several temporal readings of $\Box$ and $\Diamond$, for example, 'it has always been the case that' and 'it was at some time the case that,' or 'it will always be the case that' and 'it will at some time be the case that.' The pairs of operators mentioned above are interdefinable in the same way as $\Box$ and $\Diamond$. The latter symbols are usually reserved for alethic modalities. (Sometimes alethic necessity is expressed by **N** or **L** and possibility by **M**.)

5 Epistemic Logic

The study of epistemic logic, like many other areas of philosophical logic, goes back (at least) to the late scholastic philosophy. Many fourteenth-century treatises on philosophical logic included a section on the logic of knowledge, often entitled *De scire et dubitare* ('On knowing and doubting'), which discussed sophisms and paradoxes involving the concepts of knowledge, belief, and doubt (Boh 1993: ch. 4). At the beginning of the twentieth century Charles Peirce analyzed the semantics of modal notions, and proposed an epistemic interpretation of modality, according to which a proposition is possible if and only if "it is not known to be false in a given state of information." Peirce distinguished this epistemic concept of possibility from what he called "substantive possibility" (alethic possibility), and regarded modalities as quantifiers over "possible cases" or "possible states of things" (Peirce 1931–35: vol. II, paragraph 2.347; vol. V, paragraphs 5.454–455). Peirce and his scholastic predecessors regarded epistemic concepts as modal concepts, but epistemic logic was not developed in a systematic way as a branch of modal logic before Jaakko Hintikka's *Knowledge and Belief* (1962), the first book-length study of the subject.

The epistemic alternatives to a given possible world (or knowledge situation) u are the worlds (situations) not ruled out by what is known (or by what a certain person knows) at u. The concept of doxastic alternativeness is related to the concept of belief in a similar way. The most obvious logical difference between the concepts of knowledge and belief is that the former should satisfy the T-axiom,

(KT) $\mathbf{K}A \supset A$,

in other words, epistemic alternativeness relations must be reflexive, but the T-principle does not hold for the concept of belief. In this respect doxastic modalities resemble deontic modalities. The assumption that the epistemic alternativeness relation is transitive validates the principle that knowing entails knowing that one knows (the KK-thesis),

(4SK) $\mathbf{K}A \supset \mathbf{KK}A$,

This thesis has sometimes been called, sometimes misleadingly, "the positive introspection axiom" (Fagin et al. 1995: 32). The transitivity of the epistemic R-relation means that sentences of the form $\mathbf{K}p$ can be transferred from a given world to its epistemic alternatives: if $\mathbf{K}p$ holds at u, then $\mathbf{K}p$ (and not only p) holds in the epistemic alternatives to u. The acceptability of the KK-thesis is sensitive to variations in the meaning of 'know.' The thesis has been part of many philosophers' conception of knowledge since antiquity, and it has sometimes been thought to characterize a "strong" concept of knowledge (knowledge based on conclusive grounds). On the other hand, if knowledge is regarded simply as true belief, the validity of the thesis depends on the validity of the corresponding thesis about belief,

(4SB) $\mathbf{B}A \supset \mathbf{BB}A$.

This principle seems to hold at least for some varieties of belief. It helps to understand G. E. Moore's paradox of "saying and disbelieving." If is obvious that a sentence of the form

(5.1) $p \,\&\, \neg\mathbf{B}p$

is not inconsistent, but a first-person utterance of (5.1) seems inconsistent or paradoxical. If the BB-schema is valid (i.e. if the doxastic alternativeness relation is transitive), the proposition

(5.2) $\mathbf{B}(p \,\&\, \neg\mathbf{B}p)$

is inconsistent, in other words, a person cannot sincerely assert (5.1) about oneself if sincere assertion is regarded as an expression of belief (Hintikka 1962: 64–9). If knowledge is regarded as true and conclusively justified belief, the KK-thesis means that a person knows that p only if he is also justified in claiming that he knows that p, in other words, the evidence for p is epistemically conclusive only if it justifies the correspond-

ing knowledge-claim. The acceptance of this principle together with the epistemic versions of the rules and axioms of the modal system T amounts to the view that the logic of knowledge corresponds to the Lewis system S4. On the other hand, the epistemic versions of the modal axioms E and B do not seem to hold for the concept of knowledge: a person cannot be expected to be fully informed about his ignorance. The concept of belief obviously fails to satisfy principle T, but the doxastic counterpart of the principle D,

(DB) $\mathbf{B}A \supset \neg\mathbf{B}\neg A$,

should hold at least for the concept of consistent or rational belief.

The meaningfulness of quantifying into a modal context – that is, the interpretation of *de re* modal sentences – depends on the assumption that it is possible to make modal assertions about individuals (objects) independently of how they are described. For example, a sentence of the form

(5.3) $\exists x \Box Fx$

states that there is an individual which is F in all possible worlds (in which it exists). The existential quantifier identifies an individual across possible worlds or connects the 'appearances' of the same individual in different situations. Some philosophers have regarded such identifications as conceptually problematic. Epistemic modalities do not seem to be subject to such conceptual difficulties. The epistemic variant of (5.3),

(5.4) $\exists x \mathbf{K}_i Fx$,

says that some individual x is F in all situations not ruled out by (compatible with) what i knows in a given situation, in other words, someone (or something) is known (by i) to be F. This is of course quite different from saying that i knows that someone is F ($\mathbf{K}_i \exists x Fx$). The latter sentence is true but the former false in a situation in which it is known that there are spies, but their identity is unknown – it is not known who they are. In ordinary language, (5.4) can be expressed by saying that i knows who is F. In the same way, the sentence

(5.5) $\exists x \mathbf{K}_i (x = c)$

can be taken to mean that i knows who c is (Hintikka 1989: 20). Some more complex sentences involving quantifiers and epistemic operators do not have any counterparts in the standard first-order modal quantification theory. For example,

(5.6) Alma knows whom everyone admires most,

where every person may admire a different person (for example, his or her mother) cannot be represented in standard first-order epistemic logic. The representation of such sentences requires second-order epistemic logic or an independence-friendly logic in which logical operators (for example, quantifiers and epistemic operators) can be independent of each other (see Hintikka 1989: 27–8).

Systems of epistemic logic based on S4, or any K-system, contain the rule of inference

(RNK) $\vdash A / \vdash \mathbf{K}A$,

where $\vdash$ is the sign of provability, as well as the rule

(RMK) $\vdash A \supset B / \vdash \mathbf{K}A \supset \mathbf{K}B$.

The validity of these rules creates 'the problem of logical omniscience' for epistemic logic: according to the epistemic interpretations of the K-systems, an inquirer knows the logical consequences of whatever he knows, and belief systems are closed with respect to logical deduction. These results motivated Hintikka's reinterpretation of the concept of logical consistency as (logical) defensibility or "immunity to [logical] criticism" (Hintikka 1962: 31), and I. Levi's interpretation of the logic of belief as the logic of doxastic (or epistemic) commitments (rather than "active" beliefs, Levi 1997). Another way to deal with the problem of logical omniscience is to place suitable syntactic restrictions on knowledge-preserving deductive arguments (Hintikka 1989). On the semantical (model-theoretic) side, similar results can be obtained by generalizing the concept of possible scenario or situation to 'seemingly possible' scenarios, represented by so-called urn models (Rantala 1975).

In the past 30 years, epistemic logic has developed into a relatively autonomous field of research, directed at problems and applications with no counterparts in other areas of modal logic (see Fagin et al. 1995; Meyer and van der Hoek 1995). Epistemic logic has been applied in interesting ways to philosophical semantics, epistemology and the philosophy of science. For example, it forms the logical basis of the interrogative theory of inquiry in which questions are treated as requests for knowledge or epistemic imperatives (Hintikka 1976, 1999).

6 Deontic Logic

The logic of normative concepts began to be investigated as a branch of modal logic in the fourteenth century, when some scholastic philosophers observed the analogies between deontic and alethic modalities, and studied the deontic (normative) interpretations of various laws of modal logic (Knuuttila 1993: ch. 5). In the seventeenth century, G. W. Leibniz (1930) called the deontic categories of the obligatory, the permitted, and the prohibited "legal modalities" ("Iuris modalia"), and observed that the basic principles of modal logic hold for the legal modalities. In fact, Leibniz suggested that deontic modalities can be analyzed in terms of the alethic modalities: he suggested that the permitted (*licitum*) is "what is possible for a good man to do," and the obligatory (*debitum*) "what is necessary for a good man to do." In the twentieth century the study of deontic logic as a branch of modal logic was initiated by Georg Henrik von Wright's pioneering work in the early 1950s (1951a, 1951b).

A simple system of deontic logic can be obtained by reading Leibniz's definition of the concept of obligation (ought) as

(O.Lbnz1) A is obligatory for b if and only if A is necessary for b's being a good person,

that is,

$$\mathbf{O}_b A \text{ if and only if } \Box(G(b) \supset A),$$

where **O** is the deontic counterpart of $\Box$ and 'G(b)' means that b is 'good' (in the sense intended by Leibniz). If the explicit reference to an agent is deleted, we obtain the definition:

(O.Lbnz2) $\mathbf{O}A \equiv \Box(G \supset A)$.

The corresponding Leibnizian concept of permission is expressed by

(P.Lbnz2) $\mathbf{P}A \equiv \Diamond(G \,\&\, A)$.

These schemata can be regarded as partial reductions of deontic logic to alethic modal logic. In the twentieth-century deontic logic, the Leibnizian analysis of the concepts of obligation and permission was rediscovered by the Swedish philosopher Stig Kanger in 1950. Kanger (1971: 53) interpreted the constant G as "what morality prescribes." According to Kanger, **O**A (it ought to be the case that A) means that A follows from the requirements of morality.

If the alethic $\Box$-operator satisfies the axioms and the rules of inference of the modal system called KT (see above), the ought-operator defined by (O.Lbnz2) satisfies the deontic K-principle

(KD) $\mathbf{O}(A \supset B) \supset (\mathbf{O}A \supset \mathbf{O}B)$

and the rule of 'deontic necessitation'

(RND) $\vdash A \;/\; \vdash \mathbf{O}A$.

The additional assumption that being good is possible,

(DG) $\Diamond G$,

yields the deontic D-schema (the principle of deontic consistency),

(DD) $\mathbf{O}A \supset \neg\mathbf{O}\neg A$.

The system of (propositional) deontic logic obtained by adding to propositional logic the axiom schemata KD and DD and the rule RND is usually called the "standard system of deontic logic," abbreviated "SDL" (Føllesdal and Hilpinen 1971: 13–15). The theorems and the (derived) rules of inference of the standard system include the deontic variants of the schemata (1)–(7) and the rule

(RMD) $\vdash A \supset B \;/\; \vdash \mathbf{O}A \supset \mathbf{O}B$.

This modal system is often called the system KD or simply D (Chellas 1980: 114).

The sentences of SDL can be interpreted in terms of possible worlds (or world states) and an alternativeness relation between possible worlds in the same way as other modalities. The deontic alternatives to a given world u are worlds (or situations) in which everything that is obligatory at u is the case; thus the worlds related to u by R may be termed *deontically perfect* or *ideal* worlds (relative to u); they are worlds in which all obligations are fulfilled. If possible worlds are regarded as possible courses of events or histories which are partly constituted by an agent's actions, the semantics of SDL divides such histories into deontically acceptable and deontically unacceptable histories. An action is permitted if and only if it is part of some deontically acceptable course of events or if there is some deontically acceptable way of performing the action, and an action is obligatory if and only if no course of events is acceptable unless it exemplifies the action in question. The set of acceptable courses of action (relative to a given action situation) may be termed the *field of permissibility* (Lewis 1979). According to the deontic consistency principle (DD), the field of permissibility is never empty: some action is permissible in any situation. Additional structural assumptions about the R-relation validate further deontic principles. It is clear that sentences of the form

(6.1) $\mathbf{O}p \supset p$

are not logical truths, and therefore R cannot be regarded as a reflexive relation. However, the schema

(6.2) $\mathbf{O}(\mathbf{O}A \supset A)$

seems to hold for the concept of ought (or the concept of obligation): it ought to be the case that whatever ought to be the case is the case. The validity of (6.1) follows from the assumption that the deontic alternativeness relation is secondarily reflexive, in other words,

(C.OO) If R(u,v) for some u, then R(v,v).

SDL is quite a simple system, and cannot do justice to many complexities of normative discourse. This has been shown by various 'paradoxes' which result from attempts to formalize complex normative statements by means of SDL. (For discussions of the paradoxes of deontic logic, see Føllesdal and Hilpinen (1971: 21–6), and the articles in Hilpinen 1981). For example, SDL does not suffice for the representation of many *conditional* norms – and conditional norms abound in normative discourse. The following example about the inadequacy of SDL is analogous to an example given by Chisholm (1963); a situation of this kind is sometimes called 'Chisholm's paradox':

(Ch1) Bertie ought to confess.
(Ch2) Bertie ought to warn Corky if he is going to confess.
(Ch3) If Bertie does not confess, he ought not to warn Corky.
(Ch4) Bertie does not confess.

(Ch1)–(Ch4) seem to form a consistent set of (logically) mutually independent sentences, but in SDL they cannot be represented as such. If (Ch2) is represented as having the form

(6.3) $O(s \supset r)$,

where 's' is taken to mean that Bertie confesses and 'r' means that Bertie warns Corky, (Ch1) and (Ch2) entail

(6.4) Or

If (Ch3) is regarded as having the same form as (Ch2), that is,

(6.5) $O(\neg s \supset \neg r)$,

it is (in SDL) a logical consequence of (Ch1), and if it represented as

(6.6) $\neg s \supset O\neg r$,

(Ch3) and (Ch4) entail

(6.7) $O\neg r$,

which, according to SDL, is inconsistent with (6.4); thus the choice of (6.6) as the representation of (Ch3) would make the set (Ch1)–(Ch4) inconsistent. On the other hand, if (Ch2) is formalized as

(6.8) $s \supset Or$,

it is a logical consequence of (Ch4), which is also unacceptable.

Sentence (Ch3) tells what Bertie ought to do in a situation where he has failed to fulfill his obligation to confess; thus it can be said to express a *contrary-to-duty* obligation (abbreviated 'CTD'): Chisholm's paradox may also be called the paradox of contrary-to-duty obligation.

7 Temporal Frames

Some authors have proposed to avoid the inconsistency of between (6.4) and (6.7) by relativizing the concept of obligation (or the concept of ought) time: it has been suggested that (6.4) and (6.7) hold at different points of time (Åqvist and Hoepelman 1981). It is obvious that what is obligatory or permitted changes over time; thus it is natural to assume, quite independently of the paradox of contrary-to-duty obligation, that deontic concepts should be analyzed by means of temporally structured systems of possible worlds, and that deontic logic should be based on tense logic (Thomason 1981, 1984; Horty 2001). The temporal structures required for the semantics of

deontic modalities should involve a set W of world states or situations and a partial ordering $<$ on W such that for any u, v, w $\in$ W, if $u < w$ and $v < w$, then $u < v$ or $v < u$ or $u = w$. The relation $<$ represents the temporal precedence among world-states. According to the $<$-relation, time has a branching, tree-like structure: each world-state has a unique past, but several possible futures. Temporal frames of this kind can also be used in epistemic logic for the representation of epistemic and doxastic changes. Maximal sets of linearly ordered world-states from W are called *histories* through the tree $T = (W, <)$; a set S is linearly ordered whenever for any u, v, w $\in$ S, either $u < v$ or $v < u$ or $u = v$. Let H(u) be the set of histories that pass through u. The histories in H(u) represent the possibilities open in (or accessible from) the situation u. The truth-conditions of modal sentences can be defined for world-history pairs u/h such that h $\in$ H(u) (for details, see Thomason 1984; Horty 2001: ch. 2). For example, a temporal necessity operator $\Box$ and a future tense operator **F** can be defined in this framework as follows:

(CNtemp) M, u/h $\models \Box$A iff M,u/g $\models$ A for every g $\in$ H(u);
(CFtemp) M, u/h $\models$ **F**A iff M,v/h $\models$ A for some v such that $u < v$.

According to (CNtemp), it is clear that if there is an h $\in$ H(u) such that $\Box$p holds at u/h, $\Box$p holds at u/g for any history g $\in$ H(u); thus alethic modal sentences are *determinately* true or false at (temporary) world states or situations. The truth of $\Box$p at u can be taken to mean the truth of p is settled or fixed at u, or that p is "settled true" at u (Horty 2001: 10). The deontic alternativeness relation R may be construed as a relation between a situation u and a history g $\in$ H(u): R(u,g) can be taken to mean that g is one of the deontically preferred or deontically acceptable histories passing through u. Relative to each situation u, the field of permissibility consists of the acceptable histories in H(u). The truth-conditions of O-sentences can be defined as follows:

(COtemp) M,u/h $\models$ **O**A iff M,u/g $\models$ A for every g such that R(u,g).

According to (COtemp), p is obligatory in a given situation u if and only if p holds in every deontically acceptable history in H(u). Like alethic sentences, deontic sentences are determinately true or false at each u $\in$ W. In interesting cases (e.g. in Chisholm-type examples) the proposition in the scope of **O** is not determinately true or false at the situation of evaluation, but refers to the future, for example to the options available to the agent (see Åqvist and Hoepelman 1981: 192). In the above example, (Ch1) (i.e. **O**s) and (6.4) hold as long as confessing is one of the options available to Bertie, but as soon this option is excluded and it is 'settled' that Bertie is not going to confess, (6.7) is true.

8 Conditional Obligations and Rules of Detachment

There are also non-temporal versions of the CTD-paradox. For example, consider the following example (due to Prakken and Sergot 1997): Assume that dogs are not permitted in a certain village, but if anyone has a dog, there ought to be a warning sign

about it in front of the owner's house. Moreover, warning signs ought not to be posted without sufficient reason; thus there ought to be no warning sign if there is no dog. This example is formally analogous to Chisholm's example, and an attempt to formalize it in SDL leads to a similar inconsistency (Prakken and Sergot 1997; Carmo and Jones 2000).

The deduction of a contradiction from (6.4) and (6.7) depends on the principle of normative consistency (DD),

$$\mathbf{O}A \supset \neg\mathbf{O}\neg A.$$

This principle can be criticized independently of Chisholm's example: (DD) excludes the possibility of normative conflicts, but such conflicts are not unusual in morality and law, and it may be argued that they do not amount to paradoxes (Chellas 1974: 24). If the consistency principle is rejected, the deontic version of the aggregation principle (2.2),

$$\mathbf{O}A \,\&\, \mathbf{O}B \supset O(A \,\&\, B),$$

should be rejected as well, because the latter principle undermines the distinction between a conflict between obligations and the existence of a self-contradictory obligation. Normative conflicts can be distinguished from self-contradictory (impossible) obligations. Thus logicians have developed systems of deontic logic in which (DD) and the aggregation principle do not hold (Chellas 1980: 201–10, 272–5). Such systems represent CTD-situations as involving conflicting obligations, but they do not offer any analysis of CTD-obligations and their relationship to the 'primary' obligations.

As was observed above, the semantics of SDL is based on a division of worlds or situations into acceptable (deontically perfect) and unacceptable worlds, and the O-sentences describe how things are in the deontically faultless worlds. But sentence (Ch3) does not tell how things are in a deontically faultless world; it tells what the agent (Bertie) ought to do under imperfect conditions, that is, in situations in which Bertie does not act in accordance with his obligations. The situation could be described by saying that among the (less than ideal) scenarios where Bertie does not fulfill his obligation to confess, those in which he does not warn Corky are deontically preferable to the circumstances in which he (falsely) warns her. Thus Chisholm's example requires a distinction between different degrees of deontic perfection. (Ch2) can be taken to mean that in deontically perfect circumstances where Berie confesses, he warns Corky, and (Ch3) says that in the best worlds where he does not confess, he does not warn her (Hansson 1969). Let us express these conditional obligations by

$$(8.1) \quad O(r \,/\, s)$$

and

$$(8.2) \quad O(\neg r \,/\, \neg s),$$

respectively. Let us call the worlds where p is true, 'p-words,' and let the p-worlds which are normatively least objectionable relative to a given situation u be called deontically

optimal p-worlds relative to u. The concept of a deontically optimal p-world is a generalization of the concept of a deontically perfect world of SDL, and the assumption that for any consistent proposition p, there is a nonempty set of deontically optimal p-worlds, is a generalization of the SDL principle that any world has a nonempty set of deontic alternatives. The truth of a conditional ought-statement O(q / p) at u can be taken to mean that q is true in all deontically optimal p-worlds (relative to u). According to this interpretation of conditional obligations, the principle of 'deontic detachment,'

(DDet) $\mathbf{O}(B / A) \supset (OA \supset OB)$,

is a valid principle for conditional obligations, but the principle of 'factual detachment,'

(FDet) $\mathbf{O}(B / A) \supset (A \supset OB)$,

does not hold. If (Ch2) and (Ch3) are interpreted in this way, (Ch1)–(Ch4) do not lead to a contradiction: (Ch1) and (Ch2) entail the obligation **O**r, but (Ch3) and (Ch4) do not entail **O**–r.

Chisholm's paradox can also be avoided by replacing the truth-functional conditional in (6.6) and (6.8) by an intensional (subjunctive) conditional without introducing a special concept of conditional obligation (Mott 1973). In the representation of our example in SDL, the logical asymmetry between (6.3) and (6.6) is required by the assumption of the logical independence of (Ch1)–(Ch4), and this leads to the inconsistency (6.4)–(6.7). If the two conditionals are expressed as intensional conditionals, this problem does not arise. An intensional conditional (e.g. a subjunctive conditional) 'q if p' can be regarded as true in a situation u if and only if q is true in all possible worlds (situations) in which p is true but which resemble u in other respects as much as possible (Lewis 1973). The truth of such a conditional is not a consequence of the falsity of p (or of the truth of q).

If 'q if p' is symbolized 'p > q,' and (Ch2) and (Ch3) are represented (respectively) by

(8.3) $s > Or$

and

(8.4) $-s > O-r$,

no contradiction will arise. If the *modus ponens* rule (the rule of factual detachment) holds for the conditional connective, (Ch3) and (Ch4) entail (6.7), but (Ch1) and (Ch2) do not entail (6.4). The former analysis of conditional obligations leads in our example to the result that Bertie ought to warn Corky, but the second analysis gives the result that Bertie ought not to warn Corky. Thus the two analyses involve two different concepts of ought (or 'obligation'): the first interpretation of (Ch1)–(Ch3) takes the statements in question as expressions of *prima facie*, defeasible (ideal or sub-ideal) obligations: (Ch1)–(Ch2) can be regarded as saying that in so far as Bertie ought to confess, he ought to warn Corky. On the other hand, if he is in fact not going to confess (or if this is regarded as being *settled*), he has an actual or practical ('all-out') obligation not

to warn Corky; the second analysis concerns obligations of the latter type. The dependence of the latter type of obligation (ought) on the former presents an interesting problem for deontic logic and the theory of practical reasoning (Loewer and Belzer 1983). The paradoxes of conditional obligation and attempts to represent various CTD-obligations and other conditional obligations in formal systems of deontic logic have generated an extensive literature on the subject. (See Carmo and Jones 2000 and the articles in Nute 1997.)

As was observed above, deontic propositions are often future oriented and relative to time. This depends on another distinctive feature of deontic concepts, namely, that they are usually applied to acts, and acts normally involve change and take place in time. Philosophers and logicians have represented the concept of action in deontic logic in different ways (Hilpinen 1993, 1997). First, deontic modalities have been combined with action modalities, represented by modal operators which can be read 'i brings it about that p' or 'i sees to it that p' (Belnap 1991; Horty 2001). Another approach is to make a distinction between propositions, represented by propositional symbols, and actions, represented by action terms (action descriptions), and construe deontic concepts as operators which turn action terms into deontic propositions. The latter approach has been adopted in dynamic deontic logic (Segerberg 1982). Both approaches are based on temporal models involving temporally ordered world-states. Like epistemic logic, deontic logic has developed during the past 20–30 years into an autonomous discipline, with applications to computer science, legal informatics, moral philosophy, and other fields (see the papers in McNamara and Prakken 1999).

References

Åqvist, L. and Hoepelman, J. (1981) Some theorems about a "tree" system of deontic tense logic. In Hilpinen (1981), 187–221.

Barcan (Barcan-Marcus), R. C. (1946) A functional calculus of first order based on strict implication. *The Journal of Symbolic Logic* 11, 1–16.

Belnap, N. (1991) Backwards and forwards in the modal logic of agency. *Philosophy and Phenomenological Research*, 51, 8–37.

van Benthem, J. (1988) *A Manual of Intensional Logic*, 2nd edn. Menlo Park, CA: CSLI/SRI International.

Boh, I. (1993) *Epistemic Logic in the Later Middle Ages*. London and New York: Routledge.

Carmo, J. and Jones, A. J. I. (2000) Deontic logic and contrary-to-duties. In D. Gabbay (ed.) *Handbook of Philosophical Logic*, rev edn, vol. 4. Dordrecht: Kluwer Academic Publishers.

Chellas, B. (1974) Conditional obligation. In S. Stenlund (ed.), *Logical Theory and Semantic Analysis: Essays Dedicated to Stig Kanger on his Fiftieth Birthday*. 23–33. Dordrecht: D. Reidel.

Chellas, B. (1980) *Modal Logic: An Introduction*. Cambridge: Cambridge University Press.

Chisholm, R. M. (1963) Contrary-to-duty imperatives and deontic logic. *Analysis*, 24, 33–36.

Fagin, R., Halpern, J., Moses, Y. and Vardi, M. (1995) *Reasoning about Knowledge*. Cambridge, MA: MIT Press.

Føllesdal, D. and Hilpinen, R. (1971) Deontic logic: an introduction. In Hilpinen (1971), 1–35.

Guillaume, M. (1958) Rapports entre calculs propositionnels modaux et topologie impliqués par certaines extensions de la méthode des tableaux sémantiques. *Comptes rendus hebdomaires des séances de l'Academie des Sciences (Paris)*, 246, 1140–42, 2207–2210; 247, 1282–83. Paris: Gauthiers-Villars.

Hansson, B. (1971) An analysis of some deontic logics. *Noûs* 3 (1969), 373–98. Reprinted in Hilpinen (1971), 121–47.

Hilpinen, R. (ed.) (1971) *Deontic Logic: Introductory and Systematic Readings*. Dordrecht: Reidel.

Hilpinen, R. (ed.) (1981) *New Studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*. Dordrecht: Reidel.

Hilpinen, R. (1993) Actions in deontic logic. In J.-J. Ch. Meyer and R. J. Wieringa (eds.), *Deontic Logic in Computer Science: Normative System Specification* (pp. 85–100). Chichester and New York: John Wiley & Sons.

Hilpinen, R. (1997) On states, actions, omissions, and norms. In G. Holmström-Hintikka and R. Tuomela (eds.), *Contemporary Action Theory*, vol. I: *Individual Action* (pp. 83–108). Dordrecht: Kluwer Academic Publishers.

Hintikka, J. (1957a) Quantifiers in deontic logic. *Societas Scientiarum Fennica, Commentationes Humanarum Litterarum*, 23:4 (Helsinki).

Hintikka, J. (1957b) Modality as referential multiplicity. *Ajatus*, 20, 49–64.

Hintikka, J. (1962) *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.

Hintikka, J. (1989) Reasoning about knowledge in philosophy: the paradigm of epistemic logic. In J. and M. Hintikka (eds.), *The Logic of Epistemology and the Epistemology of Logic* (pp. 17–35). Dordrecht and Boston: Kluwer Academic Publishers. Reprinted from *Reasoning about Knowledge: Proceedings of the 1986 Conference*. Los Altos, CA: Morgan Kaufmann, 1986, 63–8.

Hintikka, J. (1976) *The Semantics of Questions and the Questions of Semantics, Acta Philosophica Fennica*, 28:4 Helsinki: Societas Philosophica Fennica.

Hintikka, J. (1999) *Inquiry as Inquiry: A Logic of Scientific Discovery*. Dordrecht: Kluwer Academic Publishers.

Horty, J. (2001) *Agency and Deontic Logic*. Oxford and New York: Oxford University Press.

Hughes, G. E. and Cresswell, M. J. (1996) *A New Introduction to Modal Logic*. London and New York: Routledge.

Kanger, S. (1957) *Provability in Logic* (*Stockholm Studies in Philosophy* 1). Stockholm: Almqvist & Wiksell.

Kanger, S. (1971) New Foundations for Ethical Theory. In Hilpinen (1971), 36–58 (originally published in monographic form in 1957).

Knuuttila, S. (1993) *Modalities in Medieval Philosophy*. London and New York: Routledge.

Kripke, S. (1963) Semantical analysis of modal logic I. Normal modal propositonal calculi. *Zeitschrift für Mathematische Logic und die Grundlagen der Mathematik*, 9, 67–96.

Leibniz, G. W. (1930) *Elementa iuris naturalis*, in G. W. Leibniz, *Sämtliche Schriften und Briefe. Sechste Reihe: Philosophische Schriften* (pp. 431–85). Vol. 1. Darmstadt: Otto Reichl Verlag.

Levi, I. (1997) *The Covenant of Reason: Rationality and the Commitments of Thought*. Cambridge: Cambridge University Press.

Lewis, C. I. and Langford, C. H. (1932) *Symbolic Logic*. New York: Dover Publications.

Lewis, D. K. (1973) *Counterfactuals*. Oxford: Basil Blackwell.

Lewis, D. K. (1979) A problem about permission. In E. Saarinen et al. (eds.), *Essays in Honour of Jaakko Hintikka on the Occasion of His Fiftieth Birthday* (pp. 163–75). Dordrecht: Reidel.

Loewer, B. and Belzer, M. (1983) Dyadic deontic detachment. *Synthese*, 54, 295–318.

McNamara, P. and Prakken, H. (eds.) (1999) *Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science*. Amsterdam: IOS Press.

Meyer, J. J. Ch. and van der Hoek, W. (1995) *Epistemic Logic for AI and Computer Science*. Cambridge: Cambridge University Press.

Montague, R. (1960) Logical necessity, physical necessity, ethics, and quantifiers. *Inquiry*, 4, 259–69. Reprinted in Montague, R. (1974), *Formal Philosophy*, R. H. Thomason (ed.). New Haven, CT: Yale University Press.

Mott, P. L. (1973) On Chisholm's paradox. *Journal of Philosophical Logic*, 2, 197–211.

Nute, D. (ed.) (1997) *Defeasible Deontic Logic*. Dordrecht and Boston: Kluwer Aacademic Publishers.

Peirce, C. S. (1931–35) *Collected Papers of Charles Sanders Peirce*, vols. I–VI, ed. C. Hartshorne and P. Weiss. Cambridge, MA: Harvard University Press.

Prakken, H. and Sergot, M. (1997) Deontic logic and contrary-to-duty obligations. In D. Nute (ed.), *Defeasible Deontic Logic* (pp. 223–62). Dordrecht: Kluwer Academic Publishers.

Rantala, V. (1975) Urn models: a new kind of non-standard model for first-order logic. *Journal of Philosophical Logic*, 4, 455–74.

Segerberg, K. (1982) A deontic logic of action. *Studia Logica*, 41, 269–82.

Thomason, R. (1981) Deontic logic founded on tense logic. In Hilpinen (1981), 165–76.

Thomason, R. (1984) 'Combinations of tense and modality. In D. Gabbay and F. Guenther (eds.), *Handbook of Philosophical Logic*, vol. II (pp. 135–65). Dordrecht: Reidel.

von Wright, G. H. (1951a) Deontic logic. *Mind*, 60, 1–15. Reprinted in G. H. von Wright, *Logical Studies*. London: Routledge and Kegan Paul, 1957, 58–74.

von Wright, G. H. (1951b) *An Essay in Modal Logic*. Amsterdam: North-Holland.

This page intentionally left blank

Part X

INTUITIONISTIC, FREE, AND MANY-VALUED LOGICS

This page intentionally left blank

32

Intuitionism

DIRK VAN DALEN AND MARK VAN ATTEN

We will view intuitionism as a philosophical program in mathematics. It was founded as such by the Dutch mathematician and philosopher L. E. J. Brouwer (1881–1966) (van Dalen 1999a). The main reference for the technical results discussed here is (Troelstra and van Dalen 1988); the original texts by Brouwer can be found in (Brouwer 1975); additional translations, and texts of other authors mentioned below, are in (van Heijenoort 1967), (Mancosu 1998), and (Gödel 1990–95).

1 Logic: The Proof Interpretation

Various arguments for intuitionistic logic have been propounded, for example by Brouwer (from the nature of mathematics), Heyting (from a concern for ontological neutrality), Dummett (from considerations on meaning) and Martin-Löf (from type theory). These are all different arguments but they lead to the same logic. We focus on Brouwer's motivation.

Brouwer thinks of mathematics first of all as an activity rather than a theory. One constructs objects in one's mind. Mathematical truth, therefore, does not consist in correspondence to an independent reality, but in the fact that a construction has been (or could be) carried out. An intuitionist accounts for the truth of $2 + 2 = 4$ by saying that if one constructs 2, constructs 2 again, and compares the overall result to a construction of 4, one sees they are the same. This construction not only establishes the truth of the proposition $2 + 2 = 4$, but is all there is to its truth.

In Brouwer's view, logic depends on mathematics and not vice versa. Logic notes and studies regularities that one observes in the mathematical construction process. For example, the logical notion of negation derives from seeing that some mathematical constructions do not go through. One constructs $2 + 3$ and sees that the outcome does not match with a construction of 4; hence $\neg(2 + 3 = 4)$.

This suggests that the construction criterion for mathematical truth also yields an interpretation of the logical connectives. We will now elaborate on this. Let us write '$a : A$' for 'a is a construction that establishes A' and call this a a *proof* of A.

A proof of $\neg A$ has to tell us that A cannot have a proof; hence we read $p : \neg A$ as 'Each proof a of A can be converted by the construction p into a proof of an absurdity (say, $0 = 1$; abbreviated $\perp$)'.

To extend this *proof interpretation* to the other connectives, it is convenient to have the following notation. (a, b) denotes the pairing of constructions, and $(c)_0$, $(c)_1$ are the first and second projections of c.

A proof of a conjunction $A \wedge B$ is a pair (a, b) of proofs such that $a : A$ and $b : B$.

Interpreting the connectives in terms of proofs means that, unlike classical logic, the disjunction has to be *effective*, one must specify for which of the disjuncts one has a proof. A proof of a disjunction $A \vee B$ is a pair (p, q) such that p carries the information which disjunct is shown correct by this proof, and q is the proof of that disjunct. We stipulate that $p \in \{0, 1\}$. So if we have $(p, q) : A \vee B$ then either $p = 0$ and $q : A$, or $p = 1$ and $q : B$.

The most interesting propositional connective is the implication. Classically, $A \to B$ is true if A is false or B is true, but this cannot be used now as it involves the classical disjunction. Moreover, it assumes that the truth values of A and B are known before one can settle the status of $A \to B$.

Heyting showed that this is asking too much. Consider A = 'there occur twenty 7's in the decimal expansion of π,' and B = 'there occur nineteen 7's in the decimal expansion of π.' $\neg A \vee B$ does not hold constructively, but in the proof interpretation, $A \to B$ is obviously correct.

It is because, if we could show the correctness of A, then a simple construction would allow us to show the correctness of B as well. Implication, then, is interpreted in terms of possible proofs: $p : A \to B$ if p transforms each possible proof $q : A$ into a proof $p(q) : B$.

The meaning of the quantifiers is specified along the same lines. Let us assume that we are dealing with a domain D of mathematical objects. A proof p of $\forall x A(x)$ is a construction which yields for every object $d \in D$ a proof $p(d) : A(d)$. A proof p of $\exists x A(x)$ is a pair (p_0, p_1) such that $p_1 : A(p_0)$. Again, note the demand of effectiviness: the proof of an existential statement requires an instance plus a proof of this instance.

The interpretation of the connectives in terms of proofs was made explicit by Heyting (1934). Around the same time, Kolmogorov gave an interpretation in terms of *problems* and *solutions*. The two are essentially the same. Note that in its dependence on the abstract concept of proof, Heyting's interpretation goes well beyond finitism (see 'The Dialectica interpretation,' below).

Here are some examples of the proof interpretation.

1. $(A \vee B) \to (B \vee A)$ Let $p : A \vee B$, then $(p)_0 = 0$ and $(p)_1 : A$, or $(p)_0 = 1$ and $(p)_1 : B$. By interchanging A and B we get, looking for $q : B \vee A$, $(q)_0 = 1$ and $(q)_1 : B$, or $(q)_0 = 0$ and $(q)_1 : A$. This comes to $\overline{\mathrm{sg}}((p)_0) = (q)_0$ and $(p)_1 : B$, or $\overline{\mathrm{sg}}((p)_0) = (q)_0$ and $(q)_1 : A$, that is, $(\overline{\mathrm{sg}}((p)_0), (p)_1) : B \vee A$. And so $\lambda p.(\overline{\mathrm{sg}}((p)_0), (p)_1) : A \vee B \to B \vee A$.
2. $A \vee \neg A$ $p : A \vee \neg A \Leftrightarrow (p)_0 = 0$ and $(p)_1 : A$ or $(p)_0 = 1$ and $(p)_1 : \neg A$.
   However, for an arbitrary proposition A we do not know whether A or $\neg A$ has a proof, and hence $(p)_0$ cannot be computed. So, in general there is no proof of $A \vee \neg A$.
3. $\neg\exists x A(x) \to \forall x \neg A(x)$
   $p : \neg\exists x A(x) \Leftrightarrow p(a) : \bot$ for a proof $a : \exists x A(x)$
   We have to find a q such that $q : \forall x \neg A(x)$, i.e. $q(d) : A(d) \to \bot$ for any $d \in D$. So pick an element d and let $r : A(d)$, then $(d, r) : \exists x A(x)$ and so $p((d, r)) : \bot$. Therefore we put $q(d)(r) = p((d, r))$, so $q = \lambda r.\lambda d.p((d, r))$ and hence
   $\lambda p.\lambda r.\lambda d.p((d, r)) : \neg\exists x A(x) \to \forall x \neg A(x)$.

Brouwer employed a characteristic technique now known as 'Brouwerian (weak) counterexamples' to show that certain classical statements are constructively untenable by reducing them to unproven statements. To illustrate, here is a Brouwerian counterexample to the classical trichotomy law $\forall x \in \mathbb{R}(x < 0 \vee x = 0 \vee x > 0)$.

We compute simultaneously the decimal expansion of π and a Cauchy sequence to be specified. We use $N(k)$ as an abbreviation for 'the decimals $p_{k-89}, \ldots, p_k$ of π are all 9.' Now we define

$$a_n = \begin{cases} (-2)^{-n} & \text{if } \forall k \leq n \neg N(k) \\ (-2)^{-k} & \text{if } k \leq n \text{ and } N(k) \end{cases}$$

a_n starts as an oscillating sequence of negative powers of -2. Should we hit upon a sequence of 90 nines in the expansion of π, a_n becomes constant from there on:

$$1, -1/2, 1/4, -1/8, \ldots, (-2)^{-k}, (-2)^{-k}, (-2)^{-k}, \ldots$$

The sequence a_n satisfies the Cauchy condition and in that sense determines a real number a. The sequence is well defined, and, in principle, for each n we can check $N(n)$.

But of this a we cannot say whether it is positive, negative, or zero:

$a > 0 \Leftrightarrow N(k)$ holds the first time for an even number
$a < 0 \Leftrightarrow N(k)$ holds the first time for an odd number
$a = 0 \Leftrightarrow N(k)$ holds for no k.

Since we as yet have no construction that determines whether $N(k)$s occur, we cannot affirm $a < 0 \vee a = 0 \vee a > 0$ and hence the trichotomy law cannot be said to have a proof.

Moreover, the number a cannot be irrational, for then $N(k)$ would never apply, and hence $a = 0$, contradiction. This shows that $\neg\neg(a$ is rational. On the other hand, there is no proof that a is rational, so $\neg\neg A \rightarrow A$ fails. Similarly, $a = 0 \vee a \neq 0$ has no proof.

This type of counterexample is called *weak* because they show that some proposition has no proof yet, but it does not at all exclude that such a proof will be found later. (A sequence that Brouwer employed in his own writings is 01234567890 in the expansion of π; but its occurence has now been proved.)

Strong counterexamples cannot always be expected. There are, for example, instances of the Principle of the Excluded Middle (PEM) that have no proof (in any case, not yet), but the negation of PEM cannot be proved! $\neg(A \vee \neg A)$ is equivalent to $\neg A \wedge \neg\neg A$, which is a contradiction. However, strong counterexamples to some other classical principles do exist, and some will be shown in next section.

Although Brouwer had little interest in developing logic for its own sake, some of the finer distinctions that are common today were introduced by him. In his 1907 thesis one can already find the explicit and fully understood notions of language, logic, metalanguage, metalogic, etc. Also, Brouwer was the first to prove a non-trivial result in intuitionistic logic, $\neg A \leftrightarrow \neg\neg\neg A$ (1923). He discussed logic in an informal manner;

Kolmogorov (1925) and Glivenko (1929) then presented formalizations of parts of intuitionistic logic. A full system was given by Heyting (1930). As such it has become a part of mathematical logic in its own right, independent of philosophical motivations. Also, semantics other than the proof interpretation were developed that allow for sharper technical results (see 'Further semantics', below).

Gödel (1933) defined a translation ° given by

$A^\circ = \neg\neg A$ for atomic A
$(A \wedge B)^\circ = A^\circ \wedge B^\circ$
$(A \vee B)^\circ = A^\circ \vee B^\circ$
$(A \to B)^\circ = A^\circ \to B^\circ$
$(\forall x A(x))^\circ = \forall x A^\circ(x)$
$(\exists x A(x))^\circ = \neg\forall x \neg A^\circ(x)$

and proved that in predicate logic we have

$$\Gamma \vdash_c A \Leftrightarrow \Gamma^\circ \vdash_i A^\circ$$

where $\Gamma^\circ = \{B \mid B \in \Gamma\}$, and $\vdash_c$ and $\vdash_i$ denote classical and intuitionistic derivation relations, respectively.

Classically, a sentence A and its translation translation A° are equivalent, $\vdash_c A \leftrightarrow A^\circ$; from an intuitionistic point of view, however, disjunctions and existential statements will be weakened by the translation. Still, Gödel's result shows that, formally, classical predicate logic can be embedded into intuitionistic predicate logic.

Taking $A = \bot$ and noting that $\bot^\circ = \bot$, it follows that classical predicate logic is consistent if and only if intuitionistic predicate logic is; so the philosophical advantages of intuitionistic over classical predicate logic must lie in its interpretation and not in its trustworthiness.

In fact, Gödel proved something stronger. Classical arithmetic (**PA**, i.e. Peano's axioms with classical logic as the underlying logic) can be embedded into intuitionistic arithmetic (**HA**, i.e. Peano's axioms with Heyting's formalized intuitionistic logic as the underlying logic):

$$\mathbf{PA} \vdash_c A \Leftrightarrow \mathbf{HA} \vdash_i A^\circ$$

In particular,

$$\begin{aligned} \mathbf{PA} \vdash_c 0 = 1 &\Leftrightarrow \mathbf{HA} \vdash_i \neg\neg 0 = 1 \\ &\Leftrightarrow \mathbf{HA} \vdash_i 0 = 1 \end{aligned}$$

So **PA** is consistent if and only if **HA** is.

However, it is not always possible to embed classical systems into their intuitionistic counterparts. In particular, it turns out that intuitionistic analysis (second-order arithmetic with function variables) contradicts classical analysis. This will be elaborated on in the next section.

2 Analysis: Choice Sequences

A *choice sequence* is a potentially infinite sequence of mathematical objects $\alpha = \alpha(0), \alpha(1), \alpha(2), \ldots$ chosen, one after the other, from a fixed collection of mathematical objects by the individual mathematician (from 1948 on, Brouwer explicitly speaks of *the creating subject* although he must have had the notion already in 1927). Here we will limit our discussion to choice sequences of natural numbers and rationals. A choice sequence is an incomplete object, for it is never finished.

Choice sequences come in many varieties, depending on how much freedom one allows oneself in making the successive choices. The two extreme cases are the lawless sequences, where there is no restriction whatsoever on future choices, and the lawlike sequences, where one simply takes the numbers generated by a law or algorithm. One may (but need not) identify 'lawlike' with 'recursive'. (A lawlike sequence need not be thought of as an incomplete object, provided one is willing to make the additional abtraction from the temporal unfolding of the sequence.)

There are various reasons why this variety is relevant. First, a type need not be closed under a given operation. Consider the sum of two lawless sequences $\gamma = \alpha + \beta$, i.e.

$$\gamma = \alpha(0) + \beta(0), \alpha(1) + \beta(1), \alpha(2) + \beta(2), \ldots$$

This γ is itself neither lawless (because it depends on α and β), nor lawlike (because α and β are lawless). Second, lawlike sequences are needed to instantiate specific existence claims. Third, lawless sequences are important for metamathematical purposes.

Brouwer probably came to accept choice sequences as objects of intuitionistic mathematics in 1914, but theory development began in 1916–17. He showed how, using choice sequences, one can formulate a theory of the continuum that does not let it dissolve into separate points. Thus, Brouwer was the first to show how to incorporate into mathematics a point already made by Aristotle and others: a set of discrete elements cannot represent the geometrical or intuitive continuum. Discreteness and continuity are inseparable, complementary notions, that cannot be reduced to one another. Neither Cantorian set theory nor earlier constructivist analyses of the continuum (e.g. Poincaré, Borel, Brouwer in his dissertation of 1907, Weyl in 1918) had been able to accommodate this insight.

How does this work? Brouwer *identifies* a 'point' with a choice sequence of numbers that represent, through some coding, rational intervals on the continuum; these intervals should satisfy the Cauchy condition. A point, then, is 'becoming' and often to some extent undetermined. Brouwer then notices that, in general, extensional identity of choice sequences is undecidable. This models the non-discrete nature of the continuum.

The undecidability of extensional identity follows from the incompleteness of choice sequences: at any particular time, all there is of a choice sequence is a finite initial segment with an open end. Even if the initial segments of two sequences are the same, still nothing can be said about whether they will always have the same values. (In the case of two lawlike sequences, one may be able to show extensional identity by proving equivalence of the laws governing them.)

Choice sequences are generated freely, and at any time we have no more than a finite initial segment of them, perhaps together with some self-imposed restrictions. But then a sequence cannot, at any stage, have (or lack) a certain property if that could not be demonstrated from the information available at that stage. It follows that bivalence, and hence PEM, does not hold generally for statements about choice sequences. For example, consider a lawless sequence α of which we have so far generated the intial segment 8, 1, 3, and the statement P = 'The number 2 occurs in α.' Then we cannot say that $P \vee \neg P$ holds. Note how this argument against the validity of PEM depends on both the freedom of generation and the potential infinity of the sequences. We see that acceptance of choice sequences as mathematical objects forces a revision of logic along the lines of the proof interpretation given above. According to Placek (1999), this is the strongest argument in favor of intuitionistic logic currently available. (The philosophical thesis that logic may vary according to the ontological region one is speaking about, has been elaborated by Tragesser (1977), taking his cue from Husserl; in category theory, the phenomenon is familiar from *topoi*.)

Just as in classical mathematics elements are collected into a set, choice sequences are held together in a *spread* ('Menge,' in Brouwer's original, somewhat confusing terminology). A *spread law*, which should be decidable, either admits an initial segment or inhibits it; a further condition on the spread law is that of each admitted segment, at least one immediate extension should be admitted as well. The admitted segments form a growing tree, hence they are also known as nodes. Because of the second conditon, there will be no finite maximal paths in the tree. Choice sequences correspond to the infinite paths, and are called the *elements* of the spread.

A special case is the universal spread, which admits all choice sequences. The spread of all choice sequences satisfying the Cauchy condition is one way to represent the continuum.

For a few particular classes of choice sequences, there are translation theorems. For simplicity, we look at the case of lawless sequences, but the arguments that follow are general. Troelstra, developing earlier work by Kreisel, presented a formal system LS describing lawless sequences, together with a mapping τ into a subsystem without variables for lawless sequences IDB_1, such that

1. $\tau(\mathrm{A}) \equiv \mathrm{A}$ for A a formula of IDB_1
2. $\mathrm{LS} \vdash \mathrm{A} \leftrightarrow \tau(\mathrm{A})$
3. $(\mathrm{LS} \vdash \mathrm{A}) \Leftrightarrow (\mathrm{IDB}_1 \vdash \tau(\mathrm{A}))$

Such translation theorems show the coherency of the translated notion as a mathematical notion, and are important for metamathematical purposes. However, it cannot be concluded right away that translations explain lawless sequences away. These translations take the form of equivalences. An interest in ontological reduction would demand that we regard them as contextual definitions of quantification over lawless sequences. However, such a demand would have to be supported by arguments against such sequences that are independent of the axiomatization, for as the translation is symmetric, it could just as well be taken to mean that in some cases, quantification over lawlike sequences is best explained as quantification over lawless sequences.

More generally, such translations depend on specific axiomatizations of choice sequences. (In fact, lawless sequences have been axiomatized in different ways (Kreisel, Myhill, Troelstra), that are not always equivalent.) But an axiomatization is a way to present mathematical content; it is not identical with it. Hence the need for independent arguments. Brouwer certainly thought of choice sequences of any type as genuine objects of mathematics, constructed by the creating subject. (A phenomenological justification of this conviction can be found in van Atten 1999.)

The incompleteness of choice sequences guarantees properties that are desirable to model the continuum, but may at the same time seem to make them unworkable in practice. For if mathematics is to be based on constructions, what place is there for objects that at no stage have been completely constructed? Fortunately, there is a continuity principle. Essentially, this says that all one has to know to make a predication of some choice sequence is an initial segment. Unlike the sequence itself, its initial segments are given in a finite construction.

(**WC-N**) $\forall\alpha\exists x\, A(\alpha, x) \Rightarrow \forall\alpha\exists m\exists x\forall\beta[\overline{\beta}m = \overline{\alpha}m \rightarrow A(\beta, x)]$

where α and β range over choice sequences of natural numbers, m and x over natural numbers, and $\overline{\alpha}m$ stands for $\langle\alpha(0), \alpha(1), \ldots, \alpha(m-1)\rangle$, the initial segment of α of length m. 'WC-N' stands for 'Weak Continuity for Numbers': weak, as it only says something about each α individually (local continuity).

From WC-N, two theorems follow that show that intuitionistic analysis is not just an amputation of classical mathematics, but contains new results that are classically not acceptable. (It is true that there is no contradiction between the classical and intuitionistic systems of analysis as such, as key terms ('point,' 'function') are defined differently; but contradiction arises when one realizes that both systems try to capture the *same*, pre-formal notions of 'continuum' and so on.)

Veldman (1982) has shown that from WC-N one can derive

THE CONTINUITY THEOREM A real function whose domain of definition is the closed segment $[0, 1]$ is continuous on $[0, 1]$:

$\forall\varepsilon\forall x_1\forall\delta\forall x_2(|x_1 - x_2| < \delta \rightarrow |f(x_1) \rightarrow f(x_2)| < \varepsilon)$
for positive δ, ε and $x_1, x_2 \in [0,1]$.

THE UNSPLITTABILITY OF THE CONTINUUM The continuum cannot be split into two nontrivial subsets: if $\mathbb{R} = A \cup B$ and $A \cap B = \emptyset$, then $A = \mathbb{R}$ or $B = \mathbb{R}$.

Weyl announced the continuity theorem in 1921, but this is not really the same strong result as Brouwer's. Weyl defined real functions in such a way that they are continuous by definition, that is via mappings of the intervals of the choice sequence determining the argument to intervals of the image sequence. This way, the function type is reduced from $\mathbb{R} \rightarrow \mathbb{R}$ to $\mathbb{N} \rightarrow \mathbb{N}$ (initial segments to initial segments). Brouwer, on the other hand, established the continuity of functions from choice sequences to choice sequences, by showing how this followed from intuitionistic principles and the functional character (the $\forall\exists!$-combination).

Brouwer did not explicitly state the continuity theorem, instead he proved the stronger

UNIFORM CONTINUITY THEOREM A real function whose domain of definition is the closed segment [0, 1] is uniformly continuous on [0, 1]

$$\forall\varepsilon\exists\delta\forall x_1\forall x_2(|x_1 - x_2| < \delta \rightarrow |f(x_1) \rightarrow f(x_2)| < \varepsilon)$$

for positive δ, ε and $x_1, x_2 \in [0, 1]$.

Brouwer used the bar theorem (see below) to prove the uniform continuity theorem and seems to have believed that the continuity theorem can only be obtained as a corollary from it. Likewise, Brouwer in his proof of the unsplittability of the continuum appealed to the fan theorem (see below), where the simpler WC-N suffices, for unsplittability is a direct consequence of the continuity theorem: suppose $\mathbb{R} = A \cup B$ and $A \cap B = \emptyset$, then f defined by

$$f(x) = \begin{cases} 0 & \text{if } x \in A \\ 0 & \text{if } x \in B \end{cases}$$

is total and therefore, by the continuity theorem, continuous. But then f must be constant, so either $\mathbb{R} = A$ or $\mathbb{R} = B$. An instance of unsplittability is that it is not true that every real number is either rational or irrational. For if it were, we could obtain a non-trivial splitting of the continuum by assigning 0 to rational, and 1 to irrational real numbers.

Also note that WC-N by itself already suffices to refute PEM: consider $\forall\alpha[\forall x(\alpha x = 0) \vee \neg\forall x(\alpha x = 0)$, 'Every choice sequence is either the constant zero sequence, or not.' This is equivalent to $\forall\alpha\exists z[(z = 0 \rightarrow \forall x(\alpha x = 0)) \wedge (z \neq 0 \rightarrow \neg\forall x(\alpha x = 0))]$. Applying WC-N to this gives:

$$\forall\alpha\exists z\exists m\forall\beta(\overline{\beta}m = \overline{\alpha}m \rightarrow [(z = 0 \rightarrow \forall x(\beta x = 0)) \wedge (z \neq 0 \rightarrow \neg\forall x(\beta x = 0))])$$

Now take $\alpha = \lambda u \cdot 0$ and determine the z and m that WC-N correlates to this α. Then the above says that each β with an initial segment of m zeros will consist of zeros throughout, which is of course not the case.

Also refuted by WC-N is Church's Thesis in the form

CT $\forall a\exists x\forall y\exists z[T(x, y, z) \wedge U(z) = a(y)]$

that says, for every sequence a there exists a Turing machine with index x that calculates, for a natural number y, the $y + 1$th member of the sequence (z represents the computation process, and $U(z)$ its result). CT fails if a ranges over the whole universe and WC-N is true. For in that case, the index x would always have to be determined from just an initial segment of a, which is impossible.

An application of WC-N to a predicate $A(\alpha, x)$ determines a set of initial segments (nodes) that suffice to calculate an x such that $A(\alpha, x)$ holds. Such a set is called a *bar*.

To see if one can arrive at stronger results in analysis, one would have to know whether bars have structural properties. Brouwer managed to find such a property. For convenience, we consider a *thin* bar B, that is one with the property that if $\vec{n} \in B$ and $\vec{m} < \vec{n}$ (in the ordering of the tree), then $\vec{m} \notin B$ (i.e. a thin bar contains no initial segments that are longer than strictly necessary). Brouwer's *bar theorem* shows that the collection of thin bars, call it *ID*, is inductively defined (they are well-ordered). The clauses are:

1. Every singleton tree is in *ID*;
2. If $T_1, T_2, T_3, \ldots$ are in *ID*, then so is the tree obtained by adding a top to the direct sum of $T_1, T_2, T_3, \ldots$

This is a powerful insight, for it allows one to use induction in reasoning about thin bars. One sees from the order of the quantifiers why uniform continuity is stronger than ordinary continuity: for a given ε, uniform continuity demands that the same δ work for the whole interval, whereas for ordinary continuity, δ may vary with each x_1. This observation makes it plausible that uniform continuity should require knowledge of the structure of the bar whereas ordinary continuity does not.

Brouwer's proof of the bar theorem strongly depends on the intuitionistic notion that truth of a proposition consists in having a construction for it, and on reflection on the available means to construct proofs concerning bars.

A bar may well be an infinite tree (not in depth, but in width). A *fan* is a finitely branching tree. A corollary of the bar theorem is the *fan theorem*: if B is a thin bar for a fan, then there is an upper bound to the length of the nodes in B. Briefly put, a thin bar for a fan is finite. (The contrapositive of the fan theorem is better known, but was proven later (1927): König's infinity lemma, which says that a fan with infinitely many nodes contains an infinite path. It is not constructively valid, for there is no effective method to pick out a path that is infinite.)

The unit continuum $[0, 1]$ can be represented by a fan, for example by demanding that for every n, the nth interval is of the form

$$\left[\frac{a}{2^{n+1}}, \frac{a+2}{2^{n+1}}\right]$$

where $2 \leq a + 2 \leq 2^{n+1}$, as then the number of alternatives at choice n is finite. Thus it is that Brouwer could prove theorems about the continuum (such as the uniform continuity theorem) from a theorem on the constructively more tractable finitary trees. This shows the power of the fan theorem.

The weak counterexamples, of which we saw an example in the section on logic, require no more than lawlike sequences and intuitionistic logic. By exploiting the presence of sequences that are not lawlike but involve genuine choice, Brouwer in 1949 found a systematic and explicit way to construct *strong* counterexamples, which show that, if one accepts non-lawlike sequences, certain classical principles are not only without proof so far but could never be proven at all, as they are contradictory. These strong counterexamples are based on the theory of the creating subject; we adopt Kreisel's terminology here.

Let $\Box_n A$ stand for 'the creating subject experiences A (has full evidence for A) at time n'. The following principles (Kripke, Kreisel) are evident:

1. $\forall n \forall m\,(\Box_n A \rightarrow \Box_{n+m} A)$
   that is evidence never gets lost;
2. $\forall n(\Box_n A \vee \neg\Box_n A)$
   that is at every moment the creating subject can decide whether it has full evidence for A or not;
3. $A \leftrightarrow \exists n \Box_n A$
   A holds exactly if the creating subject has full evidence for it at some moment. (Kreisel dubbed this the 'Principle of Christian Charity,' or, alternatively, the 'Principle of Infinite Vanity': if something is true, the creating subject will sooner or later experience this.)

These principles more or less define the intuitionistic conception of truth.

On the basis of 1–3, one can associate with each proposition A a choice sequence α that 'witnesses' A:

$$a(n) = \begin{cases} 0 & \text{if } \neg\Box_n A, \\ 1 & \text{else} \end{cases}$$

The statement that such an α exists is known as 'Kripke's Schema':

(KS) $\exists \alpha(A \leftrightarrow \exists x \cdot \alpha(x) = 1)$

Brouwer used the principles 1–3, and implicitly Kripke's Schema, to establish strong counterexamples.

For example, in 1949 he showed

$$\neg\forall x \in \mathbb{R}(\neg\neg x > 0 \rightarrow x > 0)$$

and, by an argument of the same type,

$$\neg\forall x \in \mathbb{R}(x \neq 0 \rightarrow x \# 0)$$

(# denotes *apartness* of two real numbers: $a \# b \equiv \exists n(|a - b| > 2^{-n})$. In the proof interpretation, this is stronger than $\neg(a = b)$.)

In the proofs of these counterexamples choice sequences are employed that depend on the creating subject's having experienced either the truth or the absurdity of a particular mathematical assertion; these sequences are not lawlike. These methods are very powerful, for example one can prove that already the irrationals are unsplittable (van Dalen 1999b).

Besides analysis and counterexamples, other uses of choice sequences have been found. They are used in certain completeness proofs for intuitionistic predicate logic, and, together with KS, allow the definition of the (intuitionistic) powerset of $\mathbb{N}$ as a spread.

3 Further Semantics

As remarked, it is not easy to get model-theoretic results out of the proof interpretation, as the notion of 'construction' as employed there is still informal and not very specific. Therefore, various alternative semantics for intuitionistic logic have been developed (topological models, realizability, Kripke models, Beth trees, Martin-Löf's type theory, the Dialectica interpretation, sheaf semantics, topos models). The investment into various codifications of formal proof-notions should be rewarded by perspicuous effectiveness: the first prize being the 'existence property' or 'effective definability property': if $\exists x P(x)$ is proved constructively, the interpretation should supply us with an effective procedure to compute (or define) an object a and a proof of $P(a)$.

We will present four: realizability, Kripke semantics, the Dialectica interpretation, and Martin-Löf's type theory.

Realizability

Starting considerations from the finitary standpoint of Hilbert-Bernays, Kleene suggested that provability in HA of a statement of the form $\forall x \exists y \phi(x, y)$ should be taken to mean that there exists a recursive (choice) function f such that $\forall x \phi(x, f(x))$. Thus, the original statement is only an 'incomplete communication' (a notion introduced by Weyl), a full statement gives the choice function as well. Similarly, $\exists x \phi(x)$ is an incomplete communication of a full statement that specifies an object a such that $\phi(a)$. The idea behind Kleene's *recursive realizability* (or 1945-realizability) is to code all the information necessary to prove a particular statement ϕ into a natural number n. The notation is $n \mathbf{r} \phi$, '*n realizes* ϕ'.

The defining clauses of $\mathbf{r}$ mirror those of the proof interpretation. We use some notation from recursion theory: $\{x\}y$ for application, and $\downarrow$ for convergence.

$$
\begin{array}{rcl}
x \mathbf{r} \phi & := & \text{for atomic } \phi \\
x \mathbf{r} (\phi \wedge \psi) & := & (x)_0 \mathbf{r} \phi \wedge (x)_1 \mathbf{r} \psi \\
x \mathbf{r} (\phi \vee \psi) & := & ((x)_0 = 0 \rightarrow (x)_1 \mathbf{r} \phi) \wedge ((x) \neq 0 \rightarrow (x)_1 \mathbf{r} \psi) \\
x \mathbf{r} (\phi \rightarrow \psi) & := & \forall y (y \mathbf{r} \phi \rightarrow \{x\}y \downarrow \wedge \{x\}y \mathbf{r} \psi \\
x \mathbf{r} \exists y \phi(y) & := & (x)_1 \mathbf{r} \phi((x)_0) \\
x \mathbf{r} \forall y \phi(u) & := & \forall y (\{x\}y \downarrow \wedge \{x\}y \mathbf{r} \phi(y))
\end{array}
$$

According to the first clause, any number realizes an atomic sentence; no number, however, realizes a false atomic sentence. The second clause is obvious. The third clause shows the effective nature of the disjunction: as we can effectively test whether $(x)_0 = 0$ or $(x)_0 \neq 0$, the 'realizer' of a disjunction gives us all the information needed to indicate the desired disjunct. Similarly, the fifth clause says that a realizer of $\exists y \phi(y)$ codes the required instance and the information that realizes it. The fourth and sixth clauses are like the proof interpretation: the realizer of an implication transforms any realizer of ϕ into a realizer of ψ; the realizer of a universal statement is a partial recursive function that yields a realizer for any instance.

Note that $n \mathbf{r} \phi$ is itself a formula of HA, so realizability can be viewed as an interpretation of HA in itself. Therefore, it makes sense to ask for the truth of an instance of $n \mathbf{r} \phi$, or whether it is derivable in HA.

Since the introduction of realizability by Kleene, many variations on the original notion have been developed. In particular we mention 'truth realizability' **rt**, which is defined like **r** but with an extra condition in the clause for implication:

$$x \mathbf{rt} (\phi \to \psi) \quad := \quad \forall y(y \mathbf{rt} \phi \to \{x\}y \downarrow \wedge \{x\}(y) \mathbf{rt} \psi) \wedge (\phi \to \psi)$$

Truth realizability is particularly useful in showing how realizability renders the relation between existential statements and instantiations explicit. One can prove that

$\mathrm{HA}^* \vdash t \mathbf{rt} \psi \to$ and
$\mathrm{HA}^* \vdash \psi \Rightarrow \mathrm{HA}^* \vdash t \mathbf{rt} \psi$ for a suitable term t,

where HA^* is a suitable extension of HA in which partial terms are allowed, and which allow for a formalization of the basis of recursion theory. This fact is used to obtain an effective version of the existence property

$$\mathrm{HA}^* \vdash \exists x P(x) \Rightarrow \mathrm{HA}^* \vdash P(\overline{n}) \text{ for suitable } \overline{n}$$

Moreover, HA is closed under Church's rule:

$$\mathrm{HA} \vdash \forall x \exists y P(x, y) \Rightarrow \mathrm{HA} \vdash P(x, \{e\}y) \text{ for a suitable } e.$$

Since the index of the recursive (choice) function can be effectively determined, realizability provides the (admittedly not very practical) machinery needed to extract programs from proofs.

Kripke's semantics

In Kripke's semantics, the activity of the creating subject is modeled; it strongly resembles the theory of the creating subject mentioned above. At each point in time, the subject has constructed a collection of objects and has experienced a number of truths. The subject is free to take its activity of construction to a next stage; at each moment there is a number of possible next stages (or *possible worlds*). Thus, the stages for the individual form a partially ordered set (even a tree) $\langle K, \leq \rangle$; $k \leq \ell$ is taken to mean 'k is before, or coincides with, ℓ.' We write '$k \Vdash A$' for 'A holds at stage k;' the standard terminology is 'k forces A.' With every $k \in K$ we associate its local domain of objects created so far, denoted by $D(k)$. A reasonable assumption is that objects, once created, are not destroyed later: $k \leq \ell \Rightarrow D(k) \subseteq D(\ell)$.

The interpretation of the logical connectives now consists in spelling out the clauses of the proof interpretation in this possible-world model of the subject's activity. Then the inductive definition of the forcing relation is obvious:

For atomic A, $k \Vdash A$ is given; $\perp$ is never forced.

$$\begin{aligned} k &\Vdash A \wedge B \quad \Leftrightarrow \quad k \Vdash A \text{ and } K \Vdash B \\ k &\Vdash A \vee B \quad \Leftrightarrow \quad k \Vdash A \text{ or } k \Vdash B \\ k &\Vdash A \to B \quad \Leftrightarrow \quad \forall \ell \geq k(\ell \Vdash A \Rightarrow \ell \Vdash B) \end{aligned}$$

$$
\begin{array}{lcl}
k \Vdash A & \Leftrightarrow & k \Vdash A \to \bot \\
 & \Leftrightarrow & \forall \ell \geq k (\ell \Vdash A \Rightarrow \ell \Vdash \bot) \\
 & \Leftrightarrow & \forall \ell \geq k (\ell \nVdash A) \\
k \Vdash \exists x A(x) & \Leftrightarrow & \exists \alpha \in D(k) k \Vdash A(a) \\
k \Vdash \forall x A(x) & \Leftrightarrow & \forall \ell \geq k \forall a \in D(\ell) \Vdash A(a)
\end{array}
$$

Note that the cases of $\wedge$, $\vee$ and $\exists$ are determined on the spot, whereas $\to$, $\neg$ and $\forall$ essentially refer to the future.

A *Kripke model* $\mathcal{K}$ is concrete partially ordered set with an assignment of domains and relations. A is true in a Kripke model $\mathcal{K}$ if for all $k \in K$, $k \Vdash A$. A is true, simpliciter, if A is true in all Kripke models. Semantical consequence is defined as follows: $\Gamma \Vdash A$ iff for all Kripke models $\mathcal{K}$ and all $k \in K$ $k \Vdash C$ for all $C \in \Gamma \to k \Vdash A$.

There is an extensive model theory for Kripke semantics. It is strongly complete for intuitionistic logic, that is $\Gamma \vdash_i A \Leftrightarrow \Gamma \Vdash A$, and in particular $\vdash_i A \Leftrightarrow A$ is true. Predicate logic is complete for Kripke models over trees, and for propositional logic we even have the *finite model property*: $\nVdash A \Rightarrow A$ is false in a Kripke model over a finite tree.

From the completeness over tree models, one proves the *disjunction property*:

(DP) $\vdash_i A \vee B \Rightarrow \vdash_i A$ or $\vdash_i B$

A straightforward proof of DP is as follows. Suppose $\nvdash_i A$ and $\nvdash_i B$, then there is a tree model $\mathcal{K}_1$ that does not force A, and a tree model $\mathcal{K}_2$ that does not force B. Now $\mathcal{K}_1$ and $\mathcal{K}_2$ are glued together: put the two models side by side and place a new node k below both. In k no proposition is forced. The result is a correct Kripke model, and since $\vdash_i A \vee B$ (given), $k \Vdash A \vee B$, and hence $k \Vdash A$ or $k \Vdash B$. But that contradicts the fact that A and B are not forced in $\mathcal{K}_1$ and $\mathcal{K}_2$; therefore, $\vdash_i A$ or $\vdash_i B$.

Similarly, there is the *existence property*

(EP) $\vdash_i \exists x A(x) \Rightarrow \vdash_i A(t)$ for a closed term t

The theorems support the intuitionistic intended meaning of 'existence,' but the straightforward proofs use *reductio ad absurdum* and are therefore not constructive. Here proof theoretical devices have come to the rescue (the normal form theorem):

1. If $\vee$ does not occur positively in any formula in $\Gamma \vdash_i A \vee B$, then $\Gamma \vdash_i A$ or $\Gamma \vdash_i B$
2. If $\exists$ and $\vee$ do not occur positively in any formulas in Γ and $\Gamma \vdash_i \exists x A(x)$, then $\Gamma \vdash_i A(t)$ for some closed term t.

The Dialectica *interpretation*

Gödel's *Dialectica* interpretation (1958) is an interpretation of HA where, as the primitive notion, 'construction' of the proof interpretation is traded in for 'computable function of finite type over the natural numbers,' axiomatized in his system T. The latter notion is both more specific and less abstract (i.e. closer to Hilbert's 'concrete' finitary methods). The main result can be stated

If HA $\vdash A$, then T $\vdash \exists x \forall y A_D(x, y)$ (A_D is quantifier-free; see below)

This explains Gödel's philosophical motivation to devise the interpretation; for combined with a willingness to grant that the principles of T are evident, the main result yields a consistency proof of HA (and, in combination with Gödel's embedding of PA into HA mentioned above, a consistency proof of PA). In other words, Gödel aimed to show that if one wants to go beyond Hilbert's finitary arithmetic (and to prove its consistency, one has to), the required non-finitary elements need not be as abstract as the intuitionistic notion of proof.

The interpretation A^D of a formula A is defined by induction on the number of logical operators in A ($s \ldots z$ and $V \ldots Z$ stand for finite (possibly empty) sequences of, respectively, arbitrary type or higher type; in particular, x and u denote the sequences of free variables in A and B):

$A^D := A$ for atomic A

For the induction step, suppose $A^D = \exists y \forall z A_D(y, z, x)$ and $B^D = \exists v \forall w B_D(v, w, u)$; then

$(A \wedge B)^D = \exists y v \forall z w (A_D(y, z, x) \wedge B_D(v, w, u))$
$(A \vee B)^D = \exists y v t \forall z w (t = 0 \wedge A_D(y, z, x) \vee t = 1 \wedge B_D(v, w, u))$
$(A \to B)^D = \exists V Z \forall y w (A_D(y, Z(yw), x) \to B_D(V(y), w, u))$
$(\exists s A)^D = \exists s y \forall z A_D(y, z, x)$
$(\forall s A)^D = \exists Y \forall s z A_D(Y(s), z, x)$
(negation is defined by $\neg A := A \to 0 = 1$)

The interpretation reduces the logical complexity of sentences at the cost of increasing the type of the objects. The interplay between, on the one hand, the connectives and, on the other, the quantifiers as constructively construed, introduces the higher-order functions and thereby removes quantifiers from the connected statements. As statements without quantifiers are decidable, the connectives between them become simple computable (truth) functions.

For example, $\exists x A(x) \to \exists u B(u)$ (for atomic A and B) is translated as $\exists U \forall x (A(x) \to B(U(x)))$. This renders exactly the constructive reading of the original formula: 'Given an object with property A, one can construct an object with property B', that is, there is a construction that takes an object with property A as input and yields an object with property B as output. Such constructions are the values for U in the translated formula.

It cannot be excluded that an intuitionistic proof of a statement invokes proofs of more complex statements; this exhibits a form of impredicativity in the proof interpretation. The *Dialectica* interpretation does not fare better here, as functionals of a higher type could be used to define functionals of a lower type. Also, unless one is willing to take the notion of 'computable functional' as primitive, logic will be needed again in the precise defintion of the intended class of functionals. For these reasons, it is not easy to assess the exact epistemological advantage of the *Dialectica* interpretation.

Martin-Löf's type theory

Per Martin-Löf was the first logician to see the full importance of the connection between intuitionistic logic and type theory. Indeed, in his approach the two are so closely interwoven, that they actually merge into one master system. His type systems are no mere technical innovations, but they intend to capture the foundational mean-

ing of intuitionistic logic and the corresponding mathematical universe (Martin-Löf 1975, 1984).

Martin-Löf points out that we not only consider propositions (statements) but also make judgments about them. That is, we may hold propositions true. The basic judgments we have to consider are:

1. A is a type
2. A and B are equal types
3. a is an element of the type A
4. a and b are equal terms of the type A.

We have the following correspondence between propositions and proofs on the one hand, and types and elements on the other hand:

A is a type	a is an element of the type A	A is inhabited
A is a proposition	a is a proof of the proposition A	A is true

The type formation corresponds exactly to the formation of propositions, as used in logic. It is a basic idea of Martin-Löf's type theory, that elements and types have canonical forms. This explains, for example, equality judgments. Why is $2 + 3 = 4 + 1 : \mathbb{N}$? That is to say, why are the terms $2 + 3$ and $4 + 1$ equal in the type $\mathbb{N}$? The answer is that $2 + 3$ and $4 + 1$ have the same canonical form 5 (i.e. $(1 + (1 + (1 + (1 + 1))))$). The rules for equality have to be understood in this way, for example

$$\frac{a = b : A}{b = a : A}$$

A particular feature of Martin-Löf's type theory that is the system does not take anything for granted, but always makes explicit all required assumptions. Thus, when making up a type from parts, all those parts have to satisfy the necessary requirements.

An informal example: in order to know that $a + b$ is a number, we have to know that a and b are numbers, formally stated: a and $b \in N \Rightarrow a + b \in N$. Or, considering types depending on a parameter, one has to make sure that the parameters are correctly chosen: $a \in N \Rightarrow A(a)$ is a type. Given the required rules for equality, substitution, etc., one goes on to list the various type constructions. Here are some basic rules governing judgments:

natural numbers	N *type*
product	$\dfrac{x : A \Rightarrow B(x)\,\text{type}}{\Pi x : A \cdot B(x)\,\text{type}}$
sum	$\dfrac{x : A \Rightarrow B(x)\,\text{type}}{\Sigma x : A \cdot B(x)\,\text{type}}$
disjoint sum	$\dfrac{A\,\text{type}\ B\,\text{type}}{A + B\,\text{type}}$
identity	$\dfrac{t \in A, s \in A, A\,\text{type}}{I(A, t, s)\,\text{type}}$

In the common set theoretical practice, $\Pi x : A.B(x)$ is the cartesian product, $\Sigma x : A.B(x)$ is the disjoint sum of the family $\{B(x) \mid x \in A\}$, $A + B$ is the disjoint sum of two sets. The identity type is rather unusual, it is a set which is inhabited if t and s are identical, otherwise it is empty.

Note that there is a dual reading: $\Pi x : A.B(x)$ becomes $\forall x : A.B(x)$ in the logical notation, etc.

The characteristic properties of the various types and their canonical elements are laid down by a number of rules:

Natural numbers

$$NI \quad 0:N \qquad \frac{t:N}{St:N}$$

(0 is a natural number, and if t is a natural number then its successor St is also a natural number). These rules introduce numbers.

$$NE \quad \frac{t:N \;\; t_0:A[0/x]x:N, y:A \Rightarrow t_1:A[Sx/x] \;\; x:N \Rightarrow A \text{ type}}{R_{x,y}(t,t_0,t_1):A[t/x]}$$

R_{xy} is the recursor operator, its nature will be explained below.

$$\Pi I \quad \frac{x:A \Rightarrow t:B \qquad x:A \Rightarrow B \text{ type}}{\lambda x \cdot t : \Pi x : A \cdot B}$$

$$\Pi E \quad \frac{t:(\Pi x:A)\cdot B \qquad t':A \qquad x:A \Rightarrow B \text{ type}}{App(t,t'):B[t'/x]}$$

The introduction rule is the common λ- abstraction. The elimination rule yields the application of the functional term t to the 'input' term t, usually written as $t(t')$, or tt'.

$$\Sigma I \quad \frac{t:A \qquad t':B[t/x] \qquad x:A \Rightarrow B \text{ type}}{p(t,t'):(\Sigma x:A)\cdot B}$$

(elements of the disjoint sum are thought of as pairs, the first item is from the 'parameter set' A, the second on from the parametrized set B_x)

$$\Sigma E \quad \frac{t:(\Sigma x:A)\cdot B \qquad A \text{ type}}{p_0(t):A} \qquad \frac{t:(\Sigma x:A)\cdot B \qquad x:A \Rightarrow B \text{ type}}{p_0(t):B[p_0(t)/x]}$$

For the remaining rules see for example Troelstra and van Dalen (1988: 580).

In addition one has to give rules for 'computing' terms. Here are some examples:

$$\begin{cases} R_{x,y}(0,\, t_0,\, t_1) & \rhd \;\; t_0 \\ R_{x,y}(St,\, t_0,\, t_1) & \rhd \;\; t_1[x,\, t/y,\, R_{xy}(t_1,\, t_0,\, t_1)] \end{cases}$$

$App.(\lambda x \cdot t,\, t') \rhd t[t'/x]$

$p_i(t_0,\, t_1) \rhd t_i \;\; (i = 0,\, 1),$

$(p_0(t),\, p_1(t)) \rhd t$

Where $\triangleright$ stands for 'converts to.' In the formalism these conversions are also presented in the form of rules.

The system with the above types and terms is a kind of minimal system, there are a number of meaningful types to be added to make it more convenient and to strengthen it. But as it is, one can demonstrate a few characteristic features.

The properties that one establishes for types and terms can immediately be copied for propositions and proofs. If one suppresses, as is usual, the proof terms, the old natural deduction rules reappear. Example:

$$\frac{x{:}A \Rightarrow B\,type, x{:}A \Rightarrow t{:}B}{\lambda x.t{:}A \to B} \quad \text{becomes} \quad \frac{A\,true \Rightarrow B\ true}{A \to B\,true}$$

Thus we can get the intuitionistic provable proposition by operating in type theory.

Actually we get a few extras for working in a constructive setting. For example the axiom of choice becomes derivable. In ordinary language the axiom reads:

$$\forall x \in A \exists y \in B(x) C[x, y] \to \exists f \in \Pi x{:}A.B \forall x \in A\ C[x, fx]$$

In type theory one can indeed find a term t such that:

$$t{:}\Pi x{:}A \Sigma y{:}B.C[x, y] \to \Sigma z{:}(\Pi x{:}A.B) \Pi x{:}A.C[x, zx]$$

This confirms the intuitive argument that one would make in the proof interpretation. Note that in the proper reading of the axiom of choice, one exploits the hybrid nature of the system, terms may be elements or proofs. This is a strong practical feature of Martin-Löf's type theory.

We have barely scratched the surface of the theory, but one can see the striking similarity to the proof interpretation. To some extent, choice sequences have been incorporated in this framework as well, by admitting non-standard type theories.

References

Brouwer, L. (1923) Intuitionistische splitsing van mathematische grondbegripper. *Nederl. Ak. Wetensch.* Verslagen, 32, 877–80.

Brouwer, L. (1925) Intuitionistische Zerlegung Mathematischer Grundbegriffe. *Jahresber Dtsch. Math. Ver.*, 33, 251–6.

Brouwer, L. (1975) *Collected Works* (vol. 1): *Philosophy and Foundations of Mathematics.* Amsterdam: North-Holland.

Glivenko, V. (1928) Sur la logique de M. Brouwer. *Acad. Royale Belg. Bull. Cl. Sci.* 14, 225–8.

Gödel, K. (1990–95) *Collected Works* (3 vols). Oxford: Oxford University Press.

Heyting, A. (1930) Die formalen Regeln der Intuitionistischen Mathematik II, III. *Die Preussische Akademie der Wissenschaften Sitzungsberichte, Physikalische-Mathematische Klasse*, 52, 57–71, 158–69.

Heyting, A. (1934) *Mathematische Grundlagenforschung, Intuitionismus, Beweistheorie.* Berlin: Springer.

Kolmogorov, A. D. (1925) Sur le principe de tertium non datur. *Mat. Sbornik*, 32, 646–67.

Mancosu, P. (1998) *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s*. Oxford: Oxford University Press.
Martin-Löf, P. (1975) An intuitionistic theory of types. In H. Rose and J. Shepherdson (eds.), *Logic Colloquium '73*, pp. 73–118. Amsterdam: North-Holland.
Martin-Löf, P. (1984) *Intuitionistic Type Theory*. Napoli: Bibliopolis.
Placek, T. (1999) *Mathematical Intuitionism and Intersubjectivity. A Critical Exposition of Arguments for Intuitionism*. Dordrecht: Kluwer.
Tragesser, R. (1977) *Phenomenology and Logic*. Ithaca, NY: Cornell University Press.
Troelstra, A. and D. van Dalen (1988) *Constructivism in Mathematics* (2 vols). Amsterdam: North-Holland.
van Atten, M. (1999) Phenomenology of choice sequences. PhD thesis, Utrecht University.
van Dalen, D. (1999a) *Mystic, Geometer, and Intuitionist: The Life of L.E.J. Brouwer* (vol. 1): *The Dawning Revolution*. Oxford: Clarendon Press.
van Dalen, D. (1999b) From Brouwerian counter examples to the creating subject. *Studia Logica*, 62, 305–14.
van Heijenoort, J. (1967) *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Cambridge, MA: Harvard University Press.
Veldman, W. (1982, April) On the continuity of functions in intuitionistic real analysis. Some remarks on Brouwer's paper: "Ueber Definitionsbereiche von Funktionen." Technical Report 8210, Mathematisch Instituut, Katholieke Universiteit Nijmegen.

Further Reading

Dummett, M. (1977) *Elements of Intuitionism*. Oxford: Oxford University Press. Second, rev. edn, Oxford: Clarendon Press, 2000.
Heyting, A. (1956) *Intuitionism: An Introduction*. Amsterdam: North-Holland. Third, rev. edn 1972.
Sommaruga, G. (2000) *History and Philosophy of Constructive Type Theory*. Dordrecht: Kluwer.
Troelstra, A. (1977) *Choice Sequences. A Chapter of Intuitionistic Mathematics*. Oxford: Oxford University Press.
van Stigt, W. (1990) *Brouwer's Intuitionism*. Amsterdam: North-Holland.
Weyl, H. (1949) *Philosophy of Mathematics and Natural Science*. Princeton, NJ: Princeton University Press.

33

Many-Valued, Free, and Intuitionistic Logics

RICHARD GRANDY

Standard logic is a package with two parts – a formal deductive apparatus and a conception of interpretation for the language. The deductive apparatus and the semantics are mutually reinforcing and in this chapter we examine primarily the semantic assumptions that formally justify the deductive machinery. The second part of the package, the semantics of 'standard' logic, includes the assumptions that:

- there are two and only two truth-values, True and False,
- every sentence of the language has a determinate truth-value in each interpretation,
- the truth-value of any sentence of the language in an interpretation is determined by the reference or extension of the parts of the sentence in that interpretation (together with the universe of discourse.)

This chapter concerns three historically important forms of non-standard logics:

1. *Many-valued logics* reject the assumption that there are only two truth-values – it explores the possiblities that some sentences may be neither true nor false. Among the reasons for rejecting the assumption are the belief that statements about the future, statements involving vague predicates or statements about quantum mechanical properties are always either true or false. Most many-valued logics begin by rejecting the law of excluded middle, though there are exceptions. The number of values ranges from 3 to various infinite sets. The nature of the further values varies widely from author to author as do the motivations for introducing the additonal values.
2. *Free logics* reject the assumption that truth-values depend only on the referents and extensions of the parts of the sentence. The primary motivation in this case is to give a treatment of names that have no referent.
3. *Intuitionist logic* and other constructivist logics reject the basic assumption, shared by classical logic and the alternatives listed above, that logic should be founded on truth values, and instead proposes to base logic for mathematics on the concept of a mathematical construction. The founder of intuitionistic logic, L. E. J. Brouwer, proposed this logic only for reasoning about mathematics, but various authors

have subsequently argued at length that standard logic should be replaced by intuitionistic logic in other domains.

Two other forms of nonstandard logic are discussed in Part XII: "Relevance and Paraconsistent Logics."

1 Two- and Three-Valued Logics

Frege, one of the originators of modern logic, argued that sentences designate their truth-values and assumed that there are just the two values True and False. Russell, the greatest developer and promoter of modern logic, thought of sentences as denoting propositions in his early work, including the monumental collaboration with Whitehead in *Principia Mathematica*. However, Post (1921) proved that the axiomatization for sentential logic given by Russell and Whitehead is complete with respect to a two-valued interpretation.

POST'S THEOREM Any sentence which cannot be derived from the standard axiomatization of sentential logic is false in a two-valued interpretation. The interpretation can be explicitly constructed given the sentence.

Russell and Whitehead cited Post's result approvingly in the preface to their second edition, and the two-valued interpretation of logic became standard. The introduction of the truth tables as a method of teaching and understanding the sentential connectives in place of the complicated derivations from the axioms of *Principia Mathematica* represented an enormous pedagogical gain, as well as a theoretical advance.

In the same article in which he proved the completeness of the axioms with respect to two-valued interpretations, Post explored generalizations of the truth functions to more values, and he is counted as one of the two founders of many-valued logic. Post's interests were entirely mathematical; he was interested in what happens when you generalize the two-valued interpretations to more values. His systems have been studied extensively, especially in recent decades as they provide a theoretical structure for the analysis of multi-valued switching circuits. However, they have not gathered much attention from philosophers.

The other major founder of many-valued logic is Łukasiewicz. He sketched the idea of a many-valued logic in 1920 and published a systematic account in 1930 (both are reprinted in Borkowski (1970)). Unlike Post, Łukasiewicz introduced three-valued logic for philosophical reasons, to provide a more appropriate representation for the indeterminacy of the future. He apparently was led to this concern both by a historical concern, studying Aristotle's discussion of necessity, particularly his sea battle example, and by a very contemporary concern about how to accommodate the indeterminism of modern physics within logic.

Aristotle's sea battle argument is:

1. If there will be a sea battle tomorrow, then necessarily there will be a sea battle tomorrow.

2. If there will not be a sea battle tomorrow, then necessarily there will not be a sea battle tomorrow.
3. Either there will or there will not be a sea battle tomorrow.
4. Therefore, either there will necessarily be a sea battle tomorrow or there will necessarily not be a sea battle tomorrow.

Aristotle suggested that premise (3), the principle of excluded middle, $\mathbf{A} \vee \sim \mathbf{A}$, should be rejected when A is a statement about a future contingency. Thus the motivation, if not the details, of many-valued logic are as ancient as the study of logic itself. Łukasiewicz developed this idea into a systematic logic.

In his original paper Łukasiewicz used 1 for truth and larger integers for other truth-values, but he later switched to using 1 for truth, 0 for falsity and intermediate values for other truth-values. Most, but not all other writers use this convention. Of course it is one thing to decide that 1/2 is your third truth-value and another to give a philosophical explanation of it. For Łukasiewicz the intermediate value is 'indeterminate.' Given this understanding, the most natural three-valued generalization of the two-valued truth tables are the following, in which negation reverses the value,

A	**~A**
1	0
1/2	1/2
0	1

conjunction takes the minimum value of the conjuncts.

A & B	1	1/2	0
1	1	1/2	0
1/2	1/2	1/2	0
0	0	0	0

and disjunction the maximum

A ∨ B	1	1/2	0
1	1	1	1
1/2	1	1/2	1/2
0	1	1/2	0

For example, the conjunction of a true sentence and an indeterminate one would seem to be indeterminate. It could become true if the indeterminacy was resolved in favor of truth, or false if it were resolved in favor of falsity.

Note that when all the components of a sentence formed from these connectives are all assigned value 1/2, then the entire sentence has value 1/2. If we introduce the conditional as $\sim\mathbf{A} \vee \mathbf{B}$, as is often done in two-valued logic, then conditionals would also have this property and there would be no sentences which are logical truths. More concretely, since that identification of the conditional with $\sim\mathbf{A} \vee \mathbf{B}$ makes $\mathbf{A} \rightarrow \mathbf{A}$ equivalent to *excluded middle* $\mathbf{A} \rightarrow \mathbf{A}$ would not be a logical truth.

Instead of using that traditional, if often questioned, equivalence, Łukasiewicz defined the conditional thus:

$\mathbf{A} \rightarrow \mathbf{B}$	1	1/2	0
1	1	1/2	0
1/2	1	1	1/2
0	1	1	1

One way of describing this table is that the conditional is false only in the case of True $\rightarrow$ False, and is Indeterminate only in the two cases: True $\rightarrow$ Indeterminate and Indeterminate $\rightarrow$ False. A rationale for these choices is that if **A** is true and **B** indeterminate, then the conditional $\mathbf{A} \rightarrow \mathbf{B}$ could be true if **B** were to be true, and false if it were false. The choice of the value 1 when both components have value 1/2 is required if $\mathbf{A} \rightarrow \mathbf{A}$ is to be logically true. Further, setting the value of $(\mathbf{A} \rightarrow \mathbf{B})$, which we will represent as $V(\mathbf{A} \rightarrow \mathbf{B})$, equal to 1/2 when the components are both assigned 1/2 would result in every sentence having value 1/2 when all its components do, and thus there would be no logical truths.

Equivalence can be defined as usual as $\mathbf{A} \leftrightarrow \mathbf{B}$ **iff** $(\mathbf{A} \rightarrow \mathbf{B}) \,\&\, (\mathbf{B} \rightarrow \mathbf{A})$. In all of the systems we will be considering equivalence is so treated and we will not make explicit mention of equivalence again. (In Łukasiewicz' presentation of his system, he used only negation and the conditional, having noted that $\mathbf{A} \vee \mathbf{B}$ can be defined as $(\mathbf{A} \rightarrow \mathbf{B}) \rightarrow \mathbf{B}$, and then **A&B** can be defined by using the usual DeMorgan's principle.)

In two-valued logic, we define a sentence to be logically true iff it is true in all interpretations. When we have more than two truth-values, then we must indicate which subset of the values are the designated values, those which are truth-like. Our definition now becomes

A IS A LOGICAL TRUTH iff it has a designated value in all interpretations.

Since Łukasiewicz' motivation was to deny excluded middle, he chose only 1 as a designated value. This achieves the purpose of rendering excluded middle not a logical truth. It has one somewhat counterintuitive consequence though, which is that under an interpretation in which both components are assigned value 1/2, A&~A has the same truth value as A$\vee$~A. This will be a consequence in any system of truth tables generalized along the principles above that has an odd number of truth-values, but not of those with an even number. This suggests that many-valued logics with an even number of truth-values might be preferable. Issues of the indeterminacy of the future are now generally studied within the framework of tense logic discussed in chapter 31, "Deontic, Epistemic, and Temporal Logics." Aristotle's argument is generally regarded as fallacious, but Łukasiewicz's innovations have opened the possibilities for a variety of other systems and ideas.

Another reason that has led philosophers and logicians to explore many-valued logics is to attempt to avoid paradoxes such as the Liar. The Liar sentence L is:

L. Sentence L is false.

This produces a paradox: if the sentence is false, what it asserts is correct and it is true; if the sentence is true, then what is asserts is correct and it is false. Introducing a third

truth-value 'paradoxical' gives a way out of the paradox. Bochvar was the first to suggest a three-valued logic as treatment for the paradoxes. His system differed from Łukasiewicz' since Bochvar's third value was 'paradoxical,' in contrast to Łukasiewicz' 'indeterminate.' Bochvar had a double set of connectives, but we will only mention the first set here. Since a paradoxical component, according to Bochvar infected an entire sentence, his truth table for conjunction was:

A & B	1	1/2	0
1	1	1/2	0
1/2	1/2	1/2	1/2
0	1	1/2	1

In this system, every sentence of the language has value 1/2 when all of its components are assigned 1/2, and thus there are no logical truths. There is, however, a related notion, that of a sentence which is never false. This set coincides with the classical two-valued logical truths.

However, the relief from paradox is at most temporary because the revised Liar L′:

L′: L′ is false or indeterminate.

produces a new but closely related paradox.

The other relatively well-known system of three-valued logic is due to Kleene. His motivation was to deal with statements or equations involving partially defined functions and consequently his third truth-value was 'undefined.' Since a conjunction of a false sentence and an 'undefined' could not turn out to be anything but false, his truth tables for conjunction, disjunction, and negation were the same as Łukasiewicz. However, for the conditional, Kleene regarded a conditional with both antecedent and consequent 'undefined' to have the value undefined. Thus his conditional was characterized as:

A → B	1	1/2	0
1	1	1/2	0
1/2	1	1/2	1/2
0	1	1	1

As in the Bochvar system no sentence receives value 1 on all interpretations. The most significant and plausible application of Kleene's system in philosophy was given by Korner (1966) in relation to the concept of an inexact class. Various linguists have also made use of the Kleene connectives in application to natural languages.

It is also worth mentioning that Reichenbach introduced a three-valued logic as part of an attempt to provide a better logical framework in which to understand quantum mechanics. This was a complex system with three negations and three conditionals. This approach was superseded by quantum logic; it is controversial whether quantum logic is to be considered a many-valued logic. For further discussion, we refer the reader to Part XI, "Inductive, Fuzzy, and Quantum Logics for Probability."

2 Finite Valued Systems with more than Three Values

The Łukasiewicz three-valued generalization can be systematically carried further. The n-valued generalization consists of taking the values $i/n - 1$ for $0 \leq i \leq n - 1$. Conjunction will take the minimum value of the conjuncts, and disjunction the maximum value; the value of a negation is 1 minus the value of the negated sentence. For the conditional $\mathbf{A} \rightarrow \mathbf{B}$ we have two clauses:

$V(\mathbf{A} \rightarrow \mathbf{B}) = 1$ if $V(\mathbf{A})$ is less than or equal to $V(\mathbf{B})$, and

$V(\mathbf{A} \rightarrow \mathbf{B}) = [1 - V(\mathbf{A})] + V(\mathbf{B})$ otherwise.

In all of the Łukasiewicz systems the only designated value is 1. Excluded middle will not be logically true in any of these systems, though in the even valued systems excluded middle is always truer than the contradiction **A&**–**A**. Systems with more than 1 designated value were mentioned by Post and this variation on Łukasiewicz systems was studied by Slupecki and others.

Four-valued logic was proposed for modal logic, the values being 'necessarily true,' 'contingently true,' 'contingently false,' and 'necessarily false.' The Łukasiewicz definitions of the usual connectives can be used and a modal operator added. The necessity operator will map 'necessarily true' onto itself and all other values onto 'necessarily false.' While these truth tables have some uses, they have been superseded by the possible worlds approach to modal logic discussed in chapter 29, "Alethic Modal Logics and Semantics."

3 Infinite Valued Systems

The Łukasiewicz n-valued generalization can be systematically carried further – Łukasiewicz also studied the cases where the set of truth-values consists of all rational numbers in the interval [0,1] and where the values consist of all real numbers in the same interval. As before, conjunction will take the minimum value of the conjuncts, and disjunction the maximum value; the value of a negation is 1 minus the value of the negated sentence. For the conditional $\mathbf{A} \rightarrow \mathbf{B}$ we again have the two clauses:

$V(\mathbf{A} \rightarrow \mathbf{B}) = 1$ if $V(\mathbf{A})$ is less than or equal to $V(\mathbf{B})$, and

$V(\mathbf{A} \rightarrow \mathbf{B}) = [1 - V(\mathbf{A})] + V(\mathbf{B})$ otherwise.

Some applications and extensions of these systems will be discussed in later sections. An important breaking point with respect to axiomatizability occurs in this region. All of the finite Łukasiewicz logics are axiomatizable in both their sentential and quantificational forms. In the finite-valued logics the quantifiers are straightforward generalizations of the principles for conjunction and disjunction. A universally quantified expression has as its value the minimum of the values of the Fx. However, in the infi-

nite case, the set of values of Fx may be a set whose minimum, greatest lower bound, is not a member of the set. For this reason, the rationals [0.1] are a satisfactory logic for sentential logic, but the full continuum [0,1] is required for quantificational Łukasiewicz systems. It has been shown that the infinite-valued quantified Łukasiewicz logic is not recursively axiomatizable.

4 Vagueness, Many-valued and Fuzzy Logics

Another philosophical perplexity for which many-valued logics have been prescribed as remedy concerns vagueness. A natural first step in dealing with borderline cases would be to introduce a third truth-value. However, this seems unsatisfactory for it merely replaces the unrealistically sharp boundary between true and false with two unrealistically sharp boundaries, one between true and indefinite, and the other between indefinite and false. More finite values seem only to make the problem worse, and even moving to the infinite case seems to render in appropriate results inasmuch as seems counterintuitive to suppose that a vague statement has a precise real number as its truth-value. However, an important proposal for analyzing vagueness has been based on the continuum valued Łukasiewicz logic.

Zadeh (1975) first introduced the conception of a *fuzzy set* – a set for which membership is not a dichotomous matter but where the membership can take on any of the continuum of values in [0,1] He then replaced the idea of a precise truth mathematical truth-value with fuzzy linguistic truth-values. His truth-values are the countably infinite set: {true, very true, very very true, rather true, not true, false, very false, not very true and not very false, . . . } each of which is a fuzzy subset of the continuum [0,1]. Zadeh's ideas were further developed by Goguen (1968–9) who related them to inexact concepts.

Fuzzy logic and set theory have been enormously successful as tools in engineering and artificial intelligence, and many intelligent control systems from elevators to washing machines have been designed using fuzzy logic. However, as an approach to vagueness it has not been widely accepted in the philosophical community. Part of the resistance may be due to the fact that without the 'fuzzy linguistic values' the approach imputes too much precision to vague contexts, and on the other hand the 'fuzzy linguistic values' seem too unclear and undeveloped to be philosophically respectable. It is also possible that philosophers lack the mathematical sophistication to fully appreciate the approach.

5 Boolean Valued Systems

Another family of interpretations with a different flavor are the interpretations in which the truth-values are the elements of a Boolean algebra. A Boolean algebra is a generalization of principles that are common to elementary set theory and sentential logic. A Boolean algebra consists of a set of elements $\boldsymbol{B}$ with two distinguished elements, $\mathbf{0}$ and $\mathbf{1}$, a one place operation – and two two place operations $\cup$ and $\cap$, which satisfy a set of equations to be enumerated in a moment. We are using the familiar symbols

in **bold** for the Boolean notions for heuristic reasons, but it is important to distinguish the Boolean symbol $\cup$ from the set theoretic symbol $\cup$. We will see that the set theoretic operations are one instance of the Boolean operations.

Many alternative sets of axioms are available for Boolean algebras; a simple one that is not too redundant, where $\mathbf{x}$,$\mathbf{y}$ and $\mathbf{z}$ are any elements of $\boldsymbol{B}$

B1	$-\mathbf{0} = \mathbf{1}$	$-\mathbf{1} = \mathbf{0}$
B2	$\mathbf{x} \cap \mathbf{1} = \mathbf{x}$	$\mathbf{x} \cup \mathbf{0} = \mathbf{x}$
B3	$\mathbf{x} \cap -\mathbf{x} = \mathbf{0}$	$\mathbf{x} \cup -\mathbf{x} = \mathbf{1}$
B4	$\mathbf{x} \cap \mathbf{y} = \mathbf{y} \cap \mathbf{x}$	$\mathbf{x} \cup \mathbf{y} = \mathbf{y} \cup \mathbf{x}$
B5	$(\mathbf{x} \cap \mathbf{y}) \cup \mathbf{z} = (\mathbf{x} \cap \mathbf{z}) \cup (\mathbf{y} \cap \mathbf{z})$	$(\mathbf{x} \cup \mathbf{y}) \cap \mathbf{z} = (\mathbf{x} \cup \mathbf{z}) \cap (\mathbf{y} \cup \mathbf{z})$

One example of a Boolean algebra is to take $\boldsymbol{B}$ as the pair of truth-values {T,F}, with T as $\mathbf{1}$, F as $\mathbf{0}$, and negation, conjunction, and disjunction as the operations. Another family of examples of Boolean algebras is obtained by taking any nonempty set S, and letting $\boldsymbol{B}$ be the power set, the set of all subsets of S, with S as $\mathbf{1}$, the empty set as $\mathbf{0}$, set complement, union, and intersection as the operations.

What is of interest for our purposes it that if we take the elements of any Boolean algebra as truth-values, and then let our valuation function be defined for negation, disjunction, and conjunction by the Boolean operations, we find that we have a many-valued logic which validates exactly the same set of sentences as the standard two-valued. Post's theorem is sometimes taken as establishing that standard logic is two-valued, but in fact the correct statement is that standard logic is Boolean valued, and the two-valued interpretation is just the simplest Boolean algebra.

The Boolean valued systems are importantly different from the Łukasiewicz and the many-valued approaches discussed above because the values are not linearly ordered. For example, if we take a two element set S = {a,b} we generate a Boolean algebra with the four elements {a,b}, {a}, {b} and { }. If we now consider a disjunction A $\vee$ B and give an interpretation in which V($\mathbf{A}$) = {a} and V($\mathbf{B}$) = {b}, then the disjunction will have the union of these as its value, that is {a,b}. Thus in Boolean valuations a disjunction receives the least value which is greater than or equal to the values of the two disjuncts. Unlike the other many-valued logics above, a disjunction can be truer than either disjunct.

6 Supervaluations are Boolean Valued Logics

Supervaluations are an approach that was first suggested by Mehlberg in connection with vagueness, but were first developed formally by van Fraassen in the context of free logic (to be discussed in the next section). If we consider a vague predicate such as 'bald,' there is a natural intuition that there are some clear positive applications some clear negative applications and some borderline cases. One approach to vagueness is to

use one of the Łukasiewicz systems and deny that excluded middle holds if we are considering a borderline case.

The supervaluation approach is to consider the set of all precisifications of the concept bald, that is all of the ways that the concept could be turned into a precise one by adjudicating among the borderline cases while preserving the positive and negative. We then call a statement Supertrue if it is true in all precisifications. Given our remarks above about Boolean algebras, it is evident that supervaluations are essentially a many-valued approach in which the values are members of a Boolean set algebra – the relevant set being that of the precisifications.

One of the advantages of this many-valued approach to vagueness is that we can make distinctions among the borderline cases. If Fred and Paul are both among the borderline cases of bald, but Fred has more hair than Paul, then in a supervaluation approach it will be true in fewer precisifications that 'Fred is bald,' and thus that sentence will receive a lower truth value than 'Paul is bald.' The main philosophical weakness of the approach is that the fundamental assumptions about precisifications and the specification of positive and negative cases have not yet been made sufficiently clear.

7 Free Logic

Arisototelian syllogistic logic assumed that the general terms involved in reasoning were nonempty. That is, in treating sentences of the form 'All Gs are Hs' it was assumed in evaluating the validity of arguments that there is at least one G and at least one H. Thus 'All unicorns are white' would not fall within the scope of syllogistic in spite of its form, since there are no unicorns. Modern logic does not make this assumption and sentences of the form $\forall x(Gx \rightarrow Hx)$ are permitted even when G or H are assigned the empty set.

However, standard logic does make existence assumptions in two forms. First, the domain of quantification must be *nonempty*. The symbolic representation of the assumption is the validity of the sentence $\forall xGx \rightarrow \exists xGx$. Second, it is assumed that all constants in the language denote some object. This is reflected in the validity of sentences of the form $Gc \rightarrow \exists xGx$.

Free logic dispenses with these assumptions. There are two main, and slightly different, motivations for this step. One is a methodological or ontological concern to make the foundations of logic as free from existential assumptions as possible. The second is an interest in applying logic to natural languages where, many believe, there are non-denoting terms such as 'Zeus' and 'Sherlock Holmes.' (It should be noted that there are opposing views on which 'Zeus' denotes a mythological god and 'Sherlock Holmes' a fictional detective.)

As with many-valued logics, there are a variety of proposals for free logic systems and a large and ongoing research program concerning them. In systems which include identity as a logical operation, the fact that a constant c denotes can already be expressed as $\exists x(x = c)$; in systems which do not include identity, a new logical expression, usually either '**E**' or '**E!**' is introduced as a one-place predicate. Exactly how one modifies the axioms and rules of inference of standard logic varies in detail depending

on the particular formulation of standard logic, but the basic ideas are fairly straightforward. In place of the standard rule of existential introduction, which permits the inference from Gc to $\exists$xGx, we have the slightly more complicated rule which requires an additional premise, namely $\exists$x(x = c). Universal elimination (or instantiation) is similarly modified.

This negative pruning of the derivational system is straightforward and agreed upon, but there agreement ends. The problems arise when we consider how to evaluate the truth of Gc when 'c' is a non-denoting term. *Negative free logic* declares all atomic sentences containing non-denoting terms to be false. *Positive free logics* declare at least some atomic sentences containing non-denoting terms, for example c = c, to be true. Neutral free logics are non-committal. Negative free logic satisfies the methodological concern, but is less satisfying to those who are motivated by natural language considerations because the latter often want a theory in which sentences such as 'Zeus is Zeus,' 'Sherlock Holmes is a fictional detective' and perhaps even 'Sherlock Holmes lived in London' are true.

There is also one version of positive free logic which satisfies the methodological but not the linguistic concerns. On this theory not only is 'c = c' true for all terms, 'c = d' is also true for any pair of non-denoting terms. This makes 'Zeus is Zeus' true, but also makes true the unwanted 'Zeus is Sherlock Holmes'!

Matters become even more complex if we consider a language with a definite description operator. Following Russell we use ιxGx to stand for 'the object which is G.' However, while Russell regarded statements including the description to be paraphrasable into standard logic without descriptions, free logic takes the definite description as basic. And very unlike Russell, positive free logics treat some of the atomic occurrences of non-denoting descriptions as true. One plausible further principle is to extend the validity of self-identity to all descriptions regardless of whether they denote, that is to make ιxGx = ιxGx valid regardless of the interpretation of G.

A tempting further extension would be to declare that each definite description satisfies the condition of the description, that is to say that the winged horse is winged, and so on. However this temptation must be resisted as it leads to an inconsistent system when we take G as ~x = x, because then we obtain both ιx(~x = x,) = ιx(~x = x,) from our previous principle, and ~[ιx(~x = x,) = ιx(~x = x,)] from our new principle.

Given the disagreement over which free logic principles are correct, it is not surprising that there are a variety of semantic proposals. Many of the proposals introduce a second domain to the interpretations. The first domain is the domain over which quantifiers range, but the non-denoting terms are associated with various objects in the second domain. Technically the second domain is impeccable, but the philosophical interpretations of it are varied and controversial.

A slightly different approach to free logic stems from a concern that logical principles should be true regardless of the denotation of terms, that is excluded middle should be valid even in instances like 'Either Zeus was blue-eyed or Zeus was not blue-eyed.' A method of achieving this end while avoiding issues about the truth of atomic sentences is to use supervaluations. A supervaluation in this context is a set of interpretations which assign objects to the constants which lack denotations in the starting interpretation. Since any assignment of an object 'Zeus' will make one or the other of the disjuncts true, the disjunction true though neither disjunct is. Some authors describe

supervaluations as 'non-truth functional' in this context, but the view given above seems more accurate.

All of the above discussion, however, is based on free logics which accept the two-valued assumption. That is, they reject the existence assumptions of classical logic but accept the two-valuedness assumption. More radical approaches to free logic (Jacquette 1996) also move to a many-valued set of truth-values. It is possible that the combination of these approaches will prove more philosophically compelling than the separate strands.

Further discussions of the topics of this section are to be found in Part IV: "Truth and Definite Description in Semantic Analysis" and Part VI: "Logic, Existence, and Ontology."

8 Intuitionism

Intuitionistic logic was created by L. E. J. Brouwer, a Dutch mathematician, in response to the set theoretic paradoxes, also discussed in Part VIII: "Logical Foundations of Set Theory and Mathematics," and also due to a general dissatisfaction with the understanding of the logic of mathematics as being a logic of independently existing objects, properties, and relations. In Brouwer's neo-Kantian philosophy, mathematics is a human creation and the fundamental notion is one of a mathematical construct, rather than truth and reference. For the classical logician, the statement that every natural number has a successor is true because there exist infinitely many natural numbers and the successor relation picks out a relation which holds between adjacent numbers. For Brouwer, the statement that every natural number has a successor is known because we know that there is a construction which for every natural number gives a successor natural number.

Brouwer's explanation of the logical connectives is given in terms of constructions. A construction establishes a conjunction if it consists of two parts, one of which establishes each conjunct; a construction establishes a disjunction iff it establishes one of the disjuncts and specifies which. A construction establishes a negation $\sim\mathbf{A}$ iff it is a construction which shows that if there were a construction establishing A, then we could also establish $0 = 1$. A construction establishes a conditional $\mathbf{A} \rightarrow \mathbf{B}$ iff it is a construction which, applied to any construction which establishes A, establishes that B. Note that in these last two clauses we are appealing to the application of constructions to constructions.

For the quantifiers we have, in the domain of natural numbers, a construction establishes $\forall x\mathbf{F}x$ iff it is a construction which for any natural number $\mathbf{n}$ produces a construction establishing $\mathbf{Fn}$. Analogously, in the domain of natural numbers, a construction establishes $\exists x\mathbf{F}x$ iff it is a construction which produces a natural number $\mathbf{n}$ and a construction which establishes $\mathbf{Fn}$.

Given this understanding of the connectives, instances of excluded middle such as $\exists x\mathbf{F}x \vee \sim\exists x\mathbf{F}x$, are not valid. If we take $\mathbf{F}$ to be a complex mathematical formula there is no reason to think that we can either find a specific instance of $\mathbf{F}$, or give a proof that the existence of such an $\mathbf{F}$ would imply a contradiction. Similarly, the classically valid inference from $\sim\forall x\mathbf{F}x$, which can be obtained by showing that the assumption $\forall x\mathbf{F}x$

leads to a contradiction, is insufficient to establish $\exists x \sim \mathbf{F}x$ since the proof does not typically provide a specific counterexample n.

Another classical principle which is not valid is double negation elimination: $\sim\sim\mathbf{A} \rightarrow \mathbf{A}$, although the subcase of it $\sim\sim\sim\mathbf{A} \rightarrow \sim\mathbf{A}$ is intuitionistically valid. Brouwer also opposed the then standard view that logic provided a foundation for mathematics. In Brouwer's view, mathematics required no foundation and logic was merely a reflection of mathematical practice not its basis. He also opposed the formalization of logic.

However, his student, Heyting, in an effort to generate more interest in and sympathy for intuitionism provided a formalization.

H1 $\mathbf{A \rightarrow (A \,\&\, A)}$

H2 $\mathbf{(A \,\&\, B) \rightarrow (B \,\&\, A)}$

H3 $\mathbf{(A \rightarrow B) \rightarrow ((A \,\&\, C) \rightarrow (B \,\&\, C))}$

H4 $\mathbf{((A \rightarrow B) \,\&\, (B \rightarrow C)) \rightarrow (A \rightarrow C)}$

H5 $\mathbf{A \rightarrow (B \rightarrow A)}$

H6 $\mathbf{(A \,\&\, (A \rightarrow B) \rightarrow B)}$

H7 $\mathbf{(A \rightarrow (A \vee B))}$

H8 $\mathbf{(A \vee B) \rightarrow (B \vee A)}$

H9 $\mathbf{((A \rightarrow B) \,\&\, (C \rightarrow B)) \rightarrow ((A \vee C) \rightarrow B)}$

H10 $\mathbf{\sim A \rightarrow (A \rightarrow B)}$

H11 $\mathbf{((A \rightarrow B) \,\&\, (A \rightarrow \sim B)) \rightarrow \sim A}$

Adding either excluded middle or double negation elimination, as H12, gives an axiomatization of the standard two-valued logic. Adding the usual axioms for the quantifier expressions to Heyting's system H1–11 provides an axiomatization of quantified intuitionistic logic.

How do we know that excluded middle does not follow in some subtle way from these axioms, showing that either Heyting's axiomatization is wrong or that intuitionism is incoherent? Heyting provided a three-valued interpretation in which all the Heyting axioms always have value 1 but excluded middle does not. Since modus ponens can be seen to preserve logical truth, excluded middle does not follow. This is an example of the use of many-valued logics in independence proofs we alluded to in discussing three-valued logics.

In Heyting's interpretation, conjunction and disjunctions behave as in the Łukasiewicz systems but negation and conditional are slightly different:

A	$\sim$**A**	**A** $\rightarrow$ **B**	1	1/2	0
1	0	1	1	1/2	0
1/2	0	1/2	1	1	0
0	1	0	1	1	1

Does this mean that intuitionistic logic and the rich structure of mathematical constructions can be represented by the three-value tables? No, because the Heyting interpretation gives an interpretation on which excluded middle has value 1/2 while all the axioms uniformly have value 1, but there are other schemas which receive value 1 on all interpretations but are not intuitionistic truths. Specifically, it is not intuitionistically valid to assert that for any four sentences **A**, **B**, **C**, and **D**

$$(\mathbf{A} \rightarrow \mathbf{B}) \vee (\mathbf{B} \rightarrow \mathbf{C}) \vee (\mathbf{C} \rightarrow \mathbf{D})$$

But this sentence must always receive value 1 according to the Heyting scheme.

The extension of this to n sentences is also not intuitionistically correct, but as we observed earlier an n – 1 valued logic in which conditionals have value 1 when the antecedent has a value less than or equal to that of the consequent, the principle will always have value 1. Thus no finite valued logic can correctly represent intuitionism. Jaskowski proposed an infinitely valued logic which does exactly match.

While Jaskowski's proposal provides an exact characterization of the sentences which are always true in Heyting's logic, it seems to be a technical fact and does not provide any connection with the underlying motivations. Other semantics for intuitionistic logic which are not many-valued but rely instead on tree structures or topological spaces seem somewhat more satisfying. Details can be found in Dummett (1977).

9 Conclusions

The nonstandard logics discussed above were each proposed to deal with a philosophical problem, and the innovator felt that moving beyond the standard framework would provide progress toward an answer. Many of the systems have proved to be enormously productive as applied to practical problems unforeseen by their inventors, and almost all of them have provided fruitful ground for mathematical development. However, none have succeeded in displacing standard two-valued logic based on truth and reference in the philosophical canon. In many cases, as noted above, the many-valued approaches proved to be first approximations to *extensions* or *enrichments* of classical systems rather than *replacements* for them. Łukasiewicz's concern for indeterminism is now addressed within tense logic; intuitionism is now seen, by most logicians, as providing a more refined analysis of concepts and proofs within classical mathematics rather than as challenging it. Of the major approaches discussed, free logic remains the area most likely to be adopted as a new standard approach, although it is possible that fuzzy logic or supervaluationism will become the standard in treatments of vagueness.

References

Borkowski, L. (ed.) (1970) *Jan Lukasiewicz: Selected Works*. Amsterdam: North-Holland.
Dummett, M. (1977) *Elements of Intuitionism*. Oxford: Oxford University Press.
Goguen, J. A. (1968–9) The logic of inexact concepts. *Synthese*, 19, 325–73.
Jacquette, D. (1996) *Meinongian Logic: The Semantics of Existence and Nonexistence*. The Hague: Walter de Gruyter.
Körner, S. (1966) *Experience and Theory: An Essay in the Philosophy of Science*. New York: Humanities Press.
Post, E. L. (1921) Introduction to a general theory of elementary propositions. In van Heijenoort (1967), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Cambridge, MA: Harvard University Press.
Zadeh, L. A. (1975) Fuzzy logic and approximate reasoning. *Synthese*, 30, 406–28.

Further Reading

Bolc, L. and Borowik, P. (1992) *Many-valued Logics: Theoretical Foundations*. Berlin: Springer.
Church, A. (1956) *Introduction to Mathematical Logic*. Princeton, NJ: Princeton University Press.
Dummett, M. (1991) *Logic as the Basis of Metaphysics*. Cambridge, MA: Harvard University Press.
Dunn, M. J. and Epstein, G. (1977) *Modern Uses of Multiple-Valued Logic*. Dordrecht: Reidel.
Gabbay, D. M. (1981) *Semantical Investigations in Heyting's Intuitionistic Logic*. Dordrecht: Reidel.
Heyting, A. (1966) *Intuitionism: An Introduction*. Amsterdam: North-Holland.
Heyting, A. (ed.) (1975) *L. E. J. Brouwer: Collected Works*. Amsterdam: North Holland.
Kleene, S. C. (1952) *Introduction to Metamathematics*. Princeton, NJ: D. van Nostrand.
Lambert, K. (1991) *Philosophical Applications of Free Logic*. New York: Oxford University Press.
Malinowski, G. (1993) *Many-Valued Logics*. Oxford: Clarendon Press.
Rescher, N. (1969) *Many-Valued Logics*. New York: McGraw Hill.

34

Many-Valued Logic

GRZEGORZ MALINOWSKI

1 When is a Logic Many-Valued?

The most natural and straightforward step towards the construction of a many-valued logic is to introduce logical values next to truth and falsity. Thereby, one has to reject the principle of bivalence, that every proposition has exactly one of the two logical values. Another, indirect way consists in challenging the classical laws concerning the sentence connectives and introducing non-truth-functional connectives into the language, among them the modal connectives of possibility and necessity. In either case the semantics adequate is different from the classical, that is Boolean, thus the logic under consideration is non-classical.

2 Roots, Motivations, and Early History

The roots of many-valued logics can be traced back to Aristotle (fourth century BC) who considered, within the modal framework, *future contingents* sentences. In Chapter IX of *De Interpretatione* Aristotle provides the time-honored sentence-example representing this category: 'There will be a sea-battle tomorrow.' The Philosopher from Stagira emphasizes the fact that future contingents are neither actually true nor actually false, which suggests the existence of the 'third' logical status of propositions.

The prehistory of many-valued logic falls on the Middle Ages. More serious attempts to create non-classical logical constructions, three-valued mainly, appeared only on the turn of the nineteenth century. The evaluation to what extent these different approaches by Duns Scott, William Ockham, Peter de Rivo and Hugh MacColl, Charles S. Peirce, Nicolai A.Vasil'ev were important for the topic is not easy. In most cases the division of the totality of propositions into three categories was supported by some considerations dealing with some modal or temporal concepts. Eventually, some criteria of the distinction were applied and the propositions mostly were grouped as either 'affirmative,' 'negative,' or 'indifferent.'

Philosophical motivations for logical many-valuedness may roughly be classified as *ontological* and *epistemic*. First of them focus on the nature of objects and facts, while the others refer the knowledge status of actual propositions. The 'Era of many-

valuedness' was finally inaugurated in 1920 by Łukasiewicz (1920) and Post (1920). The thoroughly successful formulations of many-valued logical constructions were possible in the result of an adaptation of the truth table method applied to the classical logic by Frege in 1879, Peirce in 1885 and others. The impetus thus given bore the Łukasiewicz and Post method of logical algebras and matrices. Apparently different proposals of the two scholars had quite different supports.

3 Łukasiewicz Three-Valuedness

Though 1920 is the year of publication of Łukasiewicz's article in an official journal *Ruch Filozoficzny* his finding was published as soon as March 7, 1918. In that paper Łukasiewicz enriched the set of the classical logical values 0 and 1 with an intermediate value 1/2 and laid down the principles of his calculus referring to Aristotle's argument. His *future contingent* proposition read "I shall be in Warsaw at noon on 21 December of the next year."

First Łukasiewicz's interpretation of the third logical value 1/2 was as a 'possibility' or 'indeterminacy.' Accordingly, the interpretation of negation and implication has been extended in the following tables:

x	$\neg x$
0	1
1/2	1/2
1	0

$\rightarrow$	0	1/2	1
0	1	1	1
1/2	1/2	1	1
1	0	1/2	1

(the truth tables of binary connectives $*$ are viewed as follows: the value of α is placed in the first vertical line, the value of β in the first horizontal line and the value of $\alpha * \beta$ at the intersection of the two lines).

The remaining standard connectives introduced through definitions

$\alpha \vee \beta = (\alpha \rightarrow \beta) \rightarrow \beta$
$\alpha \wedge \beta = \neg (\neg\alpha \vee \neg\beta)$
$\alpha \equiv \beta = (\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha)$.

have the tables:

$\vee$	0	1/2	1
0	0	1/2	1
1/2	1/2	1/2	1
1	1	1	1

$\wedge$	0	1/2	1
0	0	0	0
1/2	0	1/2	1/2
1	0	1/2	1

$\equiv$	0	1/2	1
0	1	1/2	0
1/2	1/2	1	1/2
1	0	1/2	1

A valuation of formulas in Łukasiewicz three-valued logic is any function v: *For* $\rightarrow$ {0,1/2, 1} of the set of all formulas *For* compatible with the above tables. A *tautology* is a formula which under any valuation v takes on the *designated* value 1.

The set $Ł_3$ of tautologies of three-valued logic of Łukasiewicz differs from *TAUT*. So, for instance, neither the law of the excluded middle, nor the principle of contradiction is in $Ł_3$. To see this, it suffices to assign $1/2$ for p: any such valuation also associates $1/2$ with *EM* and *CP*. The thorough-going refutation of these two laws was intended to codify the principles of indeterminism.

Another property of new semantics is that some classically inconsistent formulas are no more contradictory in $Ł_3$. One such formula:

$$(*) \quad p \equiv \neg p,$$

is connected with the famous Russell paradox of the 'set of all sets that are not their own elements.' *Russell's set* is defined by the equation

$$Z = \{x : x \notin x\}.$$

And the resulting paradox

$$Z \in Z \equiv Z \notin Z,$$

is a substitution of (*). Russell paradox ceases to be antinomy in $Ł_3$ since putting $1/2$ for p makes the formula true and therefore (*) is non-contradictory. Łukasiewicz found it a strong argument in favor of his three-valued logic.

4 Post Logics

Post's proposal was made on the margin of the completeness proof of the classical logic. It consists in defining n-valued (n finite) 'logic algebras' saving the classical property of functional completeness of the set of connectives (the property permits the definition of all other possible connectives), cf. Post (1920, 1921).

Following *Principia Mathematica* Post takes the negation ($\neg$) and disjunction ($\vee$) connectives as primitive. For any natural $n \geq 2$ he considers a linearly ordered set

$$Pn = \{t_1, t_2, \ldots, t_n\},$$

$t_n < t_j$ iff $i < j$, with two operations: unary *rotation* (or *cyclic negation*) $\neg$ and binary *disjunction* $\vee$, where

$$\neg t_i = \begin{cases} t_{i+1} \text{ if } i \neq n \\ t_1 \text{ if } i = n \end{cases} \qquad t_i \vee t_j = t_{i\max\{i,j\}}$$

Thus, for example for $n = 4$ the truth tables of these connectives are the following:

x	$\neg x$
t_1	t_2
t_2	t_3
t_3	t_4
t_4	t_1

$\vee$	t_1	t_2	t_3	t_4
t_1	t_1	t_2	t_3	t_4
t_2	t_2	t_2	t_3	t_4
t_3	t_3	t_3	t_3	t_4
t_4	t_4	t_4	t_4	t_4

It is easy to see that for $n = 2$ Post logic coincides with the negation–disjunction version of the classical logic: the set $P_2 = \{t_1, t_2\}$ may be identified as containing 0 and 1, respectively, and then the Post negation and disjunction are isomorphic variants of the classical connectives on P_2. The relation to *CPC* breaks for $n > 2$. In all these cases truth tables of negation are not compatible with the classical one is due to the fact that t_1 always corresponds to 0 and t_n to 1. Though $\neg t_n = t_1$, $\neg t_1$ equals t_2 and thus then is not t_n.

Post considers t_n as the distinguished value. Among special laws of all its logics ($n > 2$) the following many-valued counterpart of the classical law of the excluded middle

$$p \vee \neg p \vee \neg\neg p \vee \ldots \underbrace{\neg\neg \ldots}_{(n-1) \text{ times}} \neg p.$$

deserves attention. The absence of the counterparts of some other classical tautologies follows directly from the properties of negation.

The most important property of Post algebras is their functional completeness: by means of the two primitive functions, every finite-argument function on P_n can be defined. In particular, then, the constant functions may also be defined and hence the 'logical values' $t_1, t_2, \ldots, t_n$.

Post suggests interpreting the elements of P_n as $(n-1)$-element-tuples $P = (p_1, p_2, \ldots, p_{n-1})$ of ordinary two-valued propositions $p_1, p_2, \ldots, p_{n-1}$ subject to the condition that the true propositions are listed before the false. Then

($\neg$) $\neg P$ if formed by replacing the first false element by its denial, otherwise it is a sequence of false propositions.

($\vee$) When $P = (p_1, p_2, \ldots, p_{n-1})$ and $Q = (q_1, q_2, \ldots, q_{n-1})$, then $P \vee Q = (p_1 \vee q_1, p_2 \vee q_2, \ldots, p_{n-1} \vee q_{n-1})$.

For n = 4 one gets the following 3-tuples:

$$\begin{array}{ll} (0, 0, 0) & t_1 \\ (1, 0, 0) & t_2 \\ (1, 1, 0) & t_3 \\ (1, 1, 1) & t_4. \end{array}$$

This interpretation shows that the values in different Post logics should be understood differently.

5 Łukasiewicz Logics

In 1922 Łukasiewicz generalized his three-valued logic and defined the family of many-valued logics, both finite and infinite-valued, see Łukasiewicz (1970: 140). The set of logical values of n-valued logic for any natural $n \geq 2$ is

$$L_n = \{0, 1/(n-1), 2/(n-1), \ldots, (n-2)/(n-1), 1\}.$$

First infinite logic is based on the set of all fractions in the real interval [0,1],

$$L_{\aleph_0} = \{s/w: 0 \leq s \leq w, s, w \in N \text{ and } w \neq 0\}$$

and the second on the whole interval [0,1], $L_{2^{\aleph_1}} = [0,1]$. In all these cases 1 is taken as the only designated value and the connectives are defined as follows:

1. $\neg x = 1 - x$
   $x \rightarrow y = min(1, 1 - x + y)$
2. $x \vee y = (x \rightarrow y) \rightarrow y = max(x, y)$
   $x \wedge y = \neg(\neg x \vee \neg y) = min(x, y)$
   $x \equiv y = (x \rightarrow y) \wedge (y \rightarrow x) = 1 - |x - y|$.

To give an idea of what truth tables of finite valued logics look like, we now show the tables of negation and implication in the five-valued logic of Łukasiewicz:

x	$\neg x$
0	1
1/4	3/4
2/4	2/4
3/4	1/4
1	0

$\rightarrow$	0	1/4	2/4	3/4	1
0	1	1	1	1	1
1/4	3/4	1	1	1	1
2/4	2/4	3/4	1	1	1
3/4	1/4	2/4	3/4	1	1
1	0	1/4	2/4	3/4	1

Łukasiewicz matrices have this exceptional property that in all of them the set {0,1} is closed with respect to all connectives. This together with the fact that the tables for all usual connectives on this set coincide with the classical truth tables yields the fact that the set of all tautologies of every Łukasiewicz logic, $Taut_n$, is a subset of the set of tautologies of the CPC which actually is $Taut_2$. The inclusion

$$Taut_n \subseteq Taut_2$$

extends to the famous Lindenbaum's condition on mutual relations in the family of finite Łukasiewicz logic. Namely, that for any natural n, m (both ≥ 2)

$$Taut_n \subseteq Taut_m \quad \text{if and only if } m - 1 \text{ is a } \textit{divisor} \text{ of } n - 1 .$$

Infinite Łukasiewicz matrices have the same set of tautologies equal to the intersection of the contents of all finite matrices: $\cap \{ Taut_n : n \geq 2,\ n \in N \}$.

Contrary to Post none of the Łukasiewicz logics $Ł_n$ ($n \neq 2$) is functionally complete since no constant function except 0 or 1 is definable. Adding all suitable constants to the stock of connectives makes each finite logic complete. McNaughton (1951) formulated and proved an ingenious definability criterion for Łukasiewicz matrices, both finite and infinite, showing the mathematical beauty of Łukasiewicz's logic constructions.

As early as 1931 Wajsberg gave an axiomatization of $Ł_3$. Taking the rules *MP* and *SUB* he established the four axioms

W1 $p \to (q \to p)$
W2 $(p \to q) \to ((q \to r) \to (p \to r))$
W3 $(\neg p \to \neg q) \to (q \to p)$
W4 $((p \to \neg p) \to p) \to p$.

Since the other Łukasiewicz connectives are definable, the axiomatizability result obviously applies to the whole $Ł_3$. It is worth noting that W1–W4 was the first axiom system of many-valued logics. Still earlier, in 1930, Łukasiewicz conjectured that his $\mathcal{N}_0$-valued logic was axiomatizable (Łukasiewicz and Tarski 1930) by five axioms: W1, W2, and

L3 $((p \to q) \to q)\ ((q \to p) \to p)$
L4 $(\neg p \to \neg q)\ (q \to p)$
L5 $((p \to q)\ (q \to p))\ (q \to p)$.

The response came only in 1958 with two works showing the dependence, and thus, the eliminability of L5. In addition, two further completeness proofs, one syntactic and the other algebraic, were derived see Rose and Rosser (1958) and Chang (1959).

6 Kleene and Bochvar Logics

In 1938 two similar, though independent, three-valued systems of logic were invented by Kleene and Bochvar. The epistemic arguments behind their construction relate to indeterminacy or to meaninglessness.

Kleene's (1938) main assumption is that there are propositions whose logical truth (t) or falsity (f) is either undefined, undetermined by means of accessible algorithms, or is not essential. The third value of undefiniteness (u) is reserved for this category of propositions. Further to that the tables of the standard connectives save the classical behavior towards t and f and looks like:

α	$\neg\alpha$
f	t
u	u
t	f

$\to$	f	u	t
f	t	t	t
u	u	u	t
t	f	u	t

$\vee$	f	u	t
f	f	u	t
u	u	u	t
t	t	t	t

$\wedge$	f	u	t
f	f	f	f
u	f	u	u
t	f	u	t

$\equiv$	f	u	t
f	t	u	f
u	u	u	u
t	f	u	t

Kleene's logic has no tautologies. This, somewhat striking, feature follows from the fact that any valuation which assigns u to every propositional variable also assigns u to any formula.

In 1952, in his monograph *Introduction to Metamathematics* Kleene refers to the connectives of his 1938 logic as strong and introduces another set of weak connectives: retaining the negation and equivalence he defines the three others by the tables

$\rightarrow$	f	u	t
f	t	u	t
u	u	u	u
t	f	u	t

$\vee$	f	u	t
f	f	u	t
u	u	u	u
t	t	u	t

$\wedge$	f	u	t
f	f	u	f
u	u	u	u
t	f	u	t

The novel truth tables are to describe the employment of logical connectives in respect of those arithmetical propositional functions whose decidability depends on the effective recursive procedures. They are constituted according to the rule of saying that any single appearance of u results in the whole context taking u. The original arithmetic motivation states that indeterminacy occurring at any stage of computation makes the entire procedure undetermined. While the first Kleene logic was made to render the analysis of partially defined propositional functions possible, the second was inspired by the studies within the mathematical theory of recursion, see Kleene (1952).

Bochvar's (1938) approach is directed towards solving paradoxes emerging with the classical logic and set theory based on it. The propositional language of Bochvar logic has two levels corresponding to the object language and to metalanguage. They both contain connectives being counterparts of negation, implication, disjunction, conjunction, and equivalence. The *internal* connectives are conservative generalizations of the classical ones, in the sequel they will be denoted similarly. The *external* connectives are devised to characterize the relationship between logical values of propositions. Both sets are initially described using the values corresponding to two kinds of meaningful sentences that is of truth (t) and falsity (f), and the third value u reserved for meaningless sentences.

The tables of internal connectives have been set according to the rule: 'every compound proposition including at least one meaningless component is meaningless, in other cases its value is determined classically.' Consequently, the internal Bochvar logic coincides with the weak Kleene logic.

The external 'metalinguistic' connectives are supposed to express the predicates '. . . is true' and '. . . is false' and have the following 'meanings':

external nexgation:	$\neg^* \alpha$	'α is false'
external implication:	$\alpha \rightarrow^* \beta$	'if α is true, then β is true'
external disjunction:	$\alpha \vee^* \beta$	'α is true or β is true'
external implication:	$\alpha \wedge^* \beta$	'α is true and β is true'
external implication:	$\alpha \equiv^* \beta$	'α is true iff β is true'

Their truth tables are as follows:

α	$\neg^*\alpha$
f	t
u	t
t	f

$\rightarrow^*$	f	u	t
f	t	t	t
u	u	t	t
t	t	f	t

$\vee^*$	f	u	t
f	f	f	t
u	f	f	t
t	t	t	t

$\wedge^*$	f	u	t
f	f	f	f
u	f	f	f
t	f	f	t

$\equiv^*$	f	u	t
f	t	t	f
u	t	t	f
t	f	f	t

As a result, the external logic is a 'three-valued' version of the classical logic. This is due to the fact that the truth tables of all external connectives 'identify' the values u and f, whereas the behavior of these connectives with regard to f and t is classical.

7 Towards a General Framework

With a view to unification, Rosser and Turquette (1952) established some special *standard conditions* that make finitely many-valued logics resemble the classical propositional logic. This, on a certain level of investigation, permitted the simplification or solving of some metalogical questions, such as axiomatization and the extension to predicate logics.

Assume that $n \geq 2$ is a natural number and $1 \leq k < n$. Let $E_n = \{1, 2, \ldots, n\}$ be the set of logical values and $D_k = \{1, 2, \ldots, k\}$ as the set of designated values. Rosser and Turquette assume that the natural number ordering conveys decreasing degrees of truth. So, 1 always refers to 'truth' and n takes the role of falsity.

Next come the conditions concerning propositional connectives, which have to represent negation ($\neg$), implication ($\rightarrow$), disjunction ($\vee$), conjunction ($\wedge$), equivalence ($\equiv$) and special one-argument connectives $j_1, \ldots, j_n$. The respective connectives satisfy the *standard conditions* if for any $x, y \in E_n$ and $i \in \{1, 2, \ldots, n\}$

$\neg x \in D_k$	if and only if	$x \notin D_k$
$x \rightarrow y \notin D_k$	if and only if	$x \in D_k$ and $y \notin D_k$
$x \vee y \notin D_k$	if and only if	$x \in D_k$ or $y \in D_k$
$x \wedge y \in D_k$	if and only if	$x \in D_k$ and $y \in D_k$
$x \equiv y \in D_k$	if and only if	either $x, y \in D_k$ or $x, y \notin D_k$
$j_i(x) \in D_k$	if and only if	$x = i$.

Any many-valued logic $L_{n,k}$ having standard connectives as primitive or definable is called *standard*.

The class of many-valued logics, whose connectives fulfill standard conditions is quite large. It contains, 'obviously,' all Post logics since they are functionally complete. All finite Łukasiewicz logics are also standard; note that the mapping $f(x) = n - (n - 1)$ transposes the original values $\{0, 1/(n-1), 2/(n-1), \ldots, n - 2/(n-1), 1\}$ onto $\{1, 2, \ldots, n\}$. A moment's reflection shows that original Łukasiewicz disjunction and conjunction satisfy standard conditions. In turn, the other required connectives including j's, $j_i(x) = 1$ iff $x = i$, are definable.

Using their framework, Rosser and Turquette positively solved the problem of axiomatizability of known systems of many-valued logic, including n-valued Łukasiewicz and Post logics. Actually, any $\{\rightarrow, j_1, j_2, \ldots, j_n\}$ – standard logic $L_{n,k}$ is axiomatizable by means of the rule MP and SUB and the following set of axioms:

A1 $p \rightarrow (q \rightarrow p)$
A2 $(p \rightarrow (q \rightarrow r)) \rightarrow (q \rightarrow (p \rightarrow r))$
A3 $(p \rightarrow q) \rightarrow ((q \rightarrow r) \rightarrow (p \rightarrow r))$
A4 $(j_i(p) \rightarrow (j_i(p) \rightarrow q)) \rightarrow (j_i\,(p) \rightarrow q)$
A5 $(j_n(p) \rightarrow q) \rightarrow ((j_{n-1}(p) \rightarrow q)\,(\ldots \rightarrow\ ((j_1(p) \rightarrow q) \rightarrow q) \ldots))$
A6 $j_i(p)\ p \rightarrow$ for $i = 1, 2, \ldots, k$
A7 $j_{i(r)}(p_r) \rightarrow (j_{i(r-1)}\,(p_{r-1}) \rightarrow (\ldots \rightarrow (j_{i(1)}(p_1) \rightarrow j_f(F(p_1, \ldots, p_{r-1}, p_r))) \ldots))$
where $f = f(i(1), \ldots, i(r))$;

symbols F and f in A7 represent, respectively, an arbitrary connective and the function associated with it.

The first three axioms describe the properties of pure classical implication sufficient, among others, to get the deduction theorem in its classical version. The remaining axioms bridge, due to the properties of j connectives and of the implication, the semantic and syntactic properties. Checking the soundness of the axioms is easy and is heavily based on procedures known from classical logic. The completeness proof, however, requires much calculation and involves a complicated induction.

8 On Quantification

Many-valued predicate calculi are usually built along the classical pattern. In that case a first-order language with two standard quantifiers, general $\forall$ and existential $\exists$, are considered. Mostly, the starting point is the substitutional conception of quantifiers according to which $\forall$ and $\exists$ are (infinite) generalizations of conjunction and disjunction, respectively. Accordingly, for a finite domain $U = \{a_1, a_2, \ldots, a_n\}$, the commutative and associative connectives of conjunction ($\wedge$) and disjunction ($\vee$):

$\forall x F(x) \equiv_U F(a_1) \wedge F(a_2) \wedge \ldots \wedge F(a_n)$
$\exists x F(x) \equiv_U F(a_1) \vee F(a_2) \vee \ldots \vee F(a_n)$

($\equiv_U$ means the equivalence of the formulae at any interpretation in U, $a_1, a_2, \ldots, a_n$ being nominal constants ascribed to the objects of the domain). In finite-valued logical calculi constructed upon linear matrices, quantifiers are defined 'directly' through algebraic functions related to the above-mentioned connectives. Thus, for example, for finite Łukasiewicz and Post logics, for any interpretation f in a domain U

$f(\forall x f(x)) = min\{f(F(a)) : a \in U\}$
$f(\exists x F(x)) = max\{f(F(a)) : a \in U\}$.

For other calculi, the semantic description of quantifiers may vary. Thus, for example, the clauses defining quantifiers in the first-order Bochvar logic should as follows:

$$f(\forall x F(x)) = \begin{cases} \mathrm{t} & \text{when } f(F(a)) = \mathrm{t} \text{ for every } a \in U \\ \mathrm{u} & \text{when } f(F(a)) = \mathrm{u} \text{ for some } a \in U \\ \mathrm{f} & \text{otherwise} \end{cases}$$

$$f(\exists x F(x)) = \begin{cases} \mathrm{f} & \text{when } f(F(a)) = \mathrm{f} \text{ for every } a \in U \\ \mathrm{u} & \text{when } f(F(a)) = \mathrm{u} \text{ for some } a \in U \\ \mathrm{t} & \text{otherwise.} \end{cases}$$

Axiomatic systems of many-valued predicate logics are usually built as extensions of axiom systems of the grounds of propositional calculi in a similar way to classical logic, see Rasiowa and Sikorski (1963) and Rasiowa (1974). Proofs of completeness for finitely-valued calculi do not, in general, create difficulties. Axiomatizability of several important calculi of this kind are assured by Rosser and Turquette's result extending the standard condition's approach to quantifiers, see Rosser and Turquette (1952).

Introducing quantifiers to logics with infinitely many values in the semantical plane may be problematic. Thus, for example, applying the above-mentioned procedure to the $\mathcal{N}_0$-valued Łukasiewicz logic is impossible since in the case when U is infinite it may happen that the set $\{f(F(a)) : a \in U\}$ does not contain the least or the greatest element and therefore *min* and *max* functions cannot be used in the definition. In turn, in the $\mathcal{N}_1$-valued Łukasiewicz logic, the interpretations of quantifiers are introduced provided that for any interpretation in a non-empty domain U

$$f(\forall x F(x)) = inf\{f(F(a)) : a \in U\}$$
$$f(\exists x F(x)) = sup\{f(F(a)) : a \in U\},$$

see Mostowski (1961). However, it appeared that $\mathcal{N}_1$-valued predicate calculus thus obtained is not axiomatizable, Scarpelini (1962). The problem of the completeness of this logic appeared extremely complex and the experience gained while attempting to constitute such a proof raised the so-called *continuous model theory* (see Chang and Keisler 1966).

Rosser and Turquette (1952) invented a general theory of quantification for a class of finitely many-valued logics. Starting from the intuition that ordinary quantifiers are functions on the set of pairs (x, F), where x is a nominal variable and F a formula, with values in the set of formulae, Rosser and Turquette defined a *generalized quantifier* as any formula of the form:

$$Q(x_1, x_2, \ldots, x_m, F_1, F_2, \ldots, F_t),$$

where $x_1, x_2, \ldots, x_m$ are nominal variables and $F_1, F_2, \ldots, F_t$ formulae built from predicates, nominal and propositional variables, and connectives.

Carnielli (1987) admits a very general class of *distribution quantifiers* defined using multiple-valued matrices as functions mapping subsets of the set of logical values into values. This ingenious construction also directly extends a standard approach to classical quantifiers.

9 Interpretation and Justification

Scholars of the philosophical foundation of logic widely criticized many-valued constructions. The first was the explanation of the logical value 1/2 in Łukasiewicz (1920) resorting to 'future contingents' and a 'possibility' or undetermination of the 0–1 status of propositions. As shown by Gonseth (1941), such interpretation is incompatible with other principles of Łukasiewicz. Whenever α is undetermined, so is $\neg\alpha$ and then $\alpha \wedge \neg\alpha$ is undetermined. That contradicts our intuition since, independently of α's content, $\alpha \wedge \neg\alpha$ is false. The upshot discovers that Łukasiewicz interpretation neglects the mutual dependence of some 'possible' propositions.

Haack (1978) analyses Łukasiewicz's way of avoiding the fatalist conclusion derived from the assumption that the contingent statement 'I shall be in Warsaw at noon on 21 December of the next year' is either true or false in advance of the event. She remarks that this way of rejecting bivalence is wrong, since it depends on a modal fallacy of arguing from "It is necessary that (if α, then β)" to " If α, then it is necessary that β." Urquhart (1986) sees the third logical value as the set $\{0,1\}$ of two 'potential' classical values of a future contingent sentence and defines the implication as getting all possible values of implication. Thus, for example the implication having 0 as antecedent always takes value 1, the implication from 1 to $\{0,1\}$ takes $\{0,1\}$ as the value and the implication from $\{0,1\}$ to $\{0,1\}$ has the value $\{0,1\}$. The last point is inconsistent with the Łukasiewicz stipulation, since the output has to be 1. Therefore, Urquhart claims, the Łukasiewicz table is wrong. It may be of interest that the connective derived by Urquhart is the Kleene strong implication.

Reichenbach (1944) argued that adoption of three-valued logic would provide a solution to some problems raised by quantum mechanics. In order to avoid 'causal anomalies,' Reichenbach presents an extended version of the Łukasiewicz logic, adding further negation and implication connectives. He refers to the third logical value as 'indeterminate' and assigns it to anomalous statements of quantum mechanics. The weak point of Reichenbach's proposal is that certain laws are also classified as 'indeterminate', such as for example, the principle of energy.

The mathematical probability calculus in its simplest form resembles many-valued logic. Łukasiewicz, before 1918, invented a concept of *logical probability* referring to propositions and not to events, see Łukasiewicz (1913). The continuators tried to create a many-valued logic within which logical probability could find a satisfactory interpretation, see, for example, Zawirski (1934), Reichenbach (1935). The Reichenbach–Zawirski theory is based on the assumption that there is a function Pr ranging over the set of propositions of a given *standard* propositional language, with values from the real interval $[0,1]$, such that

P1 $0 \leq pr(p) \leq 1$
P2 $Pr(p \vee \neg p) = 1$
P3 $Pr(p \vee q) = Pr(p) + Pr(q)$ if p and q are mutually exclusive ($Pr(p \in q) = 0$)
P4 $Pr(p) = Pr(q)$ when p and q are logically equivalent.

Such probability, however, does not fit any ordinary extensional many-valued logic. Identifying the logical value $v(p)$ with the $Pr(p)$ for $Pr(p) = 1/2$ from P2 and P3 we would, for example, get that

$$1/2 \vee 1/2 = Pr(p \vee -p) = 1 \quad \text{and} \quad 1/2 \vee 1/2 = Pr(p \vee p) = Pr(p) = 1/2.$$

A very convincing interpretation of the $\mathcal{N}_0$-valued Łukasiewicz logic of Giles (1974) is based on a dispersive physical interpretation of standard logical language. Each prime proposition in a physical theory is associated through the rules of interpretation with a certain experimental procedure terminating in one of the two possible outcomes, 'yes' and 'no.' The tangible meaning of a proposition is related to the observers and expressed in terms of probability. In the case of prime propositions it is determined from the values of probability of success ascribed by observers in respective experiment, whereas in the case of compound propositions it is determined from the rules of obligation formulated in the nomenclature of *dialogue logic*. The inductive clauses for the connectives, and later for quantifiers, translated back to subjective probability function *pr* conform to the original Łukasiewicz definitions. The set of tautologies of the dialogue logic, that is of formulas to which any valuation assigns non-positive risk-value, coincides with the set of tautologies of the infinite-valued Łukasiewicz logic.

Elimination of the Russell paradox was among the expectations of Łukasiewicz and Bochvar. An interesting work on Łukasiewicz logics related to the question of the unlimited consistency of the comprehension axiom, that is a first-order formula with $\in$ stating the existence of all sets bearing logically expressible properties, was done. It started with Moh Shaw Kwei's (1954) result on the impossibility of the use of finite systems for the purpose, and continued in the 1960s after Skolem (1957) put forward a hypothesis that *CA* was consistent in $\mathcal{N}_1$-valued Łukasiewicz logic. Though several interesting results have been obtained, the question, in its full generality, still remains open.

Scott (1973) replaces many logical values by many valuations using the truth t and falsity f. A definite number of bivalent valuations generates a partition of the set of propositions into types (indexes) corresponding to the original logical values – Scott refers to them as to *indexes*. An n-element set of valuations can thus induce maximally 2^n types. The actual number of types depends on limiting conditions imposed on valuations. An accurate choice of these conditions leads to a relatively simple characterization of the connectives of the logic under consideration. Applying his method, Scott gets a description of the n-valued Łukasiewicz negation and implication connectives through an $(n-1)$-element set of valuations $\{v_0, v_1, \ldots, v_{n-2}\}$. The equalities of the form '$v_i(\alpha)$' should be read as '(the statement) α is true to within the degree i.' Consequently, the numbers $0, 1, \ldots, n-2$ stand for *degrees of error in deviation from the truth*. Degree 0 is the strongest and corresponds to 'perfect' truth or no error: all Łukasiewicz tautologies are schemes of the statements having 0 as their degree of error. The measure of error of the Łukasiewicz implication expresses the amount of *shift* of error between the degree of hypothesis and that of the conclusion.

Urquhart (1973) gave an interpretation motivated by the logic of tenses. The core of it is the relation $\vdash$ between natural numbers of $S_n = \{0, 1, \ldots, n-2\}$ and formulas. '$x \vdash \alpha$' expresses that 'α is true at x' satisfies

If $x \vdash \alpha$ and $x \leq y \in S_n$, then $y \vdash \alpha$.

Adopting $\vdash$ to particular finite-valued logic requires specifying n, the language, and providing recursive conditions which establish the meaning of connectives. Accordingly, each case results in some Kripke-style semantics with finite number of 'reference points' S_n. For Łukasiewicz and Post logics, Urquhart suggests a temporal interpretation: 0 is the present moment and all other points of reference are future moments. A temporal way of understanding Łukasiewicz negation and implication exhibits the sources of difficulties in getting plausibly intuitive interpretation of many-valued Łukasiewicz logic. Urquhart eventually indicates clauses which 'natural' connectives of negation and implication should satisfy.

10 Applications

Perhaps the most natural of all was the use of many-valuedness to the analysis of vagueness, inexactness, and the paradoxes, see for example Williamson (1994). This application finally gave an impetus to fuzzy set theory and, ultimately to the theory of fuzzy logics, see Zadeh (1975). Zadeh (1965) defines a fuzzy set A of a given domain U as an abstract object characterized by generalized characteristic function U_A with values in the real set $[0,1]$:

$$U_A : U \rightarrow [0,1].$$

The values of U_A are degrees of membership of elements of U to a fuzzy set A. The extreme values denote, respectively, not belonging to A and the entire membership of A. So, an ordinary set is a special fuzzy set, having only 0 and 1 as possible degrees of membership.

Fuzzy sets model inexact predicates appearing in natural languages. The values of generalized characteristic functions are logical values of propositions obtained from the predicates serving as a basis for a given fuzzy set. Consequently, with fuzzy set algebra of fuzzy (sub)sets of a given domain U can be associated with an uncountable many-valued logic. The inclusion and the operations of a (fuzzy) complement $-$, union $\cup$, and intersection $\cap$ are then stated by means of 'generalized' set-theoretic predicate $\in$ and logical constants (implication, negation, disjunction, and conjunction, respectively).

The choice of the basic logic is to a great extent prejudiced. It occurred that the $\mathcal{N}_1$-valued logic of Łukasiewicz is appropriate and it still remains favorite in the field. The early accounts yielded the (first) understanding of the term 'fuzzy logic' as a certain class of infinitely-valued logics, with Łukasiewicz logics in the foreground.

A typical case of modeling an inexact predicate within the above framework is the following attempt of modeling the classical paradox of a *bald man*. Let us take the two following, intuitively acceptable, propositions:

(1) A man with 20,000 hairs on his head is not bald
(2) A man who has one hair less than somebody who is not bald is not bald as well.

Applying the detachment rule 20,000, we get, by (1) and (2), that a man with no hair is not bald either. The paradox will vanish when the logical value of any proposition 'A

man with n hair is not bald' is identified with the degree of membership of a man with n hairs to a fuzzy set 'not-bald.' Then, (2) will have a logical value less than 1, say $1 - \varepsilon$, where $\varepsilon > 0$. And, if in basic logic we use Łukasiewicz's implication. Then as a result of 20,000 derivations we will obtain a proposition of the logical value amounting to $1 - 20{,}000\varepsilon$, thus *practically* false.

Zadeh's (1975) conception of a *fuzzy logic* conveyed the belief that thinking in terms of fuzzy sets is a typical feature of human perception. Fuzzy logic identifies predicates with fuzzy subsets of a given universe and logical values with fuzzy subsets of the set of values of the basic logic. The logical values are labeled *linguistic* entities and, similarly as predicates, may be modified by the so-called *hedges*. Finally, the procedure of linguistic approximation compensates for the lack of closure of the object language and the closure of the set of logical values onto logical connectives. Fuzzy logic is now an autonomic discipline. It seeks to formulate several rules of approximate inference.

Zadeh's fuzzy approach has found its place among accepted methods of artificial intelligence, in computer science and steering theory. It confirmed its usefulness due to reliable applications; see Turner (1984).

The use of many-valued matrices to the formalization of intensional functions, the matrix approximation of syntactically founded non-classical logics and the testing of independence of axioms are worth mentioning. The first use was already suggested by Łukasiewicz, who insisted on the formalization of possibility and necessity within the three-valued logic (see Section 2) and several years later proposed a four-valued system of modal logic in Łukasiewicz (1953). This line of approach has been in some way continued since the algebraic interpretations of Łukasiewicz and Post logics incorporated 'modal' functions in a form of the Boolean-valued endomorphisms. However, from the philosophical point of view these finite-valued interpretation of modalities have no particular value (since as already in 1940 Dugundji proved, no reasonable system of modal logic may be finite-valued), the role of their counterparts in Post algebras occurred which were crucial for the Computer Science applications.

Łoś (1948) showed that, under some reasonable assumptions, the formalization of functions of the kind 'John believes that p' naturally leads to the many-valued interpretation of the belief operators within the scope of the classical logic system. The model situation considered is the case of two persons, who do not agree on all the issues, which may be expressed in propositions. One then obtains four possible evaluations in terms of pairs of classical logic values, that is the truth or falsity, which divides the set of all propositions into four types (or classes) ultimately corresponding to non-classical values. The connectives of negation and implication defined 'naturally,' in reference to their classical counterparts in parallel use for every person, also behave classically. Accordingly, we fall in the four-valued version of *CPC*. The shifting of approach onto the case with more persons results in other formal many-valued interpretations of the classical logic with additional operators. Łoś's construction shows that it is possible to get a many-valued interpretation of some special intensional functions simultaneously adhering to the intuition of bivalence. Since many-valuedness thus received reflects certain relation of two different arguments, a person and a proposition, it has to be classified as untypical semantic correlate.

The successful use of classical logic and Boolean algebras in switching theory and in computer science became established. The algebraic approach enables the applica-

tion of several techniques for the analysis, synthesis, and minimalization of multiplex networks. And, as early as the 1950s, interests centered also on possibility of the use of many-valued logics for similar purposes. These interests brought about the birth of several techniques for the analysis and synthesis of electronic circuits and relays based mainly on Moisil's and Posts algebras, see for example Rine (1977). The *practical* switchover of two oppositely oriented contacts positioned in parallel branches of a circuit, which have to change their positions simultaneously is the simplest possible electronic circuit to consider within a three-valued framework. Namely, there are good reasons to drop the idealistic assumption affecting the circuit, for example using relays, would really change the positions of both contacts instantly, that is that the circuit would pass from state 1 to state 0. Then, obviously, we get a third state that might also obtain. A generalization of the outlined construction for the case of any number of contacts similarly results in n states. Finally, getting a description of a network composed of such switchovers is performed using Moisil algebras, that is Łukasiewicz n-valued algebras and Post algebras. The most important advantage of the many-valued approach is the possibility of eliminating switching disturbances through the algebraic synthesis of the networks, see, for example, Moisil (1972).

Post algebras found an important application in the systematization of theoretical research concerning programs and higher level programming languages which contain instruction branching programs – the constants $e_0, e_1, \ldots, e_{n-1}$ of Post algebra are then interpreted as devices which keep track of which appropriate branching conditions $W_0, W_1, \ldots, W_{n-1}$. Further to this, Post algebras of order ω^+ form a semantic base for an ω^+-valued extension of algorithmic logic adapted to arbitrary 'wide' branching programs, see Rasiowa (1977).

References

Bochvar, D. A. (1938) Ob odnom trêhznačnom isčislénii i égo priménénii k analizu paradosov klassičéskogo rassširennogo funkcjonal'nogo isčislénia [On a three-valued calculus and its application to analysis of paradoxes of classical extended functional calculus]. *Matématičéskij Sbornik*, 4, 287–308.

Carnielli, W. A. (1987) Systematization of finite many-valued logics through the method of tableaux. *Journal of Symbolic Logic*, 52 (2), 473–93.

Chang, C. C. (1959) A new proof of the completeness of the Łukasiewicz axioms. *Transactions of the American Mathematical Society*, 93, 74–80.

Chang, C. C. and Keisler, H. J. (1966) *Continuous Model Theory*. Princeton, NJ: Princeton University Press.

Giles, R. (1974) A non-classical logic for physics. *Studia Logica*, 33, 397–416.

Gonseth, F. (ed.) (1941) *Les entretiens de Zurich sur les fondements et la méthode des sciences mathématiques 6–9 décembre 1938*. Zurich.

Haack, S. (1978) *Philosophy of Logics*. Cambridge: Cambridge University Press.

Kleene, S. C. (1938) On a notation for ordinal numbers. *Journal of Symbolic Logic*, 3, 150–5.

Kleene, S. C. (1952) *Introduction to Metamathematics*. Amsterdam: North-Holland.

Łoś, J. (1948) Logiki wielowartościowe a formalizacja funkcji intensjonalnych (Many-valued logics and the formalization of intensional functions). *Kwartalnik Filozoficzny*, 17, 59–78.

Łukasiewicz, J. (1913) *Die logischen Grundlagen der Wahrscheinlichkeitsrechnung*. Kraków; English tr. Logical foundations of probability theory. In L. Borkowski (ed.), *Selected Works* (pp. 16–63). Amsterdam: North-Holland.

Łukasiewicz, J. (1920) O logice trójwartościowej. *Ruch Filozoficzny*, 5, 170–1. English tr. On three-valued logic. In L. Borkowski (ed.), *Selected Works* (pp. 87–8). Amsterdam: North-Holland.

Łukasiewicz, J. (1953) *A system of modal logic. Journal of Computing Systems*, 1, 111–49.

Łukasiewicz, J. (1970) *Selected Works*. Amsterdam: North-Holland.

Łukasiewicz, J. and Tarski, A. (1930) Untersuchungen über den Aussagenkalkül. *Comptes rendus des séances de la Société des Sciences et des Lettres de Varsovie Cl. III*, 23, 30–50.

McNaughton, R. (1951) A theorem about infinite-valued sentential logic. *Journal of Symbolic Logic*, 16, 1–13.

Moh Shaw-Kwei (1954) Logical paradoxes for many-valued systems. *Journal of Symbolic Logic*, 19, 37–40.

Moisil, G. (1972) *Essais sur les logiques non-chrisipiennes*. Bucharest: Editions de l'Académie de la Republique Socialiste de Roumanie.

Mostowski, A. (1961) Axiomatizability of some many-valued predicate calculi. *Fundamenta Mathematicae*, 50, 165–90.

Post, E. L. (1920) Introduction to a general theory of elementary propositions. *Bulletin of the American Mathematical Society*, 26, 437.

Post, E. L. (1921) Introduction to a general theory of elementary propositions. *American Journal of Mathematics*, 43, 163–85.

Rasiowa, H. (1974) *An Algebraic Approach to Non-classical Logics*. Amsterdam: North-Holland.

Rasiowa, H. (1977) Many-valued algorithmic logic as a tool to investigate programs. In J. M. Dunn and G. Epstein (eds.), *Modern Uses of Multiple-Valued Logic* (pp. 79–102). Dordrecht: Reidel.

Rasiowa, H. and Sikorski, R. (1963) *The Mathematics of Metamathematics*. Warsaw: PWN.

Reichenbach, H. (1935) *Wahrscheinlichkeitslehre*. Leiden; English tr. *The Theory of Probability*. Berkeley: University of California Press, 1949.

Reichenbach, H. (1944) *Philosophical Foundations of Quantum Mechanics*. Berkeley and Los Angeles: University of California Press.

Rescher, N. (1969) *Many-valued Logic*. New York: McGraw-Hill.

Rine, D. C. (ed.) (1977) *Computer Science and Multiple-valued Logic: Theory and Applications*. Amsterdam: North-Holland.

Rose, A. and Rosser, J. B. (1958) Fragments of many-valued statement calculi. *Transactions of the American Mathematical Society*, 87, 1–53.

Rosenbloom, P. C. (1942) Post algebra. I. Postulates and general theory. *American Journal of Mathematics*, 64, 167–88.

Rosser, J. B. and Turquette, A. R. (1952) *Many-valued Logics*. Amsterdam: North-Holland.

Scarpelini, B. (1962) Die Nichtaxiomatisierbarkeit des uendlichtwertigen Prädikatenkalküls von Łukasiewicz. *Journal of Symbolic Logic*, 17, 159–70.

Scott, D. (1973) Background to formalisation. In H. Leblanc (ed.), *Truth, Syntax and Modality* (pp. 244–73). Amsterdam: North-Holland.

Skolem, T. (1957) Bemerkungen zum Komprehensionsaxiom. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 3, 1–17.

Turner, R. (1984) *Logics for Artificial Intelligence*. Chichester: Ellis Horwood.

Urquhart, A. (1973) An interpretation of many-valued logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 19, 111–14.

Urquhart, A. (1986) Many-valued logic. In D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, vol. III (pp. 71–116). Dordrecht: Reidel.

Williamson, T. (1994) *Vagueness*. London and New York: Routledge.

Zadeh, L. A. (1965) Fuzzy sets. *Information and Control*, 8, 338–53.
Zadeh, L. A. (1975) Fuzzy logic and approximate reasoning. *Synthese*, 30, 407–28.
Zawirski, Z. (1934) Stosunek logiki wielowartościowej do rachunku prawdopodobieństwa [Relation of many-valued logic to the calculus of probability]. *Prace Komisji Filozoficznej Polskiego Towarzystwa Przyjaciół Nauk*, 4, 155–240.

Further Reading

Bolc, L. and Borowik, P. (1992) *Many-valued Logics*, vol. 1: *Theoretical Foundations*. Berlin: Springer.
Cignoli, R. L. O., D'Ottaviano, I. M. L. and Mundici D. (2000) *Algebraic Foundations of Many-valued Reasoning*. Dordrecht: Kluwer Academic.
Haack, S. (1996) *Deviant Logic, Fuzzy Logic – Beyond the Formalism*. Chicago and London: University of Chicago Press.
Malinowski, G. (1993) *Many-valued Logics*. Oxford: Clarendon Press.

This page intentionally left blank

Part XI

INDUCTIVE, FUZZY, AND QUANTUM PROBABILITY LOGICS

This page intentionally left blank

35

Inductive Logic

STEPHEN GLAISTER

All inductive logicians aim to construct a formally articulated theory of good ampliative (non-deductive) inference that parallels existing formal theories of good deductive inference. They disagree, however, about the extent and respects of that parallel, as well as about the exact formal resources that should be brought to bear. What I will call *Good Old-Fashioned Inductive Logic* (GOFIL) holds that the parallels between deductive logic and inductive logic are straightforward and extensive. ('GOFIL' and related acronyms follow a well-known model due to John Haugeland.) On this view inductive logic, like deductive logic, studies arguments, but whereas deductive logic studies the relation of deductive validity between an argument's premises and its conclusion, inductive logic studies the degree to which those premises support or confirm that conclusion.

The obvious instrumentality with which to articulate this system of degrees is probability. The basic properties of probability are codified by axiomatizations such as those of Kolmogorov and of Renyi. Advocates of GOFIL, together with many statisticians and essentially all philosophers of probability, hold that it makes perfectly good sense to ask – even that it is essential we ask – what probability *is* beyond those basic formal properties (Salmon 1967; Walley 1991; Hájek 1997). Taking existing uses of probability concepts in both commonsense and science as the first word in matters of extra-formal interpretation, GOFIL suggests that *one* thing probability is, particularly in epistemic contexts, is a parameter expressing degree of confirmation. That is, the probability $P(B \mid A)$ that conclusion proposition B is true given premise proposition A ($= \cap A_i$ if the argument has multiple premises) is understood as a measure of the objective, logical degree to which A supports or confirms B. GOFIL therefore sponsors a so-called 'logical' interpretation of (two-place) probability (Keynes 1921; Jeffreys 1957; Carnap 1962). Since Carnap's version of GOFIL is the most developed and influential we will concentrate on that account. In section 1, then, we review the principal achievements of and challenges faced by GOFIL á là Carnap.

Many of GOFIL's achievements are detachable from that program's commitment to a logical interpretation of probability. In section 2, we survey the reincarnation of GOFIL in the context of a subjectivist interpretation of probability: a development we will call *Subjectivist Inductive Logic* (SIL).

SIL rejects GOFIL's logical account of probability, but it largely perpetuates GOFIL's basic conception of inductive logic as a matter of articulating standards of coherence

and consistency that can be used to assess particular inferences or inference forms. A more radical departure from GOFIL refocuses attention away from issues of coherence and consistency, and towards the study of various sorts of logical guarantees of convergence to the truth. We discuss this *New-Fangled Inductive Logic* (NFIL) in section 3.

1 Good Old-Fashioned Inductive Logic (GOFIL): Carnap's Program

Carnap developed his account of inductive logic through a long series of important publications over 30 years, reaching an apex of both generality and compatibility with standard probabilistic terminology in the posthumously published Carnap (1971, 1980). In this section, we will employ essentially the terminology used in this later work.

Formal preliminaries

Let a *family of properties*, $\{F_1, \ldots, F_k\}$ be a set of properties that are pairwise exclusive and jointly exhaustive. Each property in the family thus functions as a complete characterization of an individual in the logic. An *atomic proposition* is a proposition that ascribes one of the properties, F_i, to an individual, for example F_3b. A *sample proposition* is a finite conjunction of atomic propositions in which each atomic proposition involves a different individual, for example $F_3b \cap F_5c \cap F_4d$. *n(E)* or n (if the reference is clear) is the total number of individuals involved in E, and *ind(E)* is the set of individuals involved in E. Let the empty sample proposition be the necessarily true proposition, Ω.

$n_i(E)$ or just n_i (if the reference is clear) is the number of individuals to which E ascribes property F_i, and $\mathbf{n}(E)$ or $\mathbf{n}$ (if the reference is clear) is the frequency vector $\langle n_1, \ldots, n_k \rangle$ for E. The number of possible frequency vectors for a sample proposition E is $\binom{n+k-1}{k}$, where $n = n(E) = \Sigma n_i$. $\left(\text{Note that } \binom{n}{r} = \frac{n!}{r!(n-r)!}.\right)$ The number of possible sample propositions involving exactly ind(E) with a given frequency vector $\mathbf{n}$ is given by the multinomial coefficient for that vector, $\frac{n!}{n_1!n_2!\ldots n_k!}$ If there are q individuals overall then there are $\binom{q}{n}$ distinct sets of n individuals available to be partitioned by $\mathbf{n}$, hence $\binom{q}{n} \cdot \frac{n!}{n_1!n_2!\ldots n_k!}$ possible realizations of $\mathbf{n}$ in that population.

Let P be a probability function on the algebra of atomic propositions for countably many individuals. A *singular predictive inference* is a conditional probability, $P(F_ia \mid E)$, where E is a sample proposition that does not involve individual a. A *rule of succession* is a general formula for P's singular predictive inferences. We define the unconditional probability $P(F_ia)$ as $P(F_ia \mid E = \Omega)$. Unconditional probabilities for arbitrary sample propositions follow immediately, for example:

$$P(F_ia \cap F_jb) = P(F_ia) \cdot P(F_jb \mid F_ia)$$
$$P(F_ia \cap F_jb \cap F_kc) = P(F_ia) \cdot P(F_jb \mid F_ia) \cdot P(F_kc \mid F_ia \cap F_jb)$$

and so on. All other unconditional probabilities follow by additivity of the probabilities of these basic possibilities.

P is *exchangeable* just in case the probabilities of sample propositions are functions of their frequency vectors, that is P(E) = P(E′) if $\mathbf{n}(E) = \mathbf{n}(E')$. Put another way, P is exchangeable just in case one doesn't change probabilities merely by altering which individuals have which properties.

To the continuum and beyond

Carnap's first major work in inductive logic (Carnap 1962) culminates in a long appendix on a particular logical probability or degree of confirmation function, c*. The rule of succession for c* is

$$c^*(F_i a \,|\, E) = \frac{n_i + 1}{n + k}$$

that is where E involves n individuals of which n_i have property F_i.

C* gives equal prior probabilities for an individual having an arbitrary property, that is for all F_i, $c^*(F_i a \mid \Omega) = 1/k$, since $n_i = n = 0$. This implies both that each possible frequency vector for the whole population is allotted the same prior probability, and that that allotment is split evenly among its realizations. For example, when the universe of individuals comprises two coin tosses, a and b, and the family of properties is just $\{F_1 =$ 'heads', $F_2 =$ 'tails'$\}$ then the c* priors are:

$$c^*(F_1 a \cap F_1 b \mid \Omega) = c^*(F_2 a \cap F_2 b \mid \Omega) = 1/3$$
$$c^*(F_1 a \cap F_2 b \mid \Omega) = c^*(F_2 a \cap F_1 b \mid \Omega) = 1/6$$

That is, the outcomes of 0, 1, and 2 heads are given equal weight, notwithstanding that there are more ways to get exactly 1 head.

Cases such as this almost immediately start one worrying that an alternative to c* that assigns equal prior probability (here 1/4) directly to all the possible realizations of all frequency vectors might be preferable. Carnap (1962) called this alternative, $c^\dagger$, and noted that Peirce, Keynes, and Wittgenstein had all succumbed to its charms. In a tour de force, however, Carnap showed that the rule of succession for $c^\dagger$ is:

$$c^\dagger(F_i a \,|\, E) = \frac{1}{k}$$

That is, $c^\dagger$ makes a certain sort of empirical learning impossible: its singular predictive inferences ignore all *observed frequency* information.

Whence comes the appeal of $c^\dagger$, say in the coin-tossing case above, if it is, in the abstract, inductively catastrophic? Evidently its appeal in the case at hand is grounded in the fact that we are jumping to the conclusion that the coin to be tossed is (close to) objectively fair. If we assume this (or indeed any other particular bias for the coin) then there is an important sense, underlined by Carnap's result for $c^\dagger$, in which we already

know everything there is to know about the coin we are tossing. The exact sequence of heads and tails remains to be determined, of course, but that's just the unfolding of a chance process: it's 'whatever happens.' And if we know the chance parameters for the overall process then no stage of the unfolding chance process tells us anything about any other stage of that process. $c^{\dagger}$ is not an inductive catastrophe in the case of tossing a coin with known bias, rather it's a legitimate expression of the fact that there's nothing left to learn about the case at hand.

Reflecting on $c^{\dagger}$ in this way helps us see that the basic inductive problem for Carnap is equivalent (given two outcomes) to trying to figure out the bias of a coin from the actual outcomes of a series of tosses. From this perspective, the degeneracy of $c^{\dagger}$ is just that it is appropriate only for a case in which exactly *that* inductive problem has already been solved. In actual empirical applications, moreover, we are always open to revisiting our estimates of a coin's bias – a string of 1000 heads from a coin we believed fair would *always* give us pause. It follows that $c^{\dagger}$ is at best a contextually specific, convenient approximation; one which Carnap himself analogizes to the use of 22/7 to approximate π.

At the opposite extreme from $c^{\dagger}$ is the so-called 'straight rule,' c^{st}. This alternative to c^{*} ignores all *prior probability* information and simply predicts the continuation of the components of the sample proposition's frequency vector into the future, that is:

$$c^{st}(F_i a \mid E) = \frac{n_i}{n}.$$

We can think of c^{st} as at one end of a spectrum, giving no weight to the prior probabilities, $c^{\dagger}$ at the other end, giving incomparably great weight to the priors, and c^{*} as somewhere in between. Carnap (1952) makes the obvious weight parameter explicit, yielding the following family of rules of succession:

$$c^{\lambda}(F_i a \mid E) = \frac{n_i + \lambda/k}{n + \lambda}$$

where $0 \leq \lambda \leq \infty$. Since c^{λ} has continuum many instances, Carnap called this system the 'Continuum of Inductive Methods.' C^{λ} reduces to c^{st}, c^{*}, and $c^{\dagger}$ when λ is 0, k, and ∞ respectively.

Carnap (1980) generalizes still further, dropping the requirement of uniformity of the prior probability, γ_i, that an individual will have property F_i:

$$c^{\lambda,\gamma}(F_i a \mid E) = \frac{n_i + \lambda\gamma_i}{n + \lambda}$$

where λ is positive and finite and $\Sigma\gamma_i = 1$. (Carnap (1980) analyzes extreme rules such as c^{st} and $c^{\dagger}$ only in the limit, as λ approaches 0 and ∞ respectively.) We can think of the probabilities generated by this sort of rule as a matter of first augmenting the n-membered sample population with a virtual population comprising λ individuals with frequency vector $\langle \lambda\gamma_1, \lambda\gamma_2, \ldots, \lambda\gamma_k \rangle$, then recalculating the relative frequencies from there. C^{*} is the case where each property gets exactly one virtual representative (Jeffrey 1980: 2–3).

Carnap presented many different sets of *qualitative* conditions on P over the years, each of which he intended to be sufficient for his favored family of logical probabilities at the time. Carnap (1980: section 19) proves that the following relatively sparse group of conditions is sufficient for the λ–γ-continuum of inductive methods:

(λ–γ1) P is exchangeable.
(λ–γ2) P is regular: $P(E) > 0$ for all E.
(λ–γ3) Strict Instantial Relevance: $P(F_ib \mid F_ia) > P(F_ib)$.
(λ–γ4) Sufficientness: $P(F_ia \mid E)$ is a function just of n and n_i.

Note that $c^{st} = c^{0,\gamma}$ conflicts with (λ–γ2) since if $n_i(E) = 0$ then $c^{0,\gamma}(E) = 0$, and $c^{\infty,\gamma}$ conflicts with (λ–γ3) since $c^{\infty,\gamma}(F_ib \mid F_ia) = c^{\infty,\gamma}(F_ib)$. Lastly, note that (λ–γ4) is vacuous if there are only two possible properties, so that, strictly speaking, the given postulates only imply the λ–γ-continuum for $k \geq 3$. Carnap saw the problem, and solved it – inelegantly – by adducing a *quantitative* axiom of linearity to cover the $k = 2$ case. We set aside this unfortunate wrinkle in Carnap's approach here.

The basic problem

Before discussing relatively technical objections to and further developments of Carnap's program, it is worth asking about the extent to which that program succeeds in meeting its original goals. Recall that the basic suggestion of GOFIL was that *one* thing probability could be, particularly in epistemic contexts, is a parameter registering degree of confirmation.

Now if, say, c^* had emerged as a uniquely compelling inductive method then it would have been possible for Carnap to declare victory: to say that probability in many epistemic contexts just *is* c^*. But if, as Carnap clearly believed by 1952, c^* is *not* uniquely compelling so that, so to speak, degrees of confirmation are *many* while probabilities are *one*, then the logical interpretation of probability has to be abandoned. $P(B \mid A)$ *can't* just be the degree to which A confirms B since degree of confirmation turns out to have additional argument places that probability lacks. Moreover, Carnap agrees that the value of λ "is fundamentally not a theoretical question" but a "practical decision," albeit one which can be importantly informed by "theoretical results concerning the properties of the various inductive methods" (Carnap 1952: 53). The λ–γ-continuum, of course, only expands the role of practical decision making. But then, whatever else he might have done, Carnap hasn't provided a logical interpretation of (two-place) probabilities, fixed as relations of entailment are fixed, simply by the nature of the underlying field of propositions. And assigning all particular degrees of support only relative to λ- and γ-values that are themselves matters of decision raises the specter of circularity, or at least of a kind of holism about probability (i.e. if those decisions, as it's natural to suppose, already involve probabilistic reasoning of some kind). This is unpromising ground on which to try to erect a strictly logical interpretation of probability.

From the late 1950s onward, and especially in Carnap (1963), Carnap strongly emphasizes the role of degrees of confirmation in helping determine expected utilities, fair betting quotients, and so on. It is not clear whether Carnap intended this new emphasis to solve or to concede defeat by the problems raised in this section. Whatever

Carnap may have intended, strongly emphasizing betting behavior and decision making *invites* exploration of how much of GOFIL can be retained given a subjectivist interpretation of probability. And historically this path has been very popular. Indeed, in one obvious respect, GOFIL apparatus immediately acquires a new luster in a subjectivist setting: symmetry arguments and principles that are endlessly controversial when wielded as additional universal postulates to help fix logical probabilities, are necessarily less objectionable when employed opportunistically, as tools for forming subjective probability models of particular cases. We consider subjectivized inductive logic in detail in section 2.

Other problems and developments

In this section, we briefly review some relatively technical problems for GOFIL.

Confirming universal generalizations

In domains where there are infinitely many individuals all of Carnap's inductive methods give zero prior probabilities (hence – except for c^{∞} – also zero posterior probabilities on finite evidence) to universal generalizations (UGs). One response is simply to accept the consequence, fashioning the point either as a sobering reminder of how far the literally universal outstrips our abilities to probabilize (R. Price, De Morgan, Jeffrey), or as a demonstration of how far the mathematics of infinity takes us away from the sorts of epistemic contexts that matter (Ramsey, Savage, T. Fine). The other main response is, of course, to try to modify Carnap's apparatus to permit assigning positive prior probability to UGs in infinite domains (and swifter confirmation in finite domains). From a subjectivist perspective, the problem is no sooner stated than it is solved: simply put finite probability where it's needed and make appropriate adjustments elsewhere (Jeffreys and Wrinch 1919; Earman 1992: 89–90). But how to justify this sort of flexibility within a GOFIL setting?

Two attempts have been made to meet this challenge, both of which centrally involve amending (λ–γ4) (actually both amendments address only the λ-continuum, but the difference doesn't matter here). Zabell (1997b) proposes the following minimal modification:

(λ–γ4.1) $P(P_i a \mid E)$ is a function just of n, n_i, *except* when E involves only a single property.

and proves a remarkable theorem showing that essentially just exchangeability of P determines both:

- the existence of prior and posterior probabilities for each UG, $(\forall j)F_i a_j$;
- a formula that makes $P(F_i a \mid E)$ in these cases a weighted average of its λ-continuum value and the posterior probability for $(\forall j)F_i a_j$.

The most extended response to the problem of confirming UGs within GOFIL is due to Hintikka and Niiniluoto (1980). The core of this response is the weakening of (λ–γ4) to:

(λ–γ4.2) $P(P_i a \mid E)$ is a function just of n, n_i, and of the number of distinct properties *not* involved in E.

The thought behind the additional argument parameter here is that "it determines the number of nonequivalent generalizations compatible with the sample" (Hintikka and Niiniluoto 1980: 160). Coordinately, the underlying technical innovation of the so-called H-N systems is to assign probabilities in the first instance directly to generalizations about which properties (and relations) are instantiated in a population (the 'constituents' of Hintikka (1966)). One can, it turns out, do this in a way that permits (1) the calculation of all sample proposition probabilities, (2) positive probability for UGs independently of the cardinality of the domain, and (3) much faster confirmation of generalizations in finite domains. Carnap's λ-continuum even emerges as the sole H-N system in which UGs fail to receive positive probability in infinite domains (Hintikka and Niiniluoto 1980: 173).

New properties/species

Carnap's inductive methods (as well as the H-N systems) suppose that we know all the basic properties, $\{F_i\}$, in advance. But this is deeply unrealistic: real-world inductions involve learning about new types almost as much as they do learning about new tokens (of pre-digested types). Zabell (1992, 1997a) shows how to make a Carnapian framework more realistic, by allowing for singular predictive inferences about novel properties or species.

Suppose for ease of exposition that the observables are letters in the alphabet. Evidently the frequency vector (from n_A to n_Z) for the sample of observations DFBBBAA, $\langle 2, 3, 0, 1, 0, 1, 0, 0, \ldots, 0\rangle$ is not an appropriate statistic unless we possess a prior enumeration of all 26 types. If we knew *just* DFBBBAA, then the only available frequency vector would seem to be the vector comprising the frequencies of the actually observed types (from n_A to n_F), $\langle 2, 3, 1, 1\rangle$. We can further imagine interpreting the vector in the following minimal fashion:

> "fourth species observed showed up twice," "third species observed showed up three times,". . . .

If we now suppose that possible observations are indexed by times $\{1, \ldots, n\}$, then the frequency vector just for observed types can be thought of as constituting a partition π of the index set $\{1, \ldots, 7\}$. We can now generate a higher-order statistic for a partition corresponding to the *frequencies of the frequencies* in the partition of the index set. Let a_j be the number of types with j observed tokens, that is the number of j-membered partition cells in the partition of the index set. For example, $a_1(\text{DFBBBAA}) = 2$, $a_2(\text{DFBBBAA}) = a_3(\text{DFBBBAA}) = 1$.

Let $\mathbf{\Pi}_n$ be a random variable taking as values possible partitions of $\{1, \ldots, n\}$ and let $\mathbf{a} = \langle a_1, \ldots, a_n\rangle$ be the *partition vector* for an n-membered sample, so that, for example, $\mathbf{a}(\text{DFBBBAA}) = \langle 2, 1, 1, 0, 0, 0, 0\rangle$. Suppose finally that one has a probability over the space of possible partitions of $\{1, \ldots, n\}$ and say that that probability is *partition exchangeable* iff all partitions with the same partition vector are equiprobable, that is $P(\mathbf{\Pi}_n = \pi_1) = P(\mathbf{\Pi}_n = \pi_2)$ if $a(\pi_1) = a(\pi_2)$, where π_1 and π_2 are partitions of $\{1, \ldots, n\}$.

Zabell (1997a) shows that a new, three-parametered continuum of inductive methods is implied by partition exchangeability, and three other conditions. The first two conditions are just partition counterparts of regularity (λ–γ2) and sufficientness (λ–γ4). The final condition governs the probability that the next individual is of a novel species:

(Z) $P(e_{n+1} \in S_{t+1} \mid \langle n_1, \ldots, n_t \rangle)$ is a function just of the number of species already observed, t, and the sample size, n.

See Zabell (1997a: section 2) for the new continuum of predictive probabilities itself. For some of the relations between the new continuum and the H-N systems, see Zabell (1992: 218).

Note finally that some authors (Salmon 1967; Fine 1973) have worried that Carnapian degree of confirmation values are inappropriately sensitive to refinements in the space of properties. Zabell's work, however, appears to demonstrate one way in which such sensitivity is not only appropriate but essential.

Analogy

Carnap's inductive logic can be understood as importantly anti-analogical. Whereas exchangeability implies that order of individuals is unimportant, proximity of individuals in some ordering or metric is often taken to be a reasonable basis for inference. Similarly, whereas sufficientness conditions insulate predictions about one type from (frequency) information about other types, proximity of types in some ordering or metric is often taken to be a reasonable basis for inferring features of one type from another.

We briefly discuss the weakenings of exchangeability that are needed for models of the first sort of analogical reasoning in the discussion of "De Finetti's exchangeability reduction" below. Notable attempts to model the second sort of analogy in a broadly Carnapian spirit include Niiniluoto (1981), Constantini (1983), Kuipers (1984), Skyrms (1993), and Maher (2000).

2 Subjectivized Inductive Logic (SIL): De Finetti Regnant

We observed in the section "To the continuum and beyond", above, that Carnap thinks of the basic inductive problem as analogous to trying to divine the bias of a coin from an actual sequence of tosses. The crucial point is that the system being theorized about is supposed, conditional on any particular bias value(s), to produce sequences of (objectively) independent, identically distributed (same objective probabilities each time for the different possible outcomes) trials.

This situation is one of the most well-studied problems in statistics, and particularly in Bayesian statistics. From a Bayesian statistical perspective, the problem is just to choose an appropriate prior probability on the possible values of the bias parameter so that (1) computation of posteriors is easy, and (2) convergence to the true bias is guaranteed. Sometimes statisticians recommend a uniform or 'flat' prior for these purposes. At least equally commonly, however, they allow any member of the family of priors that

essentially share the functional forms of the likelihoods (i.e. the probabilities of sample propositions given values of the relevant bias parameter(s)) as functions of the bias parameter(s). Most distinguished are the so-called *natural conjugate* priors. (When the product of the prior and the likelihood yields a posterior distribution in the same family as the prior, the prior is said to *conjugate* with the likelihood function. When that prior conjugates with that likelihood by essentially sharing its functional form then the prior is *natural*.) If the system generating the sample is binomial, the natural conjugate priors are the Beta distributions; if it is multinomial the natural conjugate priors are the Dirichlet distributions. These distributions themselves have multiple parameters. If these parameters have uniform values the resulting Beta and Dirichlet distributions are said to be symmetric. Flat priors result if that uniform value is 1 (Festa 1993: chapter 6; Tanner 1996).

Remarkably, the flat prior corresponds exactly to Carnap's c*, the *symmetric* natural conjugate priors to the λ-continuum, and the natural conjugate priors *in toto* to the λ–γ-continuum. GOFIL subjectivized – SIL – just is Bayesian statistics. The great connecting principle here is Carnap's requirement that degree of confirmation functions be exchangeable, that is (λ–γ1). Famously, De Finetti (1937) proved that any infinite sequence of random variables (i.e. one for each trial) for which every finite subsequence is exchangeable (i.e. according to a subjective probability P over those infinite sequences of trials) has a unique representation as a (possibly continuously) weighted average or mixture of probabilities, each one of which makes the random variables (r.v.s) independent and identically distributed (IID). We will state De Finetti's result precisely just for the binomial case:

DE FINETTI REPRESENTATION THEOREM Let $\{X_i\}_{i=1}^{\infty}$ be an infinite sequence of {0, 1}-valued random variables with $\{X_i\}_{i=1}^{n}$ exchangeable for each n (according to P); then there is a unique probability measure μ on [0,1] such that for each fixed sequence of zeros and ones $\{e_i\}_{i=1}^{n}$ we have

$$P(X_1 = e_1, \ldots, X_n = e_n) = \int_0^1 p^k (1-p)^{n-k} d\mu(p)$$

where

$$k = \sum_{i=1}^{n} e_i \text{ (i.e. k is the number of 'successes').}$$

This theorem, together with its relatives for sequences of multinomial r.v.s, real-valued r.v.s (De Finetti 1937), and beyond (Hewitt and Savage 1955), implies that choosing c*, c^{λ}, and $c^{\lambda,\gamma}$ as rules of succession is equivalent to choosing the various distributions (or distribution families) mentioned above as mixing measures for the relevant version of De Finetti's Theorem.

Probabilities with respect to infinite sequences of random variables are sometimes decried as unrealistic (Jeffrey 1992), and in part for this reason, finite exchangeable sequences of r.v.s have also been extensively studied. Probabilities for such sequences have unique representations as mixtures of (non-IID) hypergeometric sequences. Extendibility of a finite exchangeable sequence to longer and longer finite exchange-

able sequences, however, ensures convergence to representability by a mixture of IID sequences (Diaconis 1977). We set aside the case of finite exchangeability here.

Conditionalizing an exchangeable probability P on outcomes of trials leaves the subsequences of remaining outcomes exchangeable. By Bayes' theorem and the weak law of large numbers, the weights of subsequent mixing measures gradually become focused on a single IID sequence, corresponding to a single parameter value (or vector of parameter values) *unless* one's original mixing measure starts out strangely skewed away from the true parameters of the system, a possibility that natural conjugacy blocks (Diaconis and Freedman 1986). The power of De Finetti's representation theorem is that it shows how this elegant model of learning from experience is implicit in little more than an assumption of a particular sort of subjective indifference or symmetry in one's personal probabilities – exchangeability – together with the assumption that new information is assimilated via conditionalization. As we saw above, De Finetti's theorem also clarifies how to understand Carnap's efforts from a subjectivist standpoint. Over and above identifying the exchangeable probabilities, Carnap's various conditions can be seen as limning the properties of various families of prior mixing measures. De Finetti's result also suggests a more general perspective, according to which to equip a subjective probability with a symmetry of some kind just *is* to endow the agent in question with a conception of objective chance. This vision, which can be pursued through more and more abstract symmetries, often with mathematical roots independent of De Finetti's work, for example in ergodic theory, has proved tantalizing (Skyrms 1984: chapter 3, 1994).

De Finetti himself boldly made two further claims on behalf of his theorem (and its supporting materials): that it paved the way for the complete elimination of objective probability or chance parameters from statistics, and that it solved Hume's problem of induction. Let us briefly consider these claims in turn.

De Finetti's exchangeability reduction

An immediate, technical obstacle to any *general* reduction of IID notions – objective independence and objective equiprobability – to exchangeability is that if the infinite sequence of random variables take values in very rich spaces then no representation in terms of mixtures of sequences of IID trials for those r.v.s may be possible (Dubins and Freedman 1979). But let us set aside this relatively technical worry here.

In many cases, one needs to construct a subjective probability for a situation or phenomenon. And when the phenomenon is an infinite (or infinitely extendible) sequence of random variables it makes sense to ask whether exchangeability or some other related symmetry assumption is justified or reasonable. It certainly *looks* as though the better part of that justification will be an appeal to background knowledge about the phenomenon in question, to our understanding of how 'coin toss' – or 'urn model' – like the phenomenon is. If the phenomenon is judged to be 'coin toss' – like – or in the simplest case just *is* the tossing of a coin of some kind, then an IID sequence can be reasonably expected, and, in effect, only the constant probability of success parameter remains to be determined. But if we had some specific and contrary background knowledge about the coin in question, for example if we knew that the coin was made of some

highly unstable material such that every heads outcome increases the probability of heads on the next toss, then it would be perverse to assume exchangeability. In this sort of case the *order* of outcomes matters and not just the frequency vector, hence exchangeability is inappropriate.

Subjectivists have developed models of phenomena that objectivists would describe as exhibiting various sorts of parameterized dependency, under the generic heading of *partial exchangeability* (De Finetti 1938). The best explored of these is Markov exchangeability which focuses not on invariance of probabilities under permutation of trials (and on frequency vectors) but on invariance under switching of sub-sequences of trials that share starting and ending points (and on vectors of initial states and transition counts). See Diaconis and Freedman (1980) and Skyrms (1994: section 5) for further discussion.

The work in this area is impressive, and constitutes an absolutely essential broadening of the base for De Finetti's reductive proposal. It seems unlikely, however, that it does much more than push our basic objection back a step. Even with a wider arrange of symmetries to appeal to, the subjectivist still seems to have to play catch-up with respect to the objectivist. There are, after all, essentially unlimited forms of dependency and objective eccentric character, so that it is hard to see how to avoid the conclusion that symmetries in subjective probabilities are normally best seen as responses to background knowledge about objective symmetry and dependency in the target phenomenon rather than the other way around. Compare Gillies (2000: 77–84) and Walley (1991: 460–7).

Hume and grue again

We will approach the question of what De Finetti-style subjectivist inductive logic (SIL) has to say about Hume's problem of induction anachronistically, via Goodman's new riddle of induction. Let something be *grue* just in case it is green before some future date D or blue after D. Thus grass is grue before D, but not afterwards, and so on. Goodman thinks that we all agree that "Regularity in greenness confirms the prediction of further cases; regularity in grueness does not" (Goodman 1983: 82). Put probabilistically and in terms of singular predictive inferences: observing green individuals leading up to D-Day raises the probability that the next observed individual (i.e. on or after D-day) will be green whereas it does not raise the probability that it will be grue. Goodman wonders about the basis for distinctions of the green/grue kind. After reproaching Hume for failing to provide such a basis, Goodman himself offers an account that stresses the asymmetrical rootedness of the predicate 'green' in past linguistic and inductive practice (Goodman 1983: chapter 4). Let us now see whether SIL can do as well or better.

In Goodmanian D-day cases, the crucial 'next observed individual' is held fixed at D-day while we haplessly pile up observations prior to its fatefully dated occurrence. An alternative is to treat the next observed individual as a kind of moving target: as we pile up additional observations, the next observed individual, like the proverbial 'free beer tomorrow,' skips ahead always to be observed next. Call these the fixed target and moving target conceptions of 'the next observed individual' respectively (Earman 1992: section 4.7).

Jeffreys (1957) showed that inductive skepticism about the character of the next observed individual in the moving target case ('moving target inductive skepticism') is almost impossible to maintain, since

$$\lim_{n\to\infty} P(Fa_{n+1} \mid Fa_1 \cap \ldots \cap Fa_n) = 1$$

if $P((\forall i)Fa_i) > 0$. This sufficient condition surely has broad skeptical appeal – a skeptic should want to avoid having to be *a priori* certain that $(\exists i)\neg Fa_i$. This is evidently a kind of limiting answer to Hume but it is also a partial answer to Goodman. Since Jeffreys's result does not turn on what 'F' means, contrary to what Goodman might be taken to suggest, grue is on the same footing as green in the moving target sense. No contradiction results from raising the probabilities of both 'the next observed individual is green' and 'the next observed individual is grue' in the moving target sense, since their rates of convergence to the limits in question can, and indeed must be, different (Howson 1973).

Blunting inductive skepticism about the character of the next observed individual in the *fixed* target case ('fixed target inductive skepticism' – clearly the principal case for both Goodman and Hume) requires De Finetti-style symmetry assumptions, not just ringing the changes on the probability calculus. The limit we wish to evaluate in this case (where indices of the a_i now range over both positive and negative integers) is:

$$\lim_{j\to\infty} P(Fa_{n+1} \mid Fa_n \cap \ldots \cap Fa_{n-j})$$

It matters here what 'F' means since suppose day_{n+1} is the D-day for the green/grue divergence and that 'F' means 'is green.' Then

$$\lim_{j\to\infty} P(Fa_{n+1} \mid Fa_n \cap \ldots \cap Fa_{n-j}) = 1$$

implies that

$$\lim_{j\to\infty} P(F^*a_{n+1} \mid F^*a_n \cap \ldots \cap F^*a_{n-j}) = 0$$

where 'F*' means 'is grue.'

But if P is exchangeable with respect to a given property (i.e. for the infinite sequence of r.v.s comprising the indicator functions for the presence of that property) then the moving target and fixed target limits have to agree. The grue/green case therefore makes for the following inconsistent triad: (1) P is exchangeable with respect to both F and F*; (2) $P((\forall i)Fa_i) > 0$; and (3) $P((\forall i)F^*a_i) > 0$.

Basic openminded-ness militates against denying either (2) or (3), so exchangeability with respect to at least one of F and F* must go. Thus, once we grant the inductive skeptic a fixed target at which to aim, symmetries in our subjective probabilities are going to constitute most of our (broadly subjectivist) answer to that skeptic. Those symmetries constitute the respects of resemblance or uniformity that we are expecting to continue into the future, and those determinate expectations implicitly involve us in ignoring countless other abstractly possible respects of resemblance. This is De Finetti's

answer to Hume. It is conditional or coherence-minded in much the same way that Goodman's 'past practice' answer is. There's nothing in the theory of exchangeability to say which, if any, properties we should find exchangeable, just as Goodman does not presume to say what our past practices should be. De Finetti's advantage over Goodman is just the clarity afforded within a probabilistic framework for stating and relating the conditions of induction precisely: Bayesian projectibility (Skyrms 1994) is alive and well whereas Goodman's theory of projectibility, as opposed to Goodman's sensational riddle, is a philosophical and logical back-water.

3 New-Fangled Inductive Logic (NFIL)

Logical accounts of the truth-conduciveness of methods of inquiry reached technical and philosophical maturity in the 1980s and 1990s, building on the seminal work of Putnam (1965) and Gold (1965). This New-Fangled Inductive Logic (NFIL) takes guarantees of different senses of convergence to the truth to be the primary object of logical study. Considerations of coherence or consistency – probabilistic or otherwise – are distinctly secondary: they warrant study principally for whether they are likely to block, slow down, or otherwise interfere with convergence to the truth. Our bare-bones treatment of NFIL follows Kelly (1996: chapters 3 and 4).

Consider an idealized scientist trying to determine by passive observation whether some hypothesis, h, is true. We represent the scientist's background knowledge as a set of possible worlds, K, in some of which h is true and in some of which h is false. We suppose that any world in K produces a stream of data, ε, of which the scientist scans only the initial segment, $\varepsilon|n$, up to the current stage, n. The scientist is, we will assume, equipped with an inductive method, α, drawn from some larger set of methods, M, and that the scientist conjectures something about the status of the h after each new data point. We further assume that all of the worlds in K are exhaustively observable. This allows us to identify worlds with their unique data streams, and hypotheses with sets of data streams. Given these identifications, the truth of a hypothesis depends just on the data stream: h is true on ε just in case $\varepsilon \in h$. Lastly, we will assume that the data types are natural numbers and sundry other symbols, which we can think of as codes for more realistic sorts of discrete data types.

Let us now formulate four, increasingly weak senses in which inductive method α may converge to a verdict of some kind.

(C1)	α produces b *by stage n* on h, ε	iff $\alpha(h, \varepsilon\|n) = b$
(C2)	α produces b *with certainty* on h, ε	iff there is a stage n s.t. $\alpha(h, \varepsilon\|n) = !$, $\alpha(h, \varepsilon\|n + 1) = b$, and, for all $m < n$, $\alpha(h, \varepsilon\|m) \neq !$
(C3)	α produces b *in the limit* on h, ε	iff there is a stage n s.t., for all $m \geq n$, $\alpha(h, \varepsilon\|m) = b$
(C4)	α *approaches* b on h, ε	iff for each rational $s \in (0, 1]$, there is a stage n s.t. $\|b - \alpha(h, \varepsilon\|m)\| < s$, for all $m \geq n$

Intuitively speaking, (C1) is convergence by a deadline n, (C2) is convergence to when one is first prepared to say ('1') that one has the answer, (C3) is convergence in the sense of eventual stability in ones conjectures, and (C4) is convergence in the sense of getting closer and closer to some value.

The three most general notions of success for a method α on h, ε involve α conjecturing that h is true ('1') just in case h is true (verification), conjecturing that h is false ('0') just in case h is false (refutation), and both (decision). We provide clauses for the four notions of verification our four senses of convergence induce; clauses for refutation and decision are similar.

(C1v) α verifies h by stage n on ε iff [α produces 1 at n on h, $\varepsilon \leftrightarrow \varepsilon \in h$]
(C2v) α verifies h with certainty on ε iff [α produces 1 with certainty on h, $\varepsilon \leftrightarrow \varepsilon \in h$]
(C3v) α verifies h in the limit on ε iff [α produces 1 in the limit on h, $\varepsilon \leftrightarrow \varepsilon \in h$]
(C4v) α verifies h gradually on ε iff [α approaches 1 on h, $\varepsilon \leftrightarrow \varepsilon \in h$]

The reliability of an inductive method α is a matter of quantifying over the possible worlds/data streams on which α succeeds, for example:

(C1vK) α verifies h by stage n given K iff for each $\varepsilon \in K$, α verifies h at n on ε

and so on for the rest. We can also quantify over the range of hypotheses that method α can assess reliably, for example:

(C1vKH) α verifies H at stage n given K iff for each $h \in H$, α verifies h at stage n given K

Finally, one can ascend to the level of inductive problem solvability by generalizing over the collections of methods in M, for example:

(C3rKHM) H is refutable in the limit given K by a method in M iff there is an $\alpha \in M$ s.t. α refutes H in the limit given K

One can now set about exploring all these notions using the technical palette of the theory of computability and recursion theory. Elementary but important results given no restrictions on M include:

- Verifiability, refutability, and decidability are equivalent for (C1) (Kelly 1996: 45), but not for any of the weaker senses of convergence we have defined (Kelly 1996: 68). For example, the existential hypothesis that mass m is divisible is verifiable with certainty but not refutable (hence not decidable) with certainty, and the same existential within the scope of a universal (e.g. the hypothesis that mass m is infinitely divisible) is refutable in the limit but not verifiable (hence not decidable) in the limit.
- The whole structure can be characterized topologically, roughly by mapping existential and universal hypotheses onto open and closed sets respectively (Kelly 1996: 85; see also Schulte and Juhl 1996).

- Decidability in the limit and gradual decidability are equivalent (Kelly 1996: 67). The class of hypotheses that are so decidable can be classified in the finite Borel hierarchy as Δ_2^B.

NFIL affords an important, abstract yet flexible perspective on inductive inference. It reinvigorates the question of whether conditional, coherence-based answers to Humean and other skepticisms, are answers at all. If the coherencies insisted upon block us from reliably getting to the truth there's a clear sense in which they aren't. NFIL also provides some external check on what might otherwise look like relatively innocuous or 'merely technical' assumptions. Consistency, countable additivity, consideration just of properties (monadic predicates), and so on, are often adopted fairly peremptorily both inside and outside of the GOFIL/SIL tradition. But such assumptions may have hidden costs or be providing illicit benefits, and NFIL promises to help us see clearly whether this is so (Earman 1992: chapter 9; Kelly 1996: chapter 13). NFIL also has the peculiarly philosophical virtue of making strange bedfellows: from its perspective SIL approaches tend to look much more of a piece with traditional, justification-centered programs in epistemology than is usually allowed (see Earman 1992: 219). Our own view is that the articulation of the NFIL perspective on induction – a perspective for which Ramsey calls in the final two sections of his "Truth and Probability" – is a very positive development, but only time will tell whether this view is correct. In any case, both in its ingenuity and its contentiousness, we expect the future of inductive logic to resemble its past.

References

Carnap, R. (1952) *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Carnap, R. (1962) *The Logical Foundations of Probability*, 2nd edn. Chicago: University of Chicago Press.

Carnap, R. (1963) The aim of inductive logic. In E. Nagel, P. Suppes and A. Tarski (eds.), *Logic Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress* (pp. 303–18). Stanford, CA: Stanford University Press.

Carnap, R. (1964) Replies and systematic expositions. In P. A. Schillp (ed.), *The Philosophy of Rudolph Carnap* (pp. 859–1013). La Salle, IL: Open Court.

Carnap, R. (1971) A Basic system of inductive logic, part I. In Carnap and Jeffrey (eds.), *Studies in Inductive Logic and Probability* (pp. 33–165). Berkeley, CA: University of California Press.

Carnap, R. (1980) A Basic system of inductive logic, part II. In Jeffrey (ed.), 1980 (pp. 7–155).

Constantini, D. (1983) Analogy by similarity. *Erkenntnis*, 20, 103–14.

De Finetti, B. (1937) Foresight: its logical laws, its subjective sources. In H. E. Kyburg and H. E. Smokler (eds.), *Studies in Subjective Probability* (pp. 93–158). New York: John Wiley & Sons. (Originally published as *La Prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincaré*, 7, 1–68.)

De Finetti, B. (1938) On the condition of partial exchangeability. In Jeffrey (ed.), 1980 (pp. 193–205). (Originally published as *Sur la condition d'équivalence partielle. Actualités scientifiques et industrielles*, 739. Paris: Herman & Cie.)

Diaconis, P. (1977) Finite forms of De Finetti's theorem on exchangeability. *Synthese*, 36, 271–81.

Diaconis, P. and Freedman, D. (1980) De Finetti's theorem for Markov chains. *Annals of Probability*, 8, 115–30.

Diaconis, P. and Freedman, D. (1986) On the consistency of Bayes estimates. *Annals of Statistics*, 14, 1–26.

Dubins, L. and Freedman, D. (1979) Exchangeable processes need not be mixtures of independent identically distributed random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebeite*, 48, 115–32.

Festa, R. (1993) *Optimum Inductive Methods*. Dordrecht: Kluwer.

Fine, T. (1973) *Theories of Probability*. New York: Academic Press.

Gillies, D. (2000) *Philosophical Theories of Probability*. London: Routledge.

Gold, E. M. (1965) Limiting recursion. *Journal of Symbolic Logic*, 30, 27–48.

Hájek, A. (1997) 'Mises redux' – redux: fifteen arguments against finite frequentism. *Erkenntnis*, 45, 209–27.

Hewitt, E. and Savage, L. (1955) Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80, 470–501.

Hintikka, J. (1966) A two-dimensional continuum of inductive methods. In J. Hintikka and P. Suppes (eds.), *Aspects of Inductive Logic* (pp. 113–32). Amsterdam: North-Holland.

Hintikka, J. and Niiniluoto, I. (1980) An axiomatic foundation for the logic of inductive generalizations. In Jeffrey (ed.), 1980 (pp. 157–81).

Howson C. (1973) Must the logical probability of laws be zero? *British Journal of the Philosophy of Science*, 24, 153–63.

Jeffrey, R. (1980) Introduction. In Jeffrey (ed.), 1980 (pp. 1–6).

Jeffrey, R. (ed.) (1980) *Studies in Inductive Logic and Probability*, vol. 2. Berkeley, CA: University of California Press.

Jeffreys, H (1957) *Scientific Inference*, 2nd edn. Cambridge: Cambridge University Press.

Jeffreys, H. and Wrinch, D. (1919) On certain aspects of the theory of probability, *Philosophical Magazine*, 38, 715–31.

Keynes, J. M. (1921) *A Treatise on Probability*. London: Macmillan.

Kuipers, T. (1984) Two types of analogy by similarity. *Erkenntnis*, 21, 63–87.

Maher, P. (2000) Probabilities for two properties. *Erkenntnis*, 52, 63–91.

Niiniluoto, I. (1981) Analogy and inductive logic. *Erkenntnis*, 16, 1–34.

Putnam, H. (1965) Trial and error predicates and a solution to a problem of Mostowski. *Journal of Symbolic Logic*, 30, 49–57.

Salmon, W. (1967) *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Schulte, O. and Juhl, C. (1996) Topology as epistemology. *Monist*, 79, 141–7.

Skyrms, B (1993) Analogy by similarity in hyperCarnapian inductive logic. In J. Earman et al. (eds.), *Philosophical Problems of the Internal and External Worlds* (pp. 273–82). Pittsburgh: University of Pittsburgh Press.

Skyrms, B. (1994) Bayesian projectability. In D. Stalker (ed.), 1994 (pp. 241–62).

Tanner, M. A. (1996) *Tools for Statistical Inference: methods for exploration of posterior distributions and likelihood functions*, 3rd edn. Heidelberg: Springer.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman & Hall.

Zabell, S. (1992) Predicting the unpredictable. *Synthese*, 90, 205–32.

Zabell, S. (1997b) Confirming universal generalizations. *Erkenntnis*, 45, 267–83.

Further Reading

Earman, J. (1992) *Bayes or Bust: a critical examination of confirmation theory*. Cambridge, MA: MIT Press.

Goodman, N. (1983) *Fact, Fiction and Forecast*, 4th edn. Cambridge, MA: Harvard University Press.

Howson C. and Urbach P. (1993) *Scientific Reasoning: The Bayesian Approach*, 2nd edn. La Salle, IL: Open Court.

Kelly, K. (1996) *The Logic of Reliable Inquiry*. Cambridge: Cambridge University Press.

Kuipers, T. (1978) *Studies in Inductive Probability and Rational Expectation*. Dordrecht: Reidel.

Jeffrey, R. (1992) *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Maher, P. (1999) Inductive logic and the ravens paradox. *Philosophy of Science*, 66, 50–70.

Ramsey, F. P. (1926) Truth and probability. In D. H. Mellor (ed.), *F. P. Ramsey: Philosophical Papers* (pp. 52–94). Cambridge: Cambridge University Press.

Skyrms, B. (1984) *Pragmatics and Empiricism*. New Haven, CT: Yale University Press.

Stalker, D. (1994) *Grue!* La Salle, IL: Open Court.

Zabell, S. (1988) Symmetry and its discontents. In W. Harper and B. Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, vol. II (pp. 155–90). Dordrecht: Kluwer.

Zabell, S. (1989) The rule of succession. *Erkenntnis*, 31, 283–321.

Zabell, S. (1997a) The continuum of inductive methods revisited. In J. Earman and J. Norton (eds.), *The Cosmos of Science* (pp. 351–85). Pittsburgh: University of Pittsburgh Press.

36

Heterodox Probability Theory

PETER FORREST

The purpose of this chapter is to survey and assess the ways of departing from the Bayesian orthodoxy about probabilities as they apply to reasoning. Most of these departures are no longer fashionable, but they deserve reconsideration. For a more historical survey readers are referred to Hailperin (1990).

Perhaps the least controversial departure from orthodoxy is the way of adapting the standard theory to a non-classical logic. Next there is the Confirmation Theory program based upon Carnap's idea of a logical probability. This relies heavily on the standard Calculus of Probabilities but goes beyond Bayesianism. After looking at a more modest program concerning Proportional Syllogisms, Kyburg's continuation of Carnap's program will be considered. Kyburg, unlike Carnap, rejects the assumption of precise numerical probabilities in favor of ones which are fuzzy. Another approach that likewise rejects precise numerical probabilities is Levi's theory wherein probabilities are not so much fuzzy as indeterminate.

Again, there is a tradition going back to de Finetti of stating some axioms governing comparative probabilities, based upon the primitive relations: *p is more probable than q*; and *p and q are equiprobable*. The aim of this qualitative approach is to show that these comparisons may be faithfully represented by precise numerical probabilities.

Perhaps the central feature of the Bayesianism orthodoxy is the rule of Conditionalization, namely that the probability after discovering new evidence should equal the prior conditional probability on the supposition of that evidence. There are alternatives, notably Imaging.

Finally, there is the impact of quantum theory, including perhaps the most deviant theory of all, negative probabilities.

1 The Bayesian Orthodoxy

The Bayesian orthodoxy consists of four theses: (1) that probabilities are precise numerical representations of subjective degrees of confidence; (2) that the standard rules hold for the synchronic coherence of a system of subjective probabilities; (3) that the rule of Conditionalization holds for diachronic coherence; and (4) that the systematic con-

straints on the rationality of inferences are precisely the requirements of synchronic and diachronic coherence.[1]

The Dutch Book Argument for Bayesianism is based upon the assumption that subjective degrees of confidence have, for their behavioral manifestation, a willingness to take risks, which may be idealized as a public commitment to placing bets on the truth or falsity of any proposition. The probability is taken to equal the *betting quotient*, that is if the probability is p, the odds are in the ratio $p:1-p$. Your system of probabilities is then said to be *coherent* if and only if someone who knows the odds you are committed to cannot make a Dutch book, that is, cannot place a system of bets with you which is guaranteed not to make you anything and might result in your losing. As a consequence it is inferred that numerical probabilities may be represented to be real numbers in the interval 0 to 1 with the usual principles holding: If p entails q then $\text{Prob}(p) \leq \text{Prob}(q)$; $\text{Prob}(p\&\neg p) = 0$; $\text{Prob}(p) + \text{Prob}(q) = \text{Prob}(p \vee q) + \text{Prob}(p\&q)$; and $\text{Prob}(p|q) \times \text{Prob}(q) = \text{Prob}(p\&q)$. Here $\text{Prob}(p|q)$ is the conditional probability of p on the supposition that q, which is taken to be manifested by the preparedness to take a risk on p supposing q, itself idealized as the commitment to placing conditional bets on the truth or falsity of p, where in the event of $\neg q$ the bets are canceled. (See Kemeny (1955) for the Dutch Book Argument for these rules.)

The rule of Conditionalization may be stated thus: $\text{Prob}(p|q)$ is unchanged by the discovery (with certainty) that q.[2] The Dutch Book Argument for this, expounded by Teller (1976) but originally due to Lewis, depends on the rather strong assumption that if Conditionalization does not hold then before you discover that q you will have a public commitment to stating odds for non-conditional bets to be made when and if you discover that q. Hence this diachronic Dutch Book Argument seems less persuasive than the synchronic ones, which require only the commitment to bets here and now. There are, however, other arguments for Conditionalization (see Teller 1976).

The final Bayesian thesis is obviously defeasible but the Bayesian orthodoxy is that attempts, like Carnap's, to impose further systematic constraints have failed.[3]

2 Idealization

All systematic theories of probable reasoning are highly idealized. That is, various simplifications are assumed even though they are known to be false. The Bayesian assumption that betting should reflect degrees of confidence is one such idealization. Consider the bookies in the Dutch book argument who are out to make money from you no matter what eventuates. Some of their bets may seem to them much too generous when taken in isolation, but they make them out of a desire to profit come what may, which is perfectly reasonable. Hence Dutch bookies themselves provide examples of perfectly reasonable people whose degrees of confidence are not manifested by their betting behavior. None the less we might well accept the basic Bayesian idea of degrees of belief being represented by precise betting quotients as an idealization useful for modeling certain features of the belief system of a reasonable person.

Idealization can be avoided if we want to, but we usually do not. For instance, typically, the underlying language will be closed under various operations such as negation, conjunction, and disjunction and so contain propositions of arbitrary length. Now

we can easily put some restriction on the complexity of propositions considered. Moreover, Bayesianism is usually thought of as having the consequence that everyone should be certain of all mathematical theorems even those they do not understand. But if we put a restriction on the language being considered and if we weaken the principle that if p entails q $\text{Prob}(p) \leq \text{Prob}(q)$ so it holds only for various stated rules of inference then if p is a theorem whose proof using these rules cannot be stated in the restricted language we no longer require $\text{Prob}(q) = 1$ even if all the axioms from which it is proved are certain. Such modifications to the system are, however, tedious and so for most purposes may be ignored.

Another idealization worth noting is that we are not merely concerned with ideally rational changes to a system of subjective probabilities, but that we are assuming the person concerned is ideally rational at all times. Often, however, we change our mind because we recognize that our previous state was not (ideally) rational. So for instance new evidence might force us to take seriously a hypothesis we should have assigned a non-negligible probability to, but had dismissed.

Yet again for some purposes we shall need to consider not just the ordinary real numbers but nonstandard ones, obtained by adjoining infinitesimals. For instance, suppose we have a continuum of hypotheses depending on a parameter λ which could be any positive real number, but is unlikely to be either very small or very large. (This is an actual example based upon Carnap's Confirmation Theory. See section 5.) Then we might judge that it is as probable that $\lambda < 1$ as that $\lambda > 1$ but that it is more probable that $\lambda \geq 1$ than that $\lambda > 1$, and likewise more probable that $\lambda \leq 1$ than that $\lambda < 1$. In that case we seem forced to assign not zero but infinitesimal value to the probability that $\lambda = 1$. This is analogous to saying that the open interval of real numbers $(0,1) = \{x: 0 < x < 1\}$ is infinitesimally smaller in size than the closed interval $[0,1] = \{x: 0 \leq x \leq 1\}$. Perhaps the most natural theory of infinitesimals in this context is geometric measure theory (Schanuel 1982). Another common idealization, then, is that we ignore infinitesimals.

Bayesianism can and has been queried, even given its various idealizations (see section 7). None the less the consequences of first three theses cohere well with our intuitions and all the heterodox positions defended by me in this article either extend it (thus violating the fourth thesis) or are less idealized.

3 Two Approaches to a Theory of Probability

The chief topic of this article is probability theory as it applies to reasoning. Here there are two different approaches. The dominant strategy, illustrated by the Bayesian orthodoxy, is to think of probability theory as putting constraints on subjective probabilities and on how they change with time, where a subjective probability, written Prob, is a measure of just how confidently a proposition is asserted (if the probability is over 0.5) or denied (if it is less than 0.5). The constraints are usually considered to be normative and to be necessary conditions for rationality. The other approach, championed by Carnap (1950) but going back to Keynes (1921) and ultimately to Johnson (1921–4), is to think of an inference as having a probability which is 1 just in case the inference

is deductively valid and 0 just in case the conclusion is inconsistent with the premisses. This is called a logical probability and will be written $Prob_{Log}$. It is thought of as providing a degree of logical confirmation. In that case the conditional probabilities are more fundamental than the absolute ones which are defined thus: $Prob_{Log}(p) = Prob_{Log}(p|t)$ where t is any tautology.

We should anticipate a connection between logical and subjective probabilities. Consider the idealized situation in which all evidence is certain. Then someone – Carnap's 'Logically Omniscient Jones' – whose subjective probability equalled the logical probability on the evidence should not turn out to be irrational. That poses a problem if the probabilities are real numbers in the interval 0 to 1 inclusive. For then the two approaches differ in that for logical probabilities we have $Prob_{Log}(p|e) = 1$ only if p is entailed by e, whereas there is nothing irrational about someone who has evidence e being certain of some contingent truth p (e.g. that the whole universe did not come into existence 47,842 years ago) which is not strictly entailed by e and hence having Prob(p|e) = 1. However, harmony can be restored if we allow the logical probabilities to take nonstandard values which include infinitesimals. In that case we could say that $Prob_{log}(p|e)$ differs from 1 by an infinitesimal and that the corresponding subjective probability equals the logical probability *modulo* infinitesimals.

4 Adjustment for Nonclassical Logics

To illustrate the adjustments required if we reject the classical sentential calculus, suppose we are considering the subjective probabilities. Then the standard calculus of probabilities contains either as axioms or derived theorems:

$$Prob(p\&\neg p) = 0; \quad Prob\,(p \vee \neg p) = 1; \quad \text{and } Prob(p) + Prob(\neg p) = 1.$$

These rules do not presuppose bivalence but they do presuppose the excluded middle and non-contradiction. Thus suppose p is the Liar and, as dialethic logicians hold all four of p, ¬p, $p \vee \neg p$ and $p\&\neg p$ are logical truths. Then we have $Prob(p\&\neg p) = 1 \neq 0$. and $Prob(p) + Prob(\neg p) = 2 \neq 1$. Moreover, the standard calculus of probabilities implies a probabilistic version of the supposedly counter-intuitive rule of disjunctive syllogism, rejecting which is one of the motivations for relevance logic even when dialethic logic is not embraced. Thus we find that $Prob(p) \geq Prob(p \vee q) + Prob(\neg q) - 1$. So for instance if $p \vee q$ is asserted with 99 percent confidence, and ¬q asserted with confidence, it would be irrational to assert p with a confidence of less than 98 percent.

To accommodate heterodox logics we should adjust the calculus of probabilities. Instead of requiring merely that probabilities take values in the interval 0 to 1, we require in addition that the least upper bound of all probabilities is 1 and the greatest lower bound is 0. We have the addition rule: $Prob(p \vee q) + Prob(p\&q) = Prob(p) + Prob(q)$, and the usual multiplication rule: Prob(p|q), Prob(q) = Prob(p&q). Moreover if p entails q then $Prob(p) \leq Prob(q)$. Given classical sentential calculus we then recover the standard calculus of probabilities.

While these adjustments are fairly obvious there is an important feature of any system in which it can happen that Prob(p) + Prob(¬p) differs significantly from 1. For in such a system there is a difference between confidently denying p, which corresponds to Prob(p) being near 0, and asserting ¬p which corresponds to Prob(¬p) being near 1.

5 Carnap's Confirmation Theory

Confirmation theory is based upon a rather strong version of Foundationalism, according to which given the total evidence any proposition p should have a unique probability assigned to it, namely the logical probability of the inference with the evidence as premises and p as conclusion. Here we are to idealize the situation by ignoring the very real possibility of evidence which is itself merely probable. In addition it is assumed that the evidence is consistent. As mentioned above we would require these probabilities to conform to the constraints on the subjective probabilities of someone who has that evidence and no other evidence, but the latter will typically underdetermine the probabilities, as in orthodox Bayesianism. Hence Carnap's theory of logical probability committed him to a stronger theory than that provided by the Standard Calculus of Probability. Moreover, it is an attractive idea that the structure of propositions, explicated by the calculus of predicates, should interact with probability theory. Hence he embarked upon a research program to discover the correct assignment of $\mathrm{Prob}_{\mathrm{Log}}(p)$ for every p in the calculus of predicates with suitable constraints on the interpretation of the names and predicates. For then, given any inference with consistent premises, we may take p as the conclusion and q as the conjunction of the premisses, in which case $\mathrm{Prob}_{\mathrm{Log}}(p|q) = \mathrm{Prob}_{\mathrm{Log}}(p\&q)/\mathrm{Prob}_{\mathrm{Log}}(q)$ is the logical probability of the inference in question. Carnap had several attempts at providing a satisfactory confirmation theory. All are based upon the extremely plausible principle of symmetry, namely that since the names are assumed to lack content beyond the fact of their naming particulars the logical probabilities must be invariant under permutation of names. Presumably that is an idealization of the actual situation regarding the referring expressions occurring in natural languages, but given the idealization the symmetry principle should be uncontroversial. Carnap's method was to seek the simplest confirmation theory which met various intuitive constraints, such as that we can learn by ordinary induction. His first choice (the c*-function of the appendix to Carnap (1950)) was unique but on further consideration he came up with a continuum of confirmation theories, depending on a parameter λ which was not fixed by intuition (Carnap 1952).

Perhaps because of Carnap's penchant for technical exposition this continuum of confirmation theories is not widely studied. This is a shame for the intuitive ideas are both simple and appealing. What Carnap does is treat the logical probability of p on q as having both an *a posteriori* and an *a priori* component. Suppose 10 Fs have been observed and 9 were Gs. Suppose also that the classification of Fs is into five possible kinds of equal status of which the Gs are one kind. We want to find the logical probability of an inference from those suppositions to the conclusion that b, some unobserved F, is a G. Then the *a posteriori* component is 0.9 and the *a priori* component is 0.2, and the probability we are seeking is somewhere in between 0.2 and 0.9. Where it is in the

interval [0.2, 0.9] is determined by the weights assigned to the two components. Carnap took these weights to be the number of observed Fs, in this case 10 and the parameter λ, the weight of the a priori, so in this case the probability would be:

$$(10 \times 0.9 + \lambda \times 0.2)/(10 + \lambda) = (90 + 2\lambda)/(100 + 10\lambda).$$

Carnap's justification for the use of a linear weighting of the *a posteriori* and the *a priori* is an appeal to simplicity. Hence it should be treated as a hypothesis about probabilities which goes beyond our intuitions about them. If our aim is to find the best hypothesis then the appeal to simplicity is warranted. And as a hypothesis there is room for further empirical investigation of the most appropriate value for λ. For each value of λ we use the λ-confirmation theory to discover by induction the μ for which the μ-confirmation theory is most reliable. So for each λ there is a $\mu = f(\lambda)$, which is the value of the parameter implied by the original choice of the value λ. The only acceptable values of λ will be the fixed points, that is those for which $f(\lambda) = \lambda$. Following Carnap we should reject both very small and very large values for λ as neglecting either the *a priori* or the *a posteriori*. With good luck there might be jut one intermediate fixed point. And if that is very near an integer then simplicity would dictate that we round it off, obtaining the best hypothetical account of the unique logical probabilities. Perhaps we should have doubts as to whether such probabilities deserve to be called 'logical' but they would none the less provide a guide to reasoning.[4]

Carnap was criticized because his confirmation theory implied that induction does not justify universal generalizations such as 'All ravens are black' but only statistical generalizations such as 'At least 99.9 percent of ravens are black' and predictions such as 'The second raven I hear in 2005 will turn out to be black.' I urge readers to judge him to be right and his critics wrong, at least if we are ignoring such things as the purposes of an agent (human or divine) or laws of nature. None the less the Confirmation Theory research program was developed subsequently by Hintikka (1966) and more recently Zabell (1996), so as to arrive at a theory which allowed confirmation of universal generalizations. In Hintikka's theory there is a second parameter α, whose role may be illustrated by considering the simple case in which $\lambda = \infty$. Then if precisely n ravens have been observed, all of them black, the probability that all ravens are black is $1/(1 + 0.75^{1-\alpha})$. If n is 20 less than α, this probability is less than 0.4 percent. If n is 20 more than α then it is greater than 96.6 percent.

6 Proportional Syllogisms

Although Carnap's program was technically superb it suffered from the obvious defect that it was not applicable to anything other than a highly idealized language. Moreover his method was based upon the principle of selecting only the simplest out of a very many confirmation theories which were otherwise acceptable. But intuitively a slightly more complicated theory is only somewhat less probable than the simplest one. Hence even if there are precise logical probabilities even our best hypothesis about them will be too conjectural to command assent. This suggests two rather different ways of continuing something like the Carnapian program. One of these is to grant that we do not

know enough completely to constrain subjective probabilities but to insist that we can go beyond the Bayesians by adopting what is intuitively the most secure part of Carnap's Program, namely that in the absence of either certain or probable evidence to the contrary any set of m Fs are as likely to contain precisely n Gs as any other set of m Fs. From this it follows that in many situations we may assign precise logical probabilities to proportional syllogisms such as: Precisely K percent of Fs are Gs. This is an F, so this is a G. In the appropriate circumstances the probability of the conclusion of that inference on the premises (together with background evidence) is K/100. And one of the defects of strict Bayesianism is that there can be quite coherent systems of subjective probabilities which capriciously assign higher or lower probability to some given F being a G.

Proportional syllogisms illustrate the difficulty of finding a systematic theory of probability. For unlike deductive logic the probability of an inference depends not just on the inference schema but on the choice of predicates to substitute for the schematic letters. For instance suppose we know that there are far more rabbits than bandicoots but are otherwise ignorant of bandicoots. Perhaps we know that all rabbits love carrots. Then we know the vast majority of rabbits-or-bandicoots love carrots and so, by a careless proportional syllogism, we might infer that very probably the first bandicoot we ever meet will love carrots. Obviously something has gone wrong, but it is not that you have relevant evidence about bandicoots. It is that 'rabbit-or-bandicoot' is the wrong sort of predicate. This well-known problem applies also to the Carnapian program if we attempt to apply it to natural languages. What it shows, I submit, is that any attempt at a theory which extends Bayesianism must also contain a theory of natural kinds and natural properties.

Granted that in suitable circumstances we know about the logical probabilities of proportional syllogisms we may then rely on the Williams–Stove justification of ordinary induction (Stove 1986). This is based upon the quite uncontroversial mathematical fact that the vast majority of large samples are, to a good approximation, representative of the population as a whole. This can be made quite precise as in Stove's example (Stove 1986: 67–71). Provided a fair proportion of the population have been observed, and the circumstances are appropriate for making the proportional syllogism, we may conclude, with high probability, that the observed members of the population are, to a good approximation, representative. Moreover having some special information about the sample, for example that it has all been observed prior to 2010 is, in most circumstances, intuitively irrelevant and so does not defeat the proportional syllogism. We may note that even if all the observed Fs have been Gs, the approximate nature of the representation prevents any conclusion stronger than that some very high percentage of Fs are Gs, which is in agreement with Carnap's confirmation theory.

Both Carnap's Confirmation Theory and the more general reliance on proportional syllogisms has been criticized for 'generating knowledge out of ignorance.' If knowledge is meant quite literally then this is not the case. For instance, knowing only that a coin has two sides and it is possible to toss a coin so that either heads or tails come up we might well be very confident of not getting a run of 20 heads. This would not count as knowledge even though actual experience, say with a biased coin, might, after very many tosses have resulted in about the same degree of confidence, and, if we are not being pedantic, count as knowledge. The difference lies in the sensitivity to further

evidence. To count as knowledge even in a rather loose sense, a belief should not be too sensitive to further evidence. If, however, by 'knowledge' we just mean being almost certain then, far from seeing the generation of 'knowledge' out of ignorance as a defect, we should see this as a way of reconciling empiricism with the *a priori*.

7 Kyburg's Fuzzy Probabilities

Kyburg develops the Carnapian program by relying on proportional syllogisms.[5] That is, our knowledge of frequencies are taken as the sole determinant of the probabilities. (Ignoring for simplicity 'knowledge' of frequencies itself based upon probabilistic evidence.) Sometimes this results in precise numerical probabilities, but where there are rival proportional syllogisms they specify no precise probability but rather a (closed) interval. For instance, if we know that Tex is a Texan philosopher, that 30 percent of philosophers are vegetarians but only 10 percent of Texans are, then the resulting logical probability of Tex being a vegetarian is the closed interval [0.1, 0.3] or 0.2 ± 0.1. (If $a \leq b$, the closed interval $[a,b] = \{a \leq x \leq b\}$.) We may think of these as fuzzy numbers, which may be added and multiplied, using the rules $[a,b] + [c,d] = [a + c, b + d]$, $[a,b] \times [c,d] = [a \times c, b \times d]$. Now the Addition Rule still holds, namely $\text{Prob}(p\&q) + \text{Prob}(p \vee q) = \text{Prob}(p) + \text{Prob}(q)$, except that we are adding fuzzy numbers.

The most radical of Kyburg's theses is his rejection of the standard multiplication principle that $\text{Prob}(p\&q) = \text{Prob}(p|q), \text{Prob}(q)$, even in the special case in which the probabilities are precise. This is based upon the use of proportional syllogisms to specify probabilities. An example of Kyburg's (Harper 1982: 120) illustrates this nicely. Suppose in the whole population 50 percent are R (red) and 40 percent of Rs are also S (square). Now suppose a sample of 100 is taken out of this population. We are not told what the proportion of Rs is in the population but we are told that not 40 percent but 50 percent of the Rs in the sample are also Ss. Now consider one member of the sample. Because of Kyburg's reliance on proportional syllogisms he asserts that $\text{Prob}(Sx \vee Rx) = 0.5$ and $\text{Prob}(Rx) = 0.5$, but $\text{Prob}(Rx\&Sx) = 0.2$, not the 0.25 it should be according to the multiplication rule. Here I think Kyburg is mistaken. Intuitively our knowledge that in the sample 50 percent of Rs are Ss combined with what we already know about the whole population provides some information about the way in which the sample is unrepresentative. Quite how we use this information is not clear but it shows that the conditions are not appropriate for making the proportional syllogism. If so, Kyburg is mistaken about the consequences of his own reliance on proportional syllogisms, but the underlying theory of fuzzy logical probabilities is still tenable.

8 Levi's Indeterminate Systems

Kyburg's fuzzy probabilities are, as logical probabilities, assigned to inferences. As such they have their (fuzzy) values regardless of the probabilities of other propositions. By contrast Levi refines the theory of subjective probabilities by not requiring determinate numerical probabilities. Instead the actual system is represented by means of a set of

credence functions assigning precise numbers to propositions, each one satisfying the standard diachronic and synchronic principles as accepted by Bayesians.[6] So if we follow van Fraassen's method of supervaluation we would say that all the Bayesian rules are definitely true. In addition Levi requires that the family of credence functions be convex. That is, if f and g are two credence functions in the set S representing the system of subjective probabilities, and if $0 < \alpha < 1$, then $\alpha f + (1 - \alpha)g$ is also a member of S.

As in the case of precise numerical probabilities there should be a connection between subjective probabilities and logical ones. We would expect that any rational indeterminate probability of the kind described by Levi should assign probabilities to propositions within the interval assigned as a logical probability by a Kyburg-type theory. In that case we could not, of course, follow both Kyburg, in his nonstandard rule for conditional probabilities, and Levi.

9 Qualitative Theories of Probability

Many have thought that it would be more fundamental to consider a purely qualitative system of subjective probabilities based upon a ranking of propositions as more or less probable.[7] Here we have two relations: the transitive and anti-reflexive relation of being more probable ($p > q$) and the equivalence relation of being equiprobable, ($p \sim q$). Moreover if $p \sim q$, $q > r$ and $r \sim s$ then $p > s$.

Because of the usual idealization we may assume (Axiom 1) that if p entails q, then $q \geq p$ (i.e. either $p \sim q$ or $q > p$). Write $p \perp q$ if p and q are inconsistent. Then we have:

AXIOM 2 the intuitive rule that if $p \perp q$, if $r \perp s$, if $p \geq r$, if $q \geq s$ and if $r \vee s \geq p \vee q$ then $p \sim r$, $q \sim s$, and $p \vee q \sim r \vee s$.

Idealizing the situation so as to assume logical omniscience, we also have a two-part Completability Principle (modeled on Ellis 1979: 9–16), as follows.

AXIOM 3 There must be an extension of the ranking to one satisfying both the principles governing qualitative probability and *trichotomy* (i.e. given any p, q either $p > q$ or $p \sim q$ or $q > p$). Moreover, if all such extensions agree that $p > q$, or agree that $p \sim q$, then we already have $p > q$, or $p \sim q$, respectively.

From these three axioms it is easy to show that if $p \geq q$, or $p \sim q$, then $\neg q > \neg p$, or $\neg q \sim \neg p$, respectively, which would otherwise have been assumed as an axiom.

We say a credence function d is *commensurate* with the qualitative system if whenever a proposition p is more probable than proposition q then $d(p) > d(q)$ and whenever p and q are equiprobable then $d(p) = d(q)$. Unfortunately Axioms 1 to 3 for qualitative probabilities stated thus far do not ensure the existence of commensurate credence functions even for a finite system of propositions. What is required is a strengthening of Axiom 2, as follows:

AXIOM 2* For any m,n, suppose r_j, $j = 1, \ldots, n$ are pairwise inconsistent, that is $r_j \perp r_k$ if $j \neq k$. And suppose that the $p_k = /\{r_j; j \in A_{kj}\}$, $q_k = /\{r_j; j \in B_{kj}\}$, where, for all k, A_{kj} and

B_{kj} are subsets of $\{1, \ldots, n\}$. Further suppose that $p_k \geq q_k$, for all k. Finally, suppose that $Vp_k = Vq_k$, then if $\#B_{kj} \geq \#A_{kj}$ for all k, j then $p_k \sim q_k$ for all k.

It is easy to see that Axiom 2* is a necessary condition for there being a commensurate credence function. For a finite system of propositions any system of qualitative probabilities satisfying Axiom 1, Axiom 2*, and Axiom 3 has a commensurate credence function.[8] Moreover the set C of all commensurate credence functions is convex and so forms a Levi system, which is easily seen to be a faithful representation in the sense that if for all $d \in C$ $d(p) > d(q)$, or $d(p) = d(q)$ then $p > q$, or $p \sim q$ respectively. That this does not extend to the infinite case could be shown by considering an example in which it would be more appropriate to consider probabilities which can take infinitesimal values than the ones actually being considered which are real valued only.[9]

If Axiom 2* is intuitive then this is a further way of justifying Levi's system. Otherwise it suggests that qualitative probabilities form an interestingly weaker kind of system of subjective probabilities.

10 The Dynamics of Subjective Probability

Carnap's theory of logical probability leaves no room for an interesting dynamics for probabilities. For if the total evidence changes from e^- to e^+ then the probability of the inference from the total evidence to some conclusion p changes from $Prob_{Log}(p|e^-)$ to $Prob_{Log}(p|e^+)$ without any change in $Prob_{Log}(p|q)$ for any p or q. In the theory of subjective probabilities the orthodoxy is the rule of conditionalization, according to which the new subjective probability $Prob^+(p)$ on coming to be certain of new evidence e equals the old conditional probability $Prob^-(p|e)$ provided $Prob^-(e) > 0$. Now the Dutch Book Argument for Conditionalization required commitment to the same *rule* by which subjective probabilities change. Any rule other than conditionalization results in the possibility of a Dutch Book. This does not exclude a position even more subjective than Bayesianism, namely resisting the suggestion that there be a rule governing the dynamics of belief. In spite of the Dutch Book Argument alternative rules such as *imaging* have been suggested. The difference between conditionalization and imaging is most easily seen by taking the probability distributions to be given by a probability measure on the set of possible worlds. The effect of coming to be certain of e is to excise all the $\neg$e-worlds, and redistribute their probability to the e-worlds. Conditionalization does this by preserving the relative probabilities of the e-worlds.

Imaging does this by re-assigning the probability previously assigned to an $\neg$e-world to the nearest e-world(s). In addition to the Dutch Book Argument for conditionalization there is Gärdenfors' argument based upon the plausible principle that if initially $Prob(q) > 0$ and $Prob(p) = 1$ then after discovering that q $Prob(p)$ should remain equal to 1. From this it is argued that no principle governing the change of subjective probabilities can contradict conditionalization. (See Gärdenfors (1988) for this and a more general discussion of imaging versus conditionalization. See also Teller (1976) for a defence of conditionalization.)

Gärdenfors' case for conditionalization illustrates once again the role of idealization. Suppose in fact Prob(e) is initially positive but rather small. Then the discovery that e

might quite naturally prompt a reconsideration of the previous confidence that ¬e, resulting in a backdated change to the earlier subjective probabilities. It is assumed, however, that the person concerned is ideally rational at all times and so never has occasion to regret the earlier confidence.

11 Probability Theory and Quantum Theory

Quantum theory is open to many rival interpretations. But for a long time the most popular was to think of the state of a quantum theoretic system as specified by the probabilities of the affirmative answer if various 'questions' are asked, that is if various two-valued 'observations' are performed.[10] Often these 'observations' concern humanly unobservable entities such as quarks. So we are here considering an ideal observer. Obviously such two-valued questions correspond to propositions, so what started off as a physical theory is treated as if it were a theory of probability. However the underlying 'Sentential Calculus' is very far from classical. It is a Quantum Logic (qv) which in the most straightforward case may be represented as the lattice of closed subspaces of a Hilbert Space. The usual principles for the Calculus of Probability hold provided we replace conjunction by the intersection, disjunction by the sum, and negation by the orthogonal complement. In addition we require the probability distribution to be additive over the countable 'disjunction' of pairwise orthogonal observables. In these circumstances Gleason's Theorem tells us that, with the exception of the special case in which the Hilbert space has only two dimensions, any such probability distribution will be a mixture of pure states specified by vectors in the Hilbert space, as in the formalism of quantum mechanics.[11]

There is a rather commonsensical retort to this sophisticated but curious account. It is to insist that the propositions which correspond to the idealized observations be embedded in a larger system to include ones which are considered beyond even an ideal observer to observe. Then, it is hoped an ideal observer who was also an ideal reasoner could assign either precise numerical probabilities, or perhaps a Levi style family of credence functions, to all propositions so as to agree with the formalism in the case of the observables. The problem with this is that for many propositions the probabilities assigned would seem to be negative (Wigner 1932). In fact the currently popular Consistent Histories formulation (Omnès 1994: 122–43) restricts the propositions considered to just those which have probabilities no greater than 1 and no less than 0. I hold that this is all quite unnecessary because physicists have mistaken a mean value for a probability. There is some relevant quantity Q (mass, charge, or the number of particles of the kind considered minus the number in a quantum vacuum) which can be positive or negative, but whose mean value for the whole system is that of a single classical particle. Then the 'probability' assigned to the proposition that 'it' has such and such position, momentum, spin, etc. is in fact to be interpreted as the mean value of Q for all states such that p (see Forrest 1999).

If readers decline my kind offer to render quantum mechanics compatible with common sense, they might prefer a Kyburg-style alternative. We could assign to all propositions not allowed on the consistent histories approach, the default fuzzy probability [0,1].

Notes

1 For a defence of Bayesianism see Earman (1992).
2 The case in which the new evidence is merely probable is discussed by Jeffrey (1965).
3 Given the first three Bayesian theses, once the initial unconditional probabilities are specified, all future unconditional probabilities are then determined, provided none of the new evidence had previously had zero probability. So perhaps we should qualify the fourth thesis by allowing additional rules to govern the case in which the evidence did previously have zero probability. For instance there is the Levi–Gärdenfors method of Preservative Imaging (see Gärdenfors 1988: 117–18).
4 See also Carnap's remarks on what would, contingently, pick out one value of the parameter (Carnap 1952: chapter 3). For a more recent investigation of empirical constraints on logical probabilities, see Nolt (1990).
5 See Kyburg (1974), but perhaps the most accessible introduction to Kyburg's work is (Bogdan 1982), especially (Spielman 1982).
6 A good introduction to Levi's theory is Levi (1980). See also Bogdan (1982).
7 A first, inadequate, axiomatization of qualitative probabilities is found in De Finetti (1932). For a useful survey of some systems see the Appendix to Malmnäs (1981).
8 Axiom 2* is equivalent to strong coherence in the sense of (Malmnäs 1981: 17). Here we identify a proposition with the set of all the 'possible worlds' at which it is true, where the 'possible worlds' may in turn be thought of as maximal consistent sets of propositions. The existence of commensurate credence functions then follows from Theorem 1 (Malmnäs 1981: 31).
9 Consider a Boolean algebra with countably many atoms all of which are equiprobable. We may arrange for the axioms of qualitative probability to be satisfied yet there is no commensurate credence function.
10 See Mackey (1963) for one of the early expositions of this approach. See also Hooker (1973).
11 For a lucid exposition of the formalism of quantum theory, including Gleason's Theorem, see Hughes (1989).

References

Bogdan, Radu J. (1982) *Henry E. Kyburg, Jr. and Isaac Levi*. Dordrecht: Reidel.
Carnap, Rudolph (1950) *Logical Foundations of Probability*. Chicago: University of Chicago Press.
Carnap, Rudolph (1952) *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.
De Finetti, B (1932) La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7, 1–68.
Earman, John (1992) *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.
Ellis, Brian (1979) *Rational Belief Systems*. Oxford: Blackwell.
Forrest, Peter (1999) In defence of the phase space picture. *Synthese*, 119, 299–311.
Gärdenfors, Peter (1988) *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. Cambridge, MA: Bradford/MIT.
Hailperin, Theodore (1990) Probability logic in the twentieth century. *History and Philosophy of Logic*, 71–110.
Harper, William L. (1982) Kyburg on direct inference. In Bogdan (1982), 97–128.

Hintikka, Jaakko (1966) A two-dimensional continuum of inductive methods. In Jaakko Hintikka and Patrick Suppes (eds.), *Aspects of Inductive Logic* (pp. 113–32). Amsterdam: North Holland.

Hooker, Cliff (ed.) (1973) *Contemporary Research in the Foundations and Philosophy of Quantum Theory*. Boston, MA: Reidel.

Hughes, R. I. G. (1989) *The Structure and Interpretation of Quantum Mechanics*. Cambridge, MA: Harvard University Press.

Jeffrey, R. C. (1965) *The Logic of Decision*. New York: McGraw-Hill.

Johnson, W. E. (1921–4) *Logic*, vols. 1–3. Cambridge: Cambridge University Press.

Kemeny J. (1955) Fair bets and inductive probabilities. *Journal of Symbolic Logic*, 20, 263–73.

Keynes, John Maynard (1921) *A Treatise on Probability*. London: Macmillan.

Kyburg, Henry E. (1974) *The Logical Foundations of Statistical Inference*. Dordrecht: Reidel.

Levi, Isaac (1980) *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. Cambridge: MIT Press.

Mackey, George W. (1963) *The Mathematical Foundations of Quantum Mechanics*. Amsterdam: Benjamin.

Malmnäs, Per-Erik (1981) *From Qualitative to Quantitative Probability (Stockholm Studies in Philosophy 7)*. Stockholm: University of Stockholm Press.

Nolt, John (1990) A fully logical inductive logic. *Notre Dame Journal of Formal Logic*, 415–36.

Omnès, Roland (1994) *The Interpretation of Quantum Mechanics*. Princeton, NJ: Princeton University Press.

Schanuel, Stephen H. (1982) What is the length of a potato? An introduction to geometric measure theory. In A. Dold and B. Eckmann (eds.), *Categories in Continuum Physics, Lecture Notes in Mathematics* (pp. 118–25). Berlin: Springer.

Spielman, Stephen (1982) Kyburg's system of probability. In (Bogdan 1982), 57–96.

Stove, D. C. (1986) *The Rationality of Induction*. Oxford: Clarendon Press.

Teller, Paul (1976) Conditionalization, observation, and change of preference. In W. L. Harper and C. L. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*. Dordrecht: Reidel.

Wigner, E. (1932) On the quantum correction for thermodynamic equilibrium. *Physical Review*, 40, 749–59; reprinted in Kim, Y. S. and Noz, M. E. (eds.), *Phase Space Picture of Quantum Mechanics: Group Theoretical Approach*. Singapore: World Scientific (1991) 219–31.

Zabell, S. L. (1996) Confirming universal generalisations. *Erkenntnis*, 45, 267–283.

37

Why Fuzzy Logic?

PETR HÁJEK

It is generally understood that fuzzy logic deals with vague, imprecise notions and propositions. In spite of several successful applications, the logician may (and should) ask: is this really a logic? Does it have foundations, mathematical and/or philosophical? I shall try to give a positive answer to this question, at least as mathematical foundations are concerned, leaving philosophical foundations to professional philosophers. Due to space limitation, I can offer only a survey; but the interested reader will find enough references to detailed works.

1 Origin

Lotfi Zadeh is the author of the notion of a fuzzy set; his 1965 paper is a landmark (Zadeh 1965). A fuzzy subset X of a set A is given by its characteristic function μ_X assigning to each element $a \in A$ the degree $\mu_X(a)$ in which a belongs to X; $\mu_X(a)$ is a real number from the unit interval $[0, 1]$. Natural language offers plenty of examples: think, for example, of a set of people and its fuzzy subset of tall people (some are more tall, some less). Naturally, one can similarly speak on fuzzy propositions, some being more true and some less ('John is tall'). Apparently the term 'fuzzy logic' first occurs in (Goguen 1968–9) with a elucidating title "The logic of inexact concepts." The beginning of numerous applications of such fuzzy logic is Mamdani (1974), where the author describes a controller based on "fuzzy IF-THEN rules." Such rules are nowadays very popular and may look for example as follows: 'If the pressure is high and the increase of pressure is high then turn the wheel far to the left.' You see various fuzzy notions; for example the meaning of high pressure is to be understood as a fuzzy subset of the domain of pressures: each pressure is high is some degree. Observe the use of natural language (Zadeh likes to speak on "computing with worlds"). There is also some rudimentary logic ('and', 'if-then') but not much.

2 Many-Valued Logic

Clearly, the above resembles some many-valued logic; but for a long time, there were nearly no contacts between what was called fuzzy logic and the many-valued logic

entertained by logicians. Early examples are Giles (1976) and Pavelka (1979). Recall that 20th-century many-valued logic started in the 1920s and 1930s in the work of Jan Łukasiewicz (1930; Łukasiewicz and Tarski 1930); later there were works on many-valued logic related to intuitionistic logic (A. Heyting, K. Gödel (1932)). Their work was continued by several authors (Dummett, Chang, Moisil, McNaughton, Scarpelini and others), Gottwald in his 1988 German book on many-valued logic has a short chapter relating many-valued logic to fuzzy logic. Note that Gottwald's book is to appear soon in a revised English version (Gottwald forthcoming). In the meantime, plenty of papers appeared claiming to deal with fuzzy logic but being logically uninteresting. Mutual contacts developed rather slowly.

3 Fuzzy Logic in a Broad and Narrow Sense

It turned out that one has to distinguish two notions of fuzzy logic. It was again Zadeh who coined the terms "fuzzy logic in broad (or wide) and narrow sense." In a broad sense, the term 'fuzzy logic' has been used as anonymous with 'fuzzy set theory and its applications'; for good monographs on this logic see Zimmermann (1991) and Klir and Yuan (1995); in the emerging narrow sense, fuzzy logic is understood as a theory of approximate reasoning based on many-valued logic. Zadeh (1994) stresses that the questions of fuzzy logic in the narrow sense differ from usual questions of many-valued logic and concern more questions of approximate inferences than those of completeness, etc.; nevertheless, with full admiration to Zadeh's pioneering and extensive work (see Klir and Yuan 1996) a logician will *first* study classical logical questions on completeness, decidability, complexity, etc. of the symbolic calculi in question and *then* try to reduce the question of Zadeh's agenda to questions of *deduction* as far as possible. This is the approach in my monograph (Hájek 1998), which I sketch below.

4 The Basic Fuzzy Propositional Calculus

The calculus we are going describe is a result of the following 'design choices' (they are not obligatory but are apparently rather reasonable:

1. The real unit interval $[0, 1]$ is taken to be the *standard set of truth values*, 1 meaning absolute truth, 0 absolute falsity. The usual ordering $\leq$ of reals serves as a comparison of truth-values; we build the logic as a logic with a *comparative notion of truth*. Other structures of truth-values, possibly only partially ordered, are not excluded.
2. The logic is *truth-functional*, that is connectives are interpreted via their truth functions; then for example the truth-value of a conjunction $\varphi\&\psi$ is uniquely determined by the truth-value of φ, of ψ and by the chosen truth function of &.
3. *Continuous t-norms* are taken as possible truth functions of *conjunction*. These operations are broadly used by a fuzzy community; a binary operation * on $[0, 1]$ is a

t-norm if it is commutative ($x*y = y*x$), associative ($x*(y*z) = (x*y)*z$), non-decreasing in each argument (if $x \le x'$ then $x*y \le x'*y$ and dually) and 1 is a unit element ($1*x = x$). The t-norm * is a continuous t-norm if it is continuous as a real function. The three most important continuous t-norms are:

$x*y = \max(0, x + y - 1)$ (Łukasiewicz *t*-norm),
$x*y = \min(x, y)$ (Gödel *t*-norm),
$x*y = x \cdot y$ (product *t*-norm).

(For the names see Historical remarks in Hájek (1998).) Note in passing that each continuous t-norm is built from these three in a certain way.

4. The truth function of *implication* is the *residuum* of the corresponding t-norm. If * is your continuous t-norm then its residuum is the operation $\Rightarrow$ defined as follows:

$$x \Rightarrow y = \max\{z | x*z \le y\}.$$

Note that $x \Rightarrow y = 1$ iff $x \le y$; for $x > y$ the residua of the above *t*-norms are

$x \Rightarrow y = 1 - x + y$ (Łukasiewicz),
$x \Rightarrow y = y$ (Gödel),
$x \Rightarrow y = y/x$ (product).

(One calls these implications *R*-implications, *R* for residuum.)

5. The truth function of negation is $(-)x = x \Rightarrow 0$ (x implies falsity).

The resulting logic is called BL – the basic fuzzy propositional logic. We sketch its main properties.

Work with propositional variables $p_1, p_2 \ldots$ and connectives &, $\rightarrow$ (strong conjunction, implication) and truth constant $\bar{0}$ (falsity). Formulas are defined in obvious way; $\neg\varphi$ stands for $\varphi \rightarrow \bar{0}$. Given a continuous t-norm * (and thus its residuum $\Rightarrow$), each evaluation e of propositional variables by truth degrees from [0, 1] extends to an evaluation e_* of all formulas; thus $e_*(\bar{0}) = 0$, $e_*(\varphi\&\psi) = e_*(\varphi)*e_*(\psi)$, $e_*(\varphi \rightarrow \psi) = e_*(\varphi) \Rightarrow e_*(\psi)$. Call φ a **-tautology* if $e_*(\varphi) = 1$ for each evaluation e; call φ a *t-tautology* if it is a **-tautology for each * (i.e. however you interpret your propositional variables and connectives, φ is true).

The following t-tautologies are taken to be *axioms* of BL:

(A1) $(\varphi \rightarrow \psi) \rightarrow ((\psi \rightarrow \chi) \rightarrow (\varphi \rightarrow \chi))$
(A2) $(\varphi\&\psi) \rightarrow \varphi$
(A3) $(\varphi\&\psi) \rightarrow (\psi\&\varphi)$
(A4) $(\varphi\&(\varphi \rightarrow \psi)) \rightarrow (\psi\&(\psi \rightarrow \varphi))$
(A5a) $(\varphi \rightarrow (\psi \rightarrow \chi)) \rightarrow ((\varphi\&\psi) \rightarrow \chi)$
(A5b) $((\varphi\&\psi) \rightarrow \chi) \rightarrow (\varphi \rightarrow (\varphi \rightarrow \chi))$
(A6) $((\varphi \rightarrow \psi) \rightarrow \chi) \rightarrow (((\psi \rightarrow \varphi) \rightarrow \chi) \rightarrow \chi)$
(A7) $\bar{0} \rightarrow \varphi$

The deduction rule is *modus ponens* (from φ and $\varphi \to \psi$ infer ψ), proofs and provability are defined in the obvious way.

COMPLETENESS: For each formula φ, BL proves φ iff φ is a t-tautology.

(For a proof see Cignoli et al. (submitted); Hâjek (1998) presents another completeness for BL, relating provability in BL to tautologicity with respect to so-called *BL-algebras*. Each continuous t-norm defines a BL-algebra but not conversely.)

The three important t-norms defined above (Ł – Łukasiewicz, G – Gödel, Π – product) give us three important and well-known logics stronger than BL:

Łukasiewicz logic can be axiomatized by adding the schema of double negation $\varphi \equiv \neg\neg\varphi$ to BL. Formulas provable in this logic (developed also by Ł) are exactly all Ł-tautologies. (See Cignoli et al. (2000) for extensive analysis and deep theory of Łukasiewicz logic.)

Gödel logic G (related to Godel (1932)) is BL plus the schema $\varphi \equiv (\varphi\&\varphi)$ of idempotence of conjunction. Formulas provable in G are exactly all G-tautologies.

Product logic Π is BL plus two additional axioms $(\varphi \to \neg\varphi) \to \neg\varphi$ and $\neg\neg\chi \to (((\varphi\&\chi) \to (\psi\&\chi)) \to (\varphi \to \psi))$. (The latter axiom expresses cancellation by a non-zero element.) Π proves exactly all Π-tautologies.

It should be mentioned that G contains the *intuitionistic logic* (so G is an intermediate logic between intuitionistic and classical logic). Neither Ł nor Π contain intuitionistic logic since they have a non-idempotent conjunction.

In BL we may define derived connectives: min-conjunction $(\varphi \wedge \psi) \equiv \varphi\&(\varphi \to \psi)$ whose truth function is a minimum, and max-disjunction $(\varphi \vee \psi) \equiv (((\varphi \to \psi) \to \psi) \wedge ((\psi \to \varphi) \to \psi)$ (maximum).

The truth function $(-)$ of negation in Ł is $(-)x = 1 - x$; but the negation of G is Gödel negation: $(-)0 = 1$, $(-)x = 1$ for $x > 0$. Also Π has Gödel negation.

This means that in general the strong conjunction has in BL no dual disjunction; only in Ł, whose negation is involutive, that is $(-)(-)x = x$, the strong disjunction $(\varphi \oplus \psi) \equiv \neg(\neg\varphi\&\neg\psi)$ behaves well. But you may extend both G and Π by Łukasiewicz negation (if you want to work with so-called t-conorms; see Esteva et al. (2000)) for a reasonable axiomatization.

Another important extension results when we add to Łukasiewicz logic truth constant $\bar{r}$ for each rational $r \in [0, 1]$ (Pavelka logic), postulating $e_{\mathrm{t}}(\bar{r}) = r$. Then evidently $e_{\mathrm{t}}(\bar{r} \to \varphi) = 1$ iff $e_{\mathrm{t}}(\varphi) \geq r$, which gives us the possibility of expressing estimates of the truth degree of a formula. This extension of Ł has very pleasing properties; an analogous extension of G or Π is more complicated. We note in passing that for example. Novák et al. (2000) considers Pavelka logic to be *the* fuzzy logic; I do not share this opinion.

Summarizing this section, continuous t-norm propositional logics are well understood, have pleasant properties and are presently the subject of intensive study.

5 The Basic Fuzzy Predicate Calculus

Extending the developed propositional calculus to a predicate calculus is very natural and a generalization of Tarskian truth definition is immediate. Take some predicates P_1, ..., each having its arity (unary, binary, ...), object variables $x, y, \ldots$, connectives &,

$\rightarrow$, truth constant $\bar{0}$, quantifiers $\forall$, $\exists$. (We disregard object constants and function symbols for simplicity.) Formulas are defined in the usual way. An *interpretation* (of $P_1, \ldots, P_n$) is a structure $\mathbf{M} = (M, (r_P)_{P \text{ predicate}})$ where M is a nonempty set (domain) and for each predicate P of arity n, r_P is an n-ary fuzzy relation on M, that is a mapping associating with each n-tuple $(a_1, \ldots, a_n)$ of elements of M a truth degree $r_P(a_1, \ldots, a_n) \in [0, 1]$. The *truth-value* of a formula φ in $\mathbf{M}$ depends (besides $\mathbf{M}$) on a given evaluation e of object variables by elements of M (an M-evaluation, actual meaning of variables) and on the chosen semantics of connectives, that is on the t-norm $*$. We write $\|\varphi\|^*_{M,e}$ for this. It is defined inductively as follows:

$\|P(x_1, \ldots, x_n)\|^*_{M,e} = r_P(e(x), \ldots e(x_n))$;
$\|\varphi\&\psi\|^*_{M,e} = \|\varphi\|^*_{M,e} * \|\varphi\|^*_{M,e}$
$\|\varphi \rightarrow \psi\|^*_{M,e} = \|\varphi\|^*_{M,e} \Rightarrow \|\psi\|^*_{M,e}$
$\|(\forall x)\varphi\|^*_{M,e} = \inf_{e_x} \|\varphi\|^*_{M,e}$
$\|(\exists x)\varphi\|^*_{M,e} = \sup_{e_x} \|\varphi\|^*_{M,e}$

where e_x runs over all evaluations differing from e at most in the value for the argument x. The atomic case can be paraphrased thus: the formula saying that $(x_1, \ldots, x_n)$ are P has the truth-value equal to the degree in which the objects $e(x_1) \ldots e(x_n)$ (being the meanings of $x_1, \ldots x_n$) are in the relation r_P (which is the meaning of P). The definitions for $\forall$, $\exists$ naturally generalize the two-valued case. Now the reader expects the following definitions:

A formula φ is a $*$-tautology (of the predicate calculus) if $\|\varphi\|^*_{M,e} = 1$ for each interpretation $\mathbf{M}$ and M-evaluation e. φ is a *t*-tautology if it is a $*$-tautology for each $*$.

We may call φ $*$-*true* in $\mathbf{M}$ of $\|\varphi\|^*_{M,e} = 1$ for each e. Thus φ is a $*$-tautology if φ is $*$-true in each interpretation.

Note that this may be generalized from *t*-norms to BL-algebras; then r_P is a mapping into the domain of the algebra. But for quantified formulas $\|\varphi\|^*_{M,e}$ ($\mathbf{L}$ being a BL-algebra) may be defined if the corresponding infimum/supremum does not exist in $\mathbf{L}$. One defines an $\mathbf{L}$-*safe* interpretation to be an $\mathbf{L}$-interpretation in which $\|\varphi\|^{\mathbf{L}}_{M,e}$ is total; φ is an $\mathbf{L}$-*tautology* if it is $\mathbf{L}$-true in each $\mathbf{L}$-safe interpretation.

The basic fuzzy predicate logic BL$\forall$ has the above axioms for BL and the following axioms for quantifiers:

($\forall$1) $(\forall x)\varphi(x) \rightarrow \varphi(y)$
($\exists$1) $\varphi(y) \rightarrow (\forall x)\varphi(x)$
($\forall$2) $(\forall x)(\chi \rightarrow \psi) \rightarrow (\chi \rightarrow (\forall x)\psi)$
($\exists$2) $(\forall x)(\varphi \rightarrow \chi) \rightarrow ((\exists x)\varphi \rightarrow \chi)$
($\forall$3) $(\forall x)(\varphi \vee \chi) \rightarrow ((\forall x)\varphi \vee \chi)$

These formulas are well-known from classical logic; they are all predicate *t*-tautologies (and even BL-tautologies – are $\mathbf{L}$-true in each safe $\mathbf{L}$-interpretation).

Deduction rules are modus ponens and generalization (from φ infer $(\forall x)\varphi$) – as in classical logic.

EXERCISE. Just for refreshment, take the trivial example 5.1.2 from Hájek (1998): $M = \{1, 2, 3\}$, binary predicate *likes*. r_{likes} given by the table

	1	2	3
1	1	0.3	0.7
2	0.9	0.9	0
3	0.9	0.1	0.2

Compute the truth value of $(\forall x, y)(\text{likes}(x, y) \rightarrow \text{likes}(x, x))$ (saying 'everybody likes himself/herself most') for Ł, G, Π. (Hint: for Ł it is 0.9.)

What about *completeness?* Note that BL∀ is complete will respect to general interpretation: BL∀ proves φ iff φ is an **L**-tautology for each BL-algebra **L**. Similarly, the predicate versions Ł∀, G∀, Π∀ of the corresponding propositional logics *are* complete with respect to (safe) interpretations over algebras from the corresponding subclasses of the class of BL-algebras (called MV-algebras, G-algebras and product algebras for Łukasiewicz, Gödel, and product logic respectively.

With respect to interpretations over [0, 1] the situation is more complicated: the set of all predicate t-tautologies (i.e. formulas being *-tautologies for each continuous t-norm *) is not recursively enumerable (for specialists: it is Π_2-hard). Similarly, neither the set of predicate Łukasiewicz tautologies (tautologies w.r.t. Łukasiewicz t-norm) nor the set of predicate product tautologies is recursively axiomatizable. (For BL∀ not yet published; for Ł∀ first proved by Scarpelini, see Hájek (1998).) But the set of predicate G-tautologies *is* completely axiomatized by BL∀ plus the axiom schema $\varphi \equiv (\varphi\&\varphi)$.

To get a full picture of these logics one has to have some knowledge about formulas provable in them; this is found in Hájek (1998). (Such knowledge is necessary for *proofs* of completeness results.)

Again, various extensions of these logics have been described. Furthermore, there are results on theories over these logics; we have no room to go into details. Similarly as above, let us summarize that the basic fuzzy predicate calculus is reasonably well developed and well behaving. Concerning the results on non-axiomatizability, compare this with the situation of classical second order logic: in the intended standard semantics it is not recursively axiomatizable, but it has a recursive axiomatization which is complete with respect to a generalized (Henkin) semantics.

In the rest of this chapter we shall describe some uses of fuzzy logic that may be of interest for the philosophically minded reader.

6 Similarity

Similarity is a fuzzy equality; the notion appears to be well-known in the fuzzy community. Let $x \approx y$ stand for 'x is similar to y'; the following are axioms of similarity:

$x \approx x$ (reflexivity)
$x \approx y \rightarrow y \approx x$ (symmetry)
$(x \approx y \& y \approx z) \rightarrow x \approx z$ (transitivity).

What do models of these axioms look like? First observe a non-model: For M being the real line define x, y to be 'similar' if $|x - y| \leq 1$. This is a crisp relation (yes–no) and is *not* transitive: 3 and 4 are 'similar,' 4 and 5 also, but 3 and 5 not. *Make it fuzzy*: define

$$r \approx (x, y) = \max(0, 1 - |x - y|).$$

EXERCISE: Draw the graph of the function $x \approx 4$: it is zero for $x \leq 3$ and $x \geq 5$ and goes up linearly from the point (3, 0) to (4, 1) and the down linearly from (4, 1) to (5, 0).

Is this relation transitive? It depends on your logic. The axiom of transitivity does say that if $x \approx y$ and $y \approx z$ are (absolutely) true then so is $x \approx z$; but it says *much more*, namely that the truth degree of $x \approx y \,\&\, y \approx z$ is less than or equal to the truth degree of $x \approx z$. Take Łukasiewicz logic and compute:

$$\|x \approx y \& y \approx z\| = \max(0, \|x \approx y\| + \|y \approx z\| - 1).$$

If this is 0 nothing is to be proved; otherwise continue:

$$\|x \approx y\| + \|y \approx z\| - 1 = 1 - |\, x - y| + 1 - |y - z| - 1 = 1 - (|x - y| + |y - z|) \leq 1 - |x - z|$$

by the well-known triangle of inequality; and the last term equals $\|x \approx z\|$. Thus we have verified that the truth value of the transitivity axiom is 1 (the axiom is absolutely true) for our interpretation.

Similar examples for Gödel and product logic are easy to find. Now observe that if $\approx$ satisfies the axioms of similarity and we define $x \approx^2 y$ to be $x \approx y \,\&\, x \approx y$ then $\approx^2$ is again a similarity; for example in our example $r\approx 2(x, y) = \max(0, 1 - 2|x - y|)$. For more information see Hájek (1998).

7 The Liar and Dequotation

Here I assume some knowledge of Gödel's technique of self-reference in arithmetic. **N** stands for the structure of natural numbers with zero, successor, addition, and multiplication. PA is Peano arithmetic. The undefinability of truth in arithmetic means the following: Add a unary predicate Tr to the language of arithmetic and add the axiom schema of dequotation: $\varphi \equiv Tr(\overline{\varphi})$ to the axioms of PA (φ being an arbitrary sentence of the langauge of PA extended by the predicate Tr, and $\overline{\varphi}$ being the numeral naming the Gödel number of φ). Then the resulting theory (PA+Tr) is contradictory over classical logic since one can construct the liar's formula λ such that (PA+Tr) proves $\lambda \equiv \neg Tr(\overline{\lambda})$ and hence proves $\lambda \equiv \neg\lambda$, which is classically inconsistent. Over Łukasiewicz logic the last equivalence is not contradictory, it just forces the truth-value of λ to be 1/2. But we may ask more: Take, inside Łukasiewicz logic, crisp Peano arithmetic, add the predicate Tr (which may be fuzzy) and add the dequotation schema. Is this theory consistent (over Łukasiewicz)?

This was answered in Hájek et al. (2000) as follows: (PA+Tr) *is* consistent over the Łukasiewicz predicate logic, hence it has a model (which is crisp for arithmetic and fuzzy

for *Tr*); but the standard model **N** cannot be expanded by a fuzzy predicate to a model of (PA+Tr). All models of (PA+Tr) are nonstandard (not isomorphic to **N**). To prove the last claim one constructs a formula that, over **N**, behaves as a 'modest liar formula' – saying 'I am at least a little false.' A detailed analysis shows that this leads to a contradiction.

Let us call the reader's attention to the remarkable book, Grim et al. (1992), where the authors present several self-referential formulas and analyze them by the framework of Łukasiewicz (propositional) logic.

8 Very True

When describing fuzzy logic in the narrow sense, Zadeh claims that it should go beyond the usual many-valued logic, admitting fuzzy truth-values like 'very true,' 'more-or-less true,' etc. Such truth values are understood as fuzzy subsets of the set of truth-values ('true' being just the diagonal; 'very true' being for example the fuzzy set with the characteristic function x^2 on [0, 1]). This was criticized by Haack (1996) as not well founded, unnecessary, etc. Haack herself was criticized by Dubois and Prade (1993), defending Zadeh and fuzzy logic. Here I do not want to enter this discussion but only want to show that 'very true' accommodates well in the 'standard' many-valued approach to fuzzy logic not going beyond it. The idea is to understand 'very true' as a new unary connective.

Recall that in the classical (two-valued) logic we may explicitly have, besides negation (which sends 1 to 0 and 0 to 1) a unary connective t (which sends 1 to 1 and 0 to 0). The formula tφ (evidently equivalent with φ) can then be read 'yes, φ' or 'truly, φ' or just 'φ is true' (not understood as a metatheoretical statement on φ, but just as a part of the object language). In fuzzy logic each mapping of the interval [0, 1] into itself may be taken as the truth function of a unary connective (such connectives are called *hedges*); in particular the identity ($t(x) = x$ for all x) may be taken as the truth function of the fuzzy unary connective t, tφ being read 'yes, φ' or just 'φ is true.' What about a connective vt, where vt(φ) is read 'φ is very true'? What properties shoud it have? Let us call a mapping *vt* of [0, 1] into itself a *truth-stresser* (with respect to a continuous t-norm *) if the following holds for each x, y:

$$vt(1) = 1, \quad vt(x) \leq x, \quad vt(x \Rightarrow y) \leq vt(x) \Rightarrow vt(y).$$

Let BL(vt) be the extension of our logic BL by the following axioms for vt:

(VT1) $\mathrm{vt}(\varphi) \rightarrow \varphi$
(VT2) $\mathrm{vt}(\varphi \rightarrow \psi) \rightarrow (\mathrm{vt}(\varphi) \rightarrow \mathrm{vt}(\psi))$
(VT3) $\mathrm{vt}(\varphi \vee \psi) \rightarrow \mathrm{vt}(\varphi) \vee \mathrm{vt}(\psi)$.

These axioms are *-tautologies iff vt is interpreted by a *-truth stresser. (VT1) says that if φ is very true then φ (is true); (V2) says (modulo a simple transformation) that if both φ and $\varphi \rightarrow \psi$ are very true then ψ is very true. (V3) says that if a disjunction $\varphi \vee \psi$ is very true then one of the disjuncts is very true.

One can show completeness of BL(vt) with respect to a naturally defined class of BL(vt)-algebras. Several interesting examples of truth stressers (for a given t-norm) can be given. For example, one can define vt(φ) to be just φ&φ; or, independently on the t-norm take just $vt(x) = x^2$ (real square – this is the product conjunction but works as a truth stresser also for Ł and G). Proofs are found in Hájek (submitted).

The above is possibly not too surprising but hopefully the reader will agree that saying in fuzzy logic 'φ is very true' we are doing nothing mysterious or deviant. Similarly one could axiomatize other 'fuzzy truth values.'

9 Probability

We stressed that probability on formulas (of classical logic) cannot be understood as an assignment of truth-values in the sense of a (truth-functional) fuzzy logic; but still there are bridges between probability and fuzziness. We describe one of them (see Hájek (1998) originally started by Hájek et al. (1995)). Fuzzy logic speaks in a fuzzy way on some quantities (e.g. 'Temperature is high.') *Probability* is also a quantity and one may say 'The probability of . . . is high' or just '. . . is probable.' The dots stand for any formula of Boolean logic; the word 'probably' acts as a fuzzy modality. Consider a propositional language with two kinds of formulas: *non-modal* – formulas of the classical propositional calculus built from propositional variables and connectives, and *modal* formulas: atomic modal formulas have the form $P\varphi$ where φ is any non-modal formula ($P\varphi$ is read 'φ is probable') and other modal formulas are built from the atomic modal formulas using connectives of Łukasiewicz logic. A *model* of this is a (Kripke) structure $\mathbf{K} = (W, e, \mu)$ where W is a nonempty set of possible worlds, e is a Boolean evaluation assigning to each $w \in W$ and to each propositional variable p the value $e(p, w)$ (zero or one); finally μ is a probability on subsets of W (assume W finite for simplicity). Each non-modal formula has in each possible world either the value 1 or the value 0; the truth value $\|P\varphi\|_{\mathbf{K}}$ of $P\varphi$ in $\mathbf{K}$ is the probability of φ, that is $\mu\{w|\varphi \text{ true in } w\}$). Sentences built from atoms of the form $P\varphi$ are evaluated using truth functions of Łukasiewicz logic. The following formulas are then tautologies:

(FP1) $P(\neg\varphi) \equiv \neg P\varphi$
(FP2) $P(\varphi \rightarrow \psi) \rightarrow (P\varphi \rightarrow P\psi)$,
(FP3) $P(\varphi \vee \psi) \rightarrow ((P\varphi \rightarrow P(\varphi \wedge \psi)) \rightarrow P\psi)$.

EXERCISE. Denote by a, b, c, d the probability of $\varphi\wedge\psi$, $\varphi\wedge\neg\psi$, $\neg\varphi\wedge\psi$, $\neg\varphi\wedge\neg\psi$ respectively; thus for example $a + b$ is the probability of φ. Verify tautologicity of (F1) to (F3). Note that for example (F2) reads: 'If $\varphi \rightarrow \psi$ is probable then if also φ is probable then ψ is probable.'

Postulating axioms of classical logic for non-modal formulas, axioms of Łukasiewicz logic plus our (FP1) to (FP3) for modal formulas and taking as deduction rules *modus ponens* and necessitation (from φ infer $P\varphi$) you get a logic complete with respect to the above semantics.

10 Conclusion

Fuzzy logic in the narrow sense *is* a logic, a logic *with a comparative notion of truth*. It is mathematically deep, inspiring and in quick development; papers on it are appearing in respected logical journals. (Besides the monographs already mentioned, Hájek (1998), Cignoli et al. (2000), let us also mention Turunen (1999), Gottwald (submitted), Novak et al. (2000) and (slightly older) Gottwald (1993).) The bridge between fuzzy logic in the broad sense and pure symbolic logic is being built and the results are promising.

Acknowledgement

The author recognizes partial support of the grant No. IAA1030004 of the Grant agency of the Academy of sciences of the Czech Republic as well as of the CNR Italy during his stay in Italy in June 2000.

References

Cignoli, R., D'Octaviano, R. and Mundici, D. (2000) *Algebraic Foundations of Many-Valued Reasoning*. Dordrecht: Kluwer.

Cignoli, R., Esteva, F., Godo, L. and Torrens, L. (submitted) Basic logic is the logic of continuous t-norms.

Dubois, D. and Prade, H. (1993) *Readings in Fuzzy Sets for Intelligent Systems*. New York: Morgan-Kaufmann.

Esteva, F., Godo, L., Hájek, P. and Navara, M. (2000) Residuated fuzzy logics with an involutive negation. *Archive for Math. Log.*, 39, 103–24.

Gödel, K. (1932) Zum intuitionistichen Aussagenkalkül. *Anzeiger Akad. Wissenschaften Wien, Math.-Naturw. Klasse*, 69, 65–6.

Giles, R. (1976) Łukasiewicz logic and fuzzy set theory. *Int. J. Man–Machine Studies*, 8, 313–27.

Goguen, J. A. (1968–9) The logic of inexact concepts. *Synthese*, 19, 325–73.

Gottwald, S. (1988) *Mehrwertige Logik*. Berlin: Akademie-Verlag.

Gottwald, S. (1993) *Fuzzy Sets and Fuzzy Logic*. Wiesbaden: Viehweg.

Gottwald, S. (to appear) *A Treatise on Many-Valued Logic*.

Grim, P., Mar, G. and St. Denis, P. (1992) *The Philosophical Computer*. Cambridge, MA: MIT Press.

Haack, S. (1974) *Deviant Logic, Fuzzy Logic*. Cambridge University Press; 2nd edn. University of Chicago Press, 1996.

Hájek, P. (1998) *Metamathematics of Fuzzy Logic*. Dordrecht: Kluwer.

Hájek, P. (submitted) Very true.

Hájek, P., Paris, J. and Shepherdson, J. (2000) The liar paradox and fuzzy logic. *Journal of Symbolic Logic*, 65, 339–46.

Hájek, P., Esteva, F. and Godo, L. (1995) Fuzzy logic and probability. In *Proc. UAI'95*, 237–44.

Klir, G. J. and Yuan, B. (eds.) (1996) *Fuzzy Sets, Fuzzy Logic and Fuzzy Systems: Selected Papers by Lotfi A. Zadeh*. Singapore: World Scientific.

Klir, G. J. and Yuan, B. (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall.

Łukasiewicz, J. (1930) Philosophische Bemerkungen zu mehrwertigen Systemen der Aussagenlogik. *C. R. Société des sciences et des letters de Varsovie*, iii, 23, 51–77.

Łukasiewicz, J. and Tarski, A. (1930) Untersuchungen über den Aussagenkalkül. *ibid.* 1–31.

Mamdani, E. H. (1974) Application of fuzzy algorithms for the control of a simple dynamic plant. In *Proc. IEEE*, 121–58.

Novák, V., Perfilieva, I. and Močkoř, J. (2000) *Mathematical Principles of Fuzzy Logic*. Dordrecht: Kluwer.

Pavelka, J. (1979) On fuzzy logic I, II, III. *Zeitschrift für Math. Logik und Grundlagen der Math.*, 25, 45–52, 119–34, 447–64.

Turunen, E. (1999) *Mathematics Behind Fuzzy Logic*. Berlin: Physica Verlag.

Zadeh, L. (1965) Fuzzy sets. *Information and Control*, 8, 338–53.

Zadeh, L. (1994) Preface, in (Marks-II R.J.) *Fuzzy Logic Technology and Applications*. IEEE Technical Activities Board.

Zimmermann, H.-J. (1991) *Fuzzy Set Theory and its Applications*. 2nd edn. Dordrecht: Kluwer.

This page intentionally left blank

Part XII

RELEVANCE AND PARACONSISTENT LOGICS

This page intentionally left blank

38

Relevance Logic

EDWIN D. MARES

1 Non-Sequiturs are Bad

Since 1993, when Andrew Wiles completed his difficult proof of Fermat's Last Theorem, mathematicians have wanted a shorter, easier proof. Suppose when someone addressing a conference of number theorists suggests the following proof of the theorem:

The sky is blue.

$\therefore$ There is no integer n greater than or equal to 3 such that for any non-zero integers x, y, z, $x^n = y^n + z^n$.

This proof would not be well received. But it is valid, in fact sound, on the classical logicians' definition. The premise cannot be true in any possible circumstance in which the conclusion is false. For the conclusion is necessarily true. And the premise is true. Thus the argument is sound and known to be sound.[1]

The classical notion of validity does not agree with our pre-logical intuitions about where the division between good arguments and *non-sequiturs* should be. Classical logic allows connections between premises and conclusions in valid arguments that are extremely loose. There needs to be more of a connection between the content of the premises and conclusion in an argument that we are prepared to call 'valid.'

Some classical logicians have defined content semantically, usually using possible worlds, in such a way as to vindicate arguments like our proof of Fermat's Last Theorem (see e.g. Lewis 1988). On these views, there is a real relationship between the content of the premises and that of the conclusion. I don't want to argue in detail against such attempts here. Such notions of content may be fine for some purposes. But, since they approve of arguments like our proof, they do not coincide with the intuitions that we usually apply when considering whether a proof is good or bad.

Another line of reply is that our notion of good proof is not completely logical, but rather it is partly pragmatic.[2] There is probably some truth to this claim, but we should resist the temptation to push this problem completely into pragmatics. Theories of pragmatics are notoriously vague. They tell us, for example, to reject the above argument because it violates the Gricean maxim to 'be relevant.' What counts as relevant is left

unsaid in Grice's theory. Surely, if there is a theory of relevance that is more rigorous than this, it would be better, all things being equal, to appeal to the more rigorous theory. And relevant logic does provide a very specific view about what counts as a relevant deduction.

The plan of this chapter is to use the natural deduction system for the relevance logic **R** as a guide to the various elements of relevance logic – its proof theory, its semantics, and their interpretation. Later we will introduce weaker relevance logics and two applications of relevance logics: one to the problem of conditionals and the other to the theory of properties.

2 The Real Use of Premises

The problem with *non-sequiturs* like the one given above is that the premises of the inference appear to have nothing to do with the conclusion. Relevance logic attempts to repair this problem, in part, by forcing a constraint on proofs that the *premises really be used in the derivation of the conclusion.*

We will present this idea in the context of Anderson and Belnap's natural deduction system for the logic **R**. The idea is pretty simple. Each premise, or rather hypothesis, in a proof is indexed by a number. The various steps in a proof are indexed by the numbers of the premises which are used to derive the steps. For example, the following is a valid argument in this system:

1.	$A \to B_{\{1\}}$	*hyp.*
2.	$A_{\{2\}}$	*hyp.*
3.	$B_{\{1,2\}}$	$(1)(2) \times (\to E)$

where '$\to E$' is the rule of modus ponens or implication elimination. The numbers in parentheses are the indices.

Throughout this chapter, we will be using the natural deduction system for the logic **R**. This system allows free repetition of premises and the free reiteration of steps between subproofs.

3 Implication

In natural deduction systems we do not usually merely display proofs with premises. We discharge premises to prove theorems of a system. The key rule that we will use here is the rule of conditional proof, or $\to I$ (implication introduction), *viz.*,:[3]

$A_{\{k\}}$ *hyp.*
⋮
B_α
$A \to B_{\alpha - \{k\}}$ $(\to I)$

where k occurs in α. The proviso that k occur in α is essential. It ensures that, in this case, A is really used in the derivation of B.

We can think of the problem of logical relevance in terms of inference, as we have been doing, but also in terms of implication. Relevance logic was developed in part to avoid the so-called paradoxes of material implication. These are formulae that are theorems of classical logic, but are counterintuitive when we think of the arrow as meaning 'implication' in any ordinary sense, or pre-logical philosophical sense, of that term. Among these paradoxes are the following (with names given where they exist):

1. $A \rightarrow (B \rightarrow A)$ (positive paradox)
2. $A \rightarrow (\sim A \rightarrow B)$ (negative paradox)
3. $(A \rightarrow B) \vee (B \rightarrow A)$
4. $(A \rightarrow B) \vee (B \rightarrow C)$
5. $(A \wedge \sim A) \rightarrow B$ (*ex falso quodlebet*)
6. $A \rightarrow (B \rightarrow B)$
7. $A \rightarrow (B \vee \sim B)$

Consider, for example, the positive paradox. It says that any true formula is implied by any formula at all. Implication is usually thought to indicate a tighter relationship than one that can exist between any proposition and a true proposition, merely because the latter is true. Similarly, negative paradox says that any proposition is implied by a false proposition. Again, we have an indication that the material conditional (which always makes a negative paradox true) is too loose a connection to capture the intuitive sense of 'implication.'

Relevant logics were introduced to avoid the paradoxes of implication. Now, this does not mean that the semantics of relevant logics will *show* that all such paradoxes are false in every circumstance. Rather, relevant logicians have developed semantic and proof-theoretic techniques that do not force the paradoxes to be true. Thus, there are at least two notions of relevance at play in relevance logic: (a) The system of proof forces us actually to use every premise in a deduction; (b) the proof theory and semantics do not force us to accept the paradoxes of material implication.

Returning to our natural deduction system, consider the following attempt at a proof of positive paradox:

1.	$A_{\{1\}}$	*hyp.*
2.	$B_{\{2\}}$	*hyp.*
3.	$A_{\{1\}}$	(1) × (*reiteration*)
4.	$B \rightarrow A_{\{1\}}$	(2) – (3) × ($\rightarrow I$)
5.	$A \rightarrow (B \rightarrow A)_{\emptyset}$	(1) – (4) × ($\rightarrow I$)

The illegitimate move here is the use of an implication introduction in the fourth step. 2 does not belong to $\{1\}$ and so we cannot discharge the second hypothesis here. The other paradoxes are avoided in similar ways.

4 From Proof Theory to Semantics

In 1972 Alasdair Urquhart presented a semantic interpretation of relevance numerals. He begins with the notion of a 'piece of information.' A piece of information is a concept which encompasses but is more general than that of a possible world or an evidential situation (the latter is from Kripke's semantics for intuitionist logic).

Pieces of information satisfy statements. To take an example from Urquhart (1992), if we have a piece of information a that consists of the fact that *Harry is taller than Fred* and the fact that *Jim is taller than Harry*, then

$a \models$ *Jim is taller than Fred.*

The satisfaction relation ($\models$) holds between pieces of information and basic statements of a language by virtue of the meanings of those basic statements. Here the meaning of 'is taller than' includes its transitivity.

Among those facts that can be satisfied by pieces of information are what we might call 'informational links.'[4] Among informational links are laws of nature – such as the law that all pieces of matter attract all other pieces of matter – and convention connections – such as the fact that all objects that are exactly a metre in length are the same length as a particular bar in Paris. These informational links provide the truth-makers for implicational statements.

But the truth-making relation between implicational statements and informational links is not very straightforward. For example, if it is a law of nature (or, rather, an instance of a law of nature) at a that $A \to B$ obtains at piece of information a and there is a conventional link in a such that $B \to C$ holds in a, then it would also seem that $a \models A \to C$. But there is no informational link in a which directly makes true $A \to C$. Like 'taller than,' implication seems to be transitive by virtue of its meaning.

To enforce this and other features of implication on the model, Urquhart devised a truth condition for implication using what is now called 'fusion.' Pieces of information can be combined or 'fused' together. The fusion of two pieces of information a and b is written '$a \circ b$'. $a \circ b$ is itself a piece of information.

When we fuse two pieces of information a and b together, we apply the informational links from a to the information in b. Thus, for example, suppose that it is a law in a that all material objects attract all other material objects and among the facts in b are *i is a material object* and *j is a material object*. Thus, in $a \circ b$ we have the fact that *i and j attract one another*.

So now we have a connection between informational links and fusion and a connection between informational links and implication. Putting these together, we can derive the following truth condition for implication:

$a \models A \to B$ if and only if $\forall b(b \models A \Rightarrow a \circ b \models B)$.

An implication is true at a piece of information, if whenever that piece of information is fused with a piece of information which satisfies the antecedent the fusion satisfies the consequent.

There is a tidy connection between fusion and the natural deduction system. Here is an instance of our $\rightarrow E$ rule:

1. $A \rightarrow B_{\{1\}}$ *hyp.*
2. $A_{\{2\}}$ *hyp.*
3. $B_{\{1,2\}}$ $(1)(2) \times (\rightarrow E)$

We can think of the subscripts as names of pieces of information. In line one, we have the statement that $A \rightarrow B$ is true in a piece of information 1. In line two, we have A holding in piece of information 2. And in line three we have B obtaining in $1 \circ 2$. Thus, Urquhart's semantics provides us with a semantic understanding of the indices used in our natural deduction system.

The connection between informational links and implication gives us a means of interpreting relevant implication. Implications are made true by informational links or the results of implicational links under certain closure principles, like transitivity. We take from Devlin (1991) and Israel and Perry (1990) the idea that it is because of informational links and their closures that facts carry the information that other facts obtain.[5] We follow a popular tradition in philosophy of language and hold that statements express their truth conditions. We say that $A \rightarrow B$ means that *A carries the information that B* since this formula expresses the 'fact' that there is an informational connection between A and B or the result of one or more informational connection and certain closure principles.

5 Adding Conjunction

Let's move to discuss another connective. The truth condition for conjunction in this semantics is quite straightforward. That is,

$a \Vdash A \wedge B$ if and only if $a \Vdash A$ and $\alpha \Vdash B$.

This is merely the same truth condition for conjunction that one finds in Kripke's semantics for modal and intuitionist logic.

As we saw in the previous section, the Urquhart's semantics gives us a clear relationship between pieces of information and indices in the natural deduction system. If we apply this relationship to derive rules for conjunction, the truth condition given above yields both introduction and elimination rules for that connective. First the introduction rule:

From A_α and B_α, infer $A \wedge B_\alpha$.

And now the elimination rules:

From $A \wedge B_\alpha$, infer A_α

and

From $A \wedge B_\alpha$, infer B_α.

Note that in the introduction rule the subscript on the two formulae to be conjoined is the same. Before we can conjoin two formulae we have to know that they are true at the same piece of information.

The restriction in this rule that the subscript must remain the same allows us to avoid admitting a well-known proof for positive paradox (see Lemmon 1965). For if we were to allow formulae with different subscripts to be conjoined, we would allow the following proof:

1.	$A_{\{1\}}$	*hyp.*
2.	$B_{\{2\}}$	*hyp.*
3.	$A \wedge B_{\{1,2\}}$	$(1)(2) \times (\wedge I)$
4.	$A_{\{1,2\}}$	$(3) \times (\wedge E)$
5.	$B \rightarrow A_{\{1\}}$	$(2) - (4) \times (\rightarrow I)$
6.	$A \rightarrow (B \rightarrow A)_{\emptyset}$	$(1) - (5) \times (\rightarrow I)$

The illegitimate step here is step 3. It conjoins two formulae that do not have the same index.

6 The Problem of Disjunction

The use of the Urquhart's semantics for relevant logic depends crucially on the notion of a theory. We show that the Urquhart's semantics, given certain additional semantic postulates, is a semantics for a relevant logic by proving a completeness theorem. In a completeness proof for the Urquhart's semantics, we construct a model using theories as pieces of information. A theory is a set of formulae closed under conjunction and provable implication. Let us suppose that Γ is a theory. Then, if $A \in \Gamma$ and $B \in \Gamma$ then $A \wedge B \in \Gamma$. Also, if $A \in \Gamma$ and $A \rightarrow B$ is a theorem of the logic we are using then $B \in \Gamma$. To represent fusion in our model, we take $a \circ b$ to be the set of formulae B such that there is some formula A such that $A \rightarrow B$ is in a and A is in b. In other words, in constructing $a \circ b$ we take major premises from a and minor premises from b and perform *modus ponens* on them. The result is $a \circ b$. This construction crucially depends upon the result of a fusion between two theories itself being a theory. As we shall see, this is a very important fact for the Urquhart's semantics.

Disjunction adds a new and difficult dimension to the semantics. The natural truth condition for disjunction is the following:

$$a \models A \vee B \quad \text{iff} \quad a \models A \text{ or } a \models B.$$

But theories, in general, do *not* meet the corresponding inclusion condition for disjunction, *viz.*,

$$A \vee B \in \Gamma \quad \text{iff} \quad A \in \Gamma \text{ or } B \in \Gamma.$$

Theories that do meet this condition are called *prime* theories. The Urquhart semantics, however, does not work if we restrict ourselves to using prime theories. For the fusion of two prime theories is not always a prime theory.

Thus, either we are forced to modify the standard truth condition for disjunction or abandon the use of fusion in the semantics. Relevant logicians have tried both alternatives, each with success. Kit Fine (1974) has extended the Urquhart semantics to include a treatment of disjunction by altering its truth condition. Richard Routley and Robert Meyer, on the other hand, have retained the standard truth condition for disjunction and relinquished the use of fusion in their series of papers (1973, 1972a, 1972b); for a very nice presentation of this semantics see Routley *et al.* (1982). Here we will look at the Routley–Meyer semantics, also known as the *relational semantics*.

7 Routley and Meyer's Ternary Relation

To distinguish between the elements of Urquhart's semantics and those of Routley and Meyer's relational semantics, let us call the latter *situations*, although in the literature they are also called 'worlds' and 'setups.' The Routley–Meyer semantics replaces fusion with a three-place accessibility relation on situations, R. The truth condition for implication is correspondingly changed:

$$a \Vdash A \to B \quad \text{iff} \quad \forall x \forall y((Raxy \,\&\, x \Vdash A) \Rightarrow y \Vdash B)$$

This truth condition might look like a 'bolt from the blue,' but it is actually a generalization of the truth condition for necessity from Kripke's semantics for modal logic. Whereas Kripke uses a binary relation to interpret a monadic connective (necessity), Routley and Meyer use a ternary relation to interpret a binary connective (implication).

We can view the relational semantics as generalizing fusion. Recall that we said that $a \circ b$ results when the informational links from a are applied to the information in b. Adapting this idea to the ternary relation is quite easy. We say that $Rabc$ obtains when the information that *results* from the application of the links in a to the information in b is *contained in* c. To return to our previous example, suppose that a contains the law that *all matter attracts all other matter* and b contains the information that *i and j are material objects*. Then c contains the information that *i and j attract one another*.

We can force implication to have the various properties that we want by accepting certain postulates in our model theory. For example, if we want the statements satisfied by situations to be closed under *modus ponens*, we adopt the following postulate. For all situations a,

$Raaa$.

For suppose that $a \Vdash A \to B$ and $a \Vdash A$. By the truth condition for implication, for all b such that $Raab$, $b \Vdash B$. But $Raaa$, by the above postulate. So, $a \Vdash B$. Thus, a is closed under *modus ponens* as we suggested. There are similar postulates that are required to satisfy transitivity and other properties that one might desire implication to have.

8 Rules for Disjunction

Now we return to our natural deduction system and add some rules for disjunction. The introduction rules are quite obvious:

From A_α, to infer $A \vee B_\alpha$.

and

From B_α, to infer $A \vee B_\alpha$.

The elimination rule is a little more complicated:[6]

From $A \vee B_\alpha$, $A \to C_\beta$, and $B \to C_\beta$,
infer $C_{\alpha \cup \beta}$.

There is also a rule that ensures the distribution of conjunction over distribution (it simply states that one can infer from $A \wedge (B \vee C)_\alpha$ to $(A \wedge B) \vee (A \wedge C)_\alpha$).[7]

9 The Semantics of Negation

The treatment of negation is one of the most controversial elements of relevant logic (see Copeland 1979). The key here is that in order to block *ex falso quodlebet* we need situations in our semantics that make contradictions true. In addition, in order to reject the paradox $A \to (B \vee \sim B)$ we need situations at which bivalence fails. Thus, we need a semantics for negation that does not force bivalence or consistency on us.

Routley and Meyer's model theory incorporates a device from a semantics developed originally for relevant logic by Richard and Val Routley in 1972. This device is an operator on situations, which has become known as the *Routley star*. For each situation in the semantics, there is a situation that is its 'star.' The star of a situation a is a^*. Some situations (in some models) are identical with their stars, but not all. The truth condition for negation then becomes:

$a \Vdash_v \sim A$ iff $a^* \nVdash A$.

The relative independence of a situation and its star allows inconsistencies and failures of bivalence.

On the other hand, Routley and Meyer relate worlds to their pairs in order to satisfy the various postulates governing negation. For example, they set $a = a^*$ in order to satisfy $A \to \sim\sim A$ and $\sim\sim A \to A$.

The problem with the Routley star is that many philosophers have had trouble understanding what it is and what it is supposed to do with negation. Here we will use

an explanation due to J. M. Dunn (1993). Dunn does not begin with the Routley star. Instead, he postulates a binary relation C on situations that is supposed to relate situations to situations with which they are 'compatible.' For example, suppose there is information in the present situation that a particular table is completely red. In another situation, there is the information that it is completely green. These two situations are incompatible with each other. If there are no such conflicts, then situations are compatible.

Note that situations need not be compatible with themselves. Situations are abstract ('ersatz') entities and can contain conflicting information (e.g. that the table is red and that it is green all over).

Now we have the following truth clause for negation:

$$a \models_\omega {\sim}A \quad \text{iff} \quad \forall b(Cab \supset b \not\models A)$$

Now we can use the compatibility relation to define the star operator. a^* is the largest situation that is compatible with a. a^* is largest in the sense that it contains more information than any other situation compatable with a.[8]

10 Rules for Negation

The introduction rule for negation is a form of the *reductio* rule:

From $A \rightarrow {\sim}A_\alpha$, infer ${\sim}A_\alpha$. $(\sim I)$

The elimination rule is a form of *modus tollens*:

From ${\sim}B_\alpha$ and $A \rightarrow B_\beta$ infer ${\sim}A_{\alpha\circ\beta}$. $(\sim E)$

We also add two double negation rules: From ${\sim}{\sim}A_\alpha$ to infer A_α and the converse of this rule.

11 Disjunctive Syllogism

The rule of disjunctive syllogism (DS) is

From $A \vee B_\alpha$ and ${\sim}A_\alpha$, infer B_α.

This is an intuitive rule of inference. We use it to 'deduce' the identity of the murderer when reading mystery novels – we eliminate all but the guilty party. We use it when determine who sits on university committees ('X are going on leave. Y will say crazy things. So, it has to be Z.'). In fact, we use DS all the time. But, if we add DS we get back one of the paradoxes of implication. The following proof is due to C. I. Lewis (see Lewis and Langford 1959):

1.	$A \wedge {\sim}A_{\{1\}}$	*hyp.*
2.	$A_{\{1\}}$	$(1) \times (\wedge E)$
3.	$A \vee B_{\{1\}}$	$(2) \times (\vee I)$
4.	${\sim}A_{\{1\}}$	$(1) \times (\wedge E)$
5.	$B_{\{1\}}$	$(3)(4) \times (DS)$
6.	$(A \wedge {\sim}A) \rightarrow B_{\emptyset}$	$(1) - (5) \times (- I)$

So, adding DS gets us EFQ.

Reactions to this proof are varied. Some, like Stephen Read (1988), claim that the problem is not with disjunctive syllogism, but rather with our understanding of disjunction in natural language. We should interpret natural language disjunctions as an intensional disjunction. In particular, we should treat it as *fission*, sybolised $\oplus$. In **R** and closely related relevant logics, fission can be defined as

$$A \oplus B =_{df} {\sim}A \rightarrow B$$

Clearly, disjunctive syllogism is valid for fission. It is just a form of modus ponens ($\rightarrow E$). What is not valid for fission is addition. We cannot infer from A to $A \oplus B$. This blocks the step from lines 2 to 3 in the proof above. Thus the proof is blocked and so we can create a form of disjunctive syllogism without buying into a paradox of implication.

Read's solution has definite merits. I have found that students, who have not yet become accustomed to the quirks of classical logic, find $\vee I$ a rather strange rule. When I have shown Lewis's proof during seminars, I have the audience vote to decide which rule to reject. $\vee I$ is always the one they choose.

But there are problems with taking natural language disjunction to be intensional. In particular, the truth condition for intensional disjunction does not look like anything we would identify with natural language disjunction. That is,

$$a \Vdash A \oplus B \quad \text{if and only if} \quad \forall b \forall c (Rbca \supset (b \Vdash A \text{ or } c \Vdash B)).$$

Many relevant logicians have been reluctant to accept Read's view because they do not think that this truth condition looks like it explicates the meaning of natural language disjunction.

There are other ways of dealing with the apparent validity of DS. We'll take a look at one due to Chris Mortensen (1986), with some minor changes made to fit the current chapter.

Mortensen holds that we can use DS with extensional conjunction under certain circumstances. According to Mortensen, the problem with DS is that we cannot use it when reasoning about inconsistent situations. When we reason about consistent situations, we can use it. For, the following argument is valid (Mortensen 1986: 196):

1. If a situation a is consistent, $a \Vdash A \vee B$ and $a \Vdash {\sim} A$, then $a \Vdash B$. (hypothesis)
2. a is consistent hypothesis
3. $a \Vdash A \vee B$ hypothesis

4. $a \models \sim A$ hypothesis
5. $a \models B$ (1)(2)(3)(4) × (*modus ponens*)

There are at least two interesting features of this argument. First, it is done within the semantic metalanguage. Second, the nature of the first premise is worth nothing. It tells us that consistent situations are closed under disjunctive syllogism. Note that this is a premise – it is not proven by the argument. Nor does Mortensen establish it in any other way. Mortensen says that it is justified by intuition (Mortensen 1986).

I do not have space here to argue the merits of either of these treatments of DS, nor do I have space to discuss other alternative approaches. I direct the interested reader to the bibliographies in Anderson *et al.* (1992) and Read (1988) for readings on this topic.

12 Logics Stronger than **R**

So far, we have been motivating the logic **R**. But there are many other relevant logics that have been studied.

Logics stronger than (that contain more theorems than) but close to **R**, tend to be only marginally relevant. For example, the logic **R** Mingle (**RM**) of Dunn and Storrs McCall contains all the axioms and rules of **R**, plus the mingle axiom:

$$A \rightarrow (A \rightarrow A)$$

For a logic to be counted as relevant it must have the variable sharing property. That is, if a formula $A \rightarrow B$ is a theorem, then formulae A and B must contain at least one propositional variable in common. **RM** does not have the variable sharing property, but does have a property very close to it, *viz.*,[9]

If $A \rightarrow B$ is a theorem of RM, then either
A and B share a propositional variable
or $\sim A$ and B are both theorems of RM.

RM contains theorems that seem paradoxical from a relevant point of view. Among these is '$(A \rightarrow B) \vee (B \rightarrow A)$'. Despite the semi-relevance of **RM**, it has its attractions and it and systems very similar to it have their advocates (see, e.g. Avron 1990a, 1990b).

Slightly stronger than **RM** is the elegant logic **RM3**. **RM3** is characterized by three-valued truth tables. The three values that we will use are T, F, and B. T and F are true and false respectively, and B is the value 'both true and false.' Dialetheists are philosophers that hold that some sentences can be both true and false. Even if we are not dialetheists, we can make sense of this truth-value by thinking about inconsistent fictional stories. We can understand such stories by taking the inconsistencies in them to be both true and false of the story. Both T and B are designated values in this semantics. That is, a statement is considered to be (at least) true if it is T (perhaps best thought of as 'merely true') or B.

The truth tables for **RM3** are the following:

$\wedge$	T	B	F
T	T	B	F
B	B	B	F
F	F	F	F

$\vee$	T	B	F
T	T	T	T
B	T	B	B
F	T	B	F

$\rightarrow$	T	B	F
T	T	B	F
B	T	T	B
F	T	T	T

$\sim$	
T	F
B	B
F	T

This is a simple, elegant logic that comes very close to being relevant.

13 Logics Weaker than **R**

On the other hand, many relevant logicians have argued for systems weaker than **R**. In fact, the preferred logic of the *Entailment* volumes (Anderson and Belnap (1975) and Anderson *et al.* (1992)) is the logic **E** of relevant entailment. **E** is meant to capture a *strict* relevant implication. Given this, it was conjectured that **E** would be captured by **R** extended by the addition of a necessity operator and some (S4-ish) axioms governing that operator. Unfortunately, **NR** or $\mathbf{R}^{\Box}$, as the resulting logic was called, was shown to be somewhat stronger than **E**. As a result, relevant logicians could not accept both that **R** is the logic of relevant implication and that **E** is the logic of necessary relevant implication. (See Anderson and Belnap (1975) for a more detailed history.)

Other relevant logicians have given various reasons for adopting weaker logics, but we will only look at one such motivation.

Some logicians have used relevant logics to develop naïve theories of truth and naïve set theories. A naïve theory of truth both contains its own truth predicate and admits Tarski's truth schema, *viz.*,

$$True(\ulcorner A \urcorner) \leftrightarrow A.$$

As we shall soon see, if we base a naïve theory of truth on the logic **R**, we end up with a trivial system. That is, the resulting logic can prove every proposition.

A similar state of affairs holds for naïve theory of sets. A naïve set theory contains an unrestricted comprehension principle, such as, for each formula A,

$$\exists x \forall y (A(y) \leftrightarrow y \in x).$$

Again, adding this principle (along with other standard principles of set theory) to **R** yields a trivial theory.

Let's use contraction to prove 'Curry's paradox' in **R** with the naïve theory of truth. My proof follows Meyer *et al.* (1979), and is done in the axiom system.

We start with a definition of a proposition C:

$C =_{df} Tr(\ulcorner C \urcorner) \rightarrow p$

where p is any arbitrary proposition. We also define a biconditional ($A \leftrightarrow B =_{df} (A \rightarrow B) \wedge (B \rightarrow A)$).

1.	$C \leftrightarrow (Tr(\ulcorner C \urcorner) \rightarrow p)$	definition of C
2.	$C \rightarrow (Tr(\ulcorner C \urcorner) \rightarrow p)$	(1) $\times$ (simplification)
3.	$Tr(\ulcorner C \urcorner) \leftrightarrow C$	(T-schema)
4.	$C \rightarrow (C \rightarrow p)$	(2)(3) $\times$ (replacement of equivalents)
5.	$\vdash_{\mathbf{R}} (C \rightarrow (C \rightarrow p)) \rightarrow (C \rightarrow p)$	(contraction)
6.	$C \rightarrow p$	(4)(5) $\times$ (*modus ponens*)
7.	$(Tr(\ulcorner C \urcorner) \rightarrow p) \rightarrow C$	(1) $\times$ (simplification)
8.	$(C \rightarrow p) \rightarrow C$	(7)(3) $\times$ (replacement of equivalents)
9.	C	(6)(8) $\times$ (*modus ponens*)
10.	p	(6)(9) $\times$ (*modus ponens*)

Thus, we can prove any arbitrary proposition in **R** with the naïve theory of truth. A very similar proof can be used to show that every proposition is provable in **R** with a naïve set theory.

Some relevant logicians hold that *one* problem with **R** is that it contains the principle of *contraction*, used in step three in the argument. In schematic form, contraction is

$(A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B)$.

Here is a proof of contraction in our natural deduction system:

1.	$A \rightarrow (A \rightarrow B)_{\{1\}}$	*hyp.*
2.	$A_{\{2\}}$	*hyp.*
3.	$A \rightarrow B_{\{1,2\}}$	(1)(2) $\times$ ($\rightarrow E$)
4.	$B_{\{1,2\}}$	(2)(3) $\times$ ($\rightarrow E$)
5.	$A \rightarrow B_{\{1\}}$	(2) – (4) $\times$ ($\rightarrow I$)
6.	$(A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B)_{\emptyset}$	(1) – (5) $\times$ ($\rightarrow I$)

How should we block this proof? Note that hypothesis 2 is used in *modus ponens* at both lines 3 and 4. One step towards rejecting contraction in a natural deduction system is to restrict the use of premises in a proof. That is, we allow each hypothesis to be discharged only once. This blocks the proof. (Of course, in setting up a natural deduction system that does not prove contraction, we should be sure that there is no other way to prove that thesis.)

There has been some success in developing naïve theories using weak relevant logics. Ross Brady has shown that a weak relevant logic can support a consistent naïve class theory and a consistent naïve set theory (see Brady 1983, 1989, and forthcoming).

14 Relevant Logics and Natural Language Conditionals

In this and the next section, we will look at two applications of relevant logic and its semantics.

As Dunn (1986) argues, we need a relevant theory of conditionals. For example, consider the following example:

> If you pick up a pregnant guinea pig by the tail,
> all her babies will be born without tails.[10]

Intuitively, this statement is false. Taking the conditional to be a material implication, however, makes it true.

Probabilistic treatments of the conditional, quite popular at the moment, do not capture relevance phenomena adequately either. These treatments, for the most part, accept *Adam's thesis*. This thesis says that a conditional $A \Rightarrow B$ is assertable if and only if the conditional probability $\Pr(B/A)$ is high. But there are cases in which the consequent of a conditional has a high probability independently of the probability of the antecedent. For example,

> If John dropped this piece of chalk, Einstein's theory of gravity holds.

There is clearly something wrong with this conditional. It makes it seem as if the truth of Einstein's theory of gravity was caused by John's dropping this piece of chalk. Conditional probability $\Pr(E/C)$ is high because the probability of E is high and is independent of the probability of C. On the probabilistic theory of conditionals, this makes the conditional assertable.

These brief, and rather dogmatic, remarks motivate a relevant treatment of the condition. But we cannot merely take the conditional to be a form of relevant implication. For the implications of the various logics we have seen have properties that the natural language conditional does not have. For instance, the principle of strengthening the antecedent is valid for all these conditionals, that is, we can infer from the truth of $(A \rightarrow C)$ to $((A \wedge B) \rightarrow C)$. But we cannot infer from 'If Ramsey got a new chew toy this afternoon, he is now happy' to 'If Ramsey got a new chew toy this afternoon and he had a bath, he is now happy'.[11]

Instead, we can formulate another connective, which shares some properties with relevant implication. The view that I present here is a simplification of that of Mares (forthcoming), which in turn is a development of the theory of Dov Gabbay (1972, 1976). Here I develop this idea semantically.

In addition to the elements from the Routley–Meyer semantics, we add a four-place accessibility relation, I. This relation holds between two propositions and three situations. Here a proposition will merely be a set of situations. Thus, $|A|$ is the set of situations at which the formula A is true. The conditional is represented by the symbol $\Rightarrow$.

Our truth condition for the conditional mirrors the Routley–Meyer truth condition for implication.

$$a \Vdash A \Rightarrow B \quad \text{iff} \quad \forall x \forall y((I|A||B|axy \;\&\; x \Vdash A) \supset y \Vdash B)$$

The difference here, of course, is the insertion of the antecedent and consequent propositions into the accessibility relation. Why we need the consequent proposition represented here will be treated later. The inclusion of the antecedent proposition allows us to block unwanted inferences such as strengthening of the antecedent. For we cannot infer on this semantics from $I|A \wedge B||C|abc$ to $I|A||C|abc$. This blocks the inference from $a \vDash A \Rightarrow C$ to $a \vDash (A \wedge B) \Rightarrow C$.

There are some inferences, however, that we do want to hold of the conditional. One way of understanding implication is as an idealization of the conditional. An analogy might help here. Recall that we interpreted the implication of **R** in terms of *universal* laws of nature, *sufficient* causal statements, and so on. Yet our standard laws of science and causal statements are not universal or sufficient. They have *ceteris paribus* clauses built into them. I suggest that the relationship between implication and the conditional is akin to that between universal laws of nature and sufficient causal statements, on the one hand, and *ceteris paribus* laws and normal causal statements on the other. The latter have built into their interpretation unstated conditions that indicate where they do and do not apply. Similarly, standard conditionals have unstated restrictions of these sorts built into their interpretation. Thus, in the semantics of conditionals, the conditional $I|A||B|abc$ only holds when b and c satisfy the restrictions associated with the proposition $|A|$ and $|B|$ at the situation a.

The inclusion of the consequent proposition is supposed to prevent certain irrelevances for appearing. For example, when I talk about what happens if a piece of chalk is dropped, one of the background assumptions that I make is that the laws of gravity will hold when I drop the chalk. But as we have said, we don't want to accept the conditional, 'If John dropped this piece of chalk, the theory of gravity holds.'[12] Rather, on the present view the consequent cannot count as a background assumption. Thus the consequent helps to determine which situations are used in the evaluation of the conditional.

We can represent the idea that the conditional is a relevant implication plus some restrictions by the following principle:

$$\frac{I|A||B|abc}{\therefore Rabc}.$$

This principle makes valid the thesis below:

$$(A \rightarrow B) \rightarrow (A \Rightarrow B)$$

We can also force the conditional to satisfy *modus ponens* by adding the principle that $I|A||B|aaa$, for all propositions $|A|$ and $|B|$ and all situations a. Given the previous principle, it makes sense to have the conditional satisfying modus ponens if the corresponding implication also satisfies *modus ponens*.

15 Theory of Properties

The doctrine of relevant predication is due to Dunn (1987, 1990a, 1990b). Philosophers distinguish between those properties that an object really has from those

that it has in a rather tenuous way. Consider the problem of Cambridge change. Two hours ago, in Toronto (on the other side of the world from me), it was raining. Now it has stopped. So we have a change from

Ed is such that it is raining in Toronto

to

Ed is such that it is not raining in Toronto.

Here there is a change, but not a real change in me. The change doesn't really affect me.

The standard treatment of lambda abstraction in logic does not distinguish between those properties that an object really has from those that it has only in this incidental way. For '$\lambda x(Raining(toronto))ed$' is usually treated as a legitimate predication.

Dunn (1987, 1990a, 1990b) distinguish between ordinary predication and relevant predication. A property φ is had by an entity i relevantly if a thing's being i implies that it has φ. In Dunn's formalism,

$$(\rho x \varphi x)i =_{df} \forall x(x = i \rightarrow \varphi x).$$

The idea is that the property here is had by the thing by virtue of its being that thing. Relevant implication is used to formalise the 'in virtue of' here.

As Dunn puts it in (1990b), if i has a φ relevantly, then for anything x,

$$\sim\varphi x \rightarrow x \neq i.$$

So not to have φ is to make a thing, or in our own terms, to carry the information that it is, not i.

To take an example, consider

Ronald Reagan is such that Socrates is wise.

We can ask what this 'is such that' is doing here. If it is there to indicate that 'Socrates is wise' is relevantly predicated of Reagan it would seem to be false, for

$$(\rho x S)ron \leftrightarrow \forall x(x = ron \rightarrow S).$$

There is no reason to believe that $x = ron \rightarrow S$ for any x here, since just because S is true does not mean that arbitrary statements imply it. And the sentence '$x = ron$' seems to have nothing to do with S. Thus, we can reject '$(\rho x S)ron$.'

16 Summary

This has been a rather opinionated introduction to relevant logic. I have used the natural deduction system for **R** as a guide, since it is rather elegant and because *I* like

it. And I have used a particular reading of the semantics in order to give the philosophically inclined reader a way of understanding the system. There are many other ways of understanding relevant logic, but in a short chapter one cannot cover them all. So I have decided to treat only one interpretation at some length.

Acknowledgements

An early version of this chapter was presented to a group of staff and students at Monash University. I am grateful for their comments.

Notes

1 It won't help to argue that in a proof, all the steps in the proof must be transparent to the people to whom the proof is presented, for even after it is pointed out to us that from it is a theorem that A we can infer that B, therefore A, we still feel quite cheated by the so-called proof.

2 For a very clear version of this approach, see, Robert Fogelin (1978).

3 Anderson and Belnap (1975) present their rules in 'horizontal' form: From $A_1, \ldots, A_n$ infer B. I will use both their method of presentation and the current 'vertical' method, depending on which is easier in a given context.

4 Here I am not presenting Urquhart's theory. Rather, the present theory, in effect, is that of Mares (1996). This view is an elaboration of Meyer's interpretation of Urquhart's semantics. Meyer (in conversation) takes fusion to be the application of the 'laws' of one piece of information to the facts in another piece of information.

5 Note that for Devlin and Israel and Perry (as for Mares 1996) the list of sorts of informational links is much longer than the one that I gave earlier. A reasonably good list is at Devlin (1991: 12). I don't agree with all the types of links that Devlin includes, especially empirical generalizations, but the reader can get the general idea of what I am talking about from that list.

6 I use what Anderson and Belnap (1975) call $\vee E^s$, because it is simpler than the usual rule.

7 Ross Brady has developed a natural deduction system that is supposed to eliminate the need for this additional rule for distribution. It has not yet been published.

8 In the full Routley–Meyer semantics there is a binary relation $\leq$ on situations. If Cab, then either $b < a^*$ or $b = a$.

9 This is stronger than the closest property had by classical logic. For classical logic, if $A \rightarrow B$ is a theorem, then either A and B share a variable or $\sim A$ is a theorem *or* B is a theorem. This property is called 'Halldén reasonableness.'

10 This example is courtesy of A. R. Anderson, but comes originally from a children's story. If any reader knows the title and author of this story, I would appreciate the information.

11 Ramsey is a dog.

12 If we add 'still' in the consequent of the above conditional, it becomes acceptable; 'still' and 'even' would seem to be 'relevance breakers' and a non-relevant analysis of them is appropriate.

References

Anderson, Alan R. and Nuel D. Belnap, Jr. (1975) *Entailment: Logic of Relevance and Necessity*, vol. I. Princeton, NJ: Princeton University Press.

Anderson, Alan R., Nuel D. Belnap, Jr. and Dunn, J. M. (1992) *Entailment: Logic of Relevance and Necessity*, vol. II. Princeton, NJ: Princeton University Press.

Avron, Arnon (1990a) Relevance and paraconsistency – a new approach. *Journal of Symbolic Logic*, 55, 707–32.

Avron, Arnon (1990b) Relevance and paraconsistency – a new approach II: the formal systems. *Notre Dame Journal of Formal Logic*, 31, 168–202.

Brady, Ross T. (1983) The simple consistency of a set theory based on the logic CSQ. *Notre Dame Journal of Formal Logic*, 24, 431–49.

Brady, Ross T. (1989) The non-triviality of dialectical set theory. In G. Priest, R. Routley and J. Norman (eds.), *Paraconsistent Logic* (pp. 437–71). Munich: Philosophia Verlag.

Brady, Ross T. (forthcoming) *Universal Logic*. Stanford, CA: CSLI.

Copeland, B. J. (1979) On when a semantics is not a semantics: some reasons for disliking the Routley–Meyer semantics for relevance logic. *Journal of Philosophical Logic*, 8, 399–413.

Devlin, Keith (1991) *Logic and Information*. Cambridge: Cambridge University Press.

Dunn, J. M. (1986) Relevance logic and entailment. In D. Gabbay and F. Geunthner (eds.), *Handbook of Philosophical Logic III* (pp. 117–224). Dordrecht: Reidel.

Dunn, J. M. (1987) Relevant predication I: the formal theory. *Journal of Philosophical Logic*, 16, 347–81.

Dunn, J. M. (1990a) Relevant predication II: intrinsic properties and internal relations. *Philosophical Studies*, 60, 177–206.

Dunn, J. M. (1990b) Relevant predication III: essential properties. In J. M. Dunn and A. Gupta (eds.), *Truth or Consequences: Essays in Honour of Nuel Belnap* (pp. 77–95). Dordrecht: Kluwer.

Dunn, J. M. (1993) Star and perp. *Philosophical Perspectives*, 7, 331–57.

Fine, Kit (1974) Models for entailment. *Journal of Philosophical Logic*, 3, 347–72.

Fogelin, Robert (1978) *Understanding Arguments: An Introduction to Informal Logic*. New York: Harcourt, Brace, Jovanovich.

Gabbay, Dov (1972) A general theory of the conditional in terms of a ternary operator. *Theoria*, 38, 97–104.

Gabbay, Dov (1976) *Investigations in Modal and Tense Logics with Applications to Problems in Philosophy and Linguistics*. Dordrecht: Reidel.

Israel, David and John Perry (1990) What is information? In P. P. Hanson (ed.), *Information, Language, and Cognition* (pp. 1–19). Vancouver: University of British Columbia Press.

Lemmon, E. J. (1965) *Beginning Logic*. London: Nelson.

Lewis, C. I. and Langford, C. H. (1959) *Symbolic Logic*, 2nd end. New York: Dover.

Lewis, David K. (1988) Relevant implication. *Theoria*, 54, 161–74. Reprinted in D. K. Lewis, *Papers in Philosophical Logic*. Cambridge: Cambridge University Press, 1998, 111–24.

Mares, Edwin D. (1996) Relevant logic and the theory of information. *Synthese*, 109, 345–60.

Mares, Edwin D. (forthcoming) Relevant implication and the indicative conditional. *Synthese*.

Mares, Edwin D. and Robert K. Meyer (2001) Relevant logics – an overview, in L. Goble (ed.), *A Guide to Philosophical Logic*. Oxford: Blackwell.

Meyer, Robert K., Richard Routley and Dunn, J. M. (1979) Curry's Paradox. *Analysis*, n.s. 39, 124–8.

Mortensen, Chris (1986) Reply to Burgess and to Read. *Notre Dame Journal of Formal Logic*, 27, 195–200.

Restall, Greg (2000) *An Introduction to Substructural Logics*. London: Routledge.
Read, Stephen (1988) *Relevant Logic*. Oxford: Blackwell.
Routley, Richard and Robert K. Meyer (1973) Semantics of entailment. In H. Leblanc (ed.), *Truth, Syntax, and Modality* (pp. 199–243). Amsterdam: North Holland.
Routley, Richard and Robert K. Meyer (1972a) Semantics of entailment II. *Journal of Philosophical Logic*, 1, 53–73.
Routley, Richard and Robert K. Meyer (1972b) Semantics of Entailment III. *Journal of Philosophical Logic*, 1, 192–208.
Routley, Richard, Robert K. Meyer, Ross T. Brady and Val Plumwood (1982) *Relevant Logics and its Rivals*, vol. 1. Atascardero: Ridgeview.
Routley, Richard and Val Routley (1972) Semantics for first degree entailment. *Noûs*, 6, 335–59.
Urquhart, Alasdair (1972) Semantics for relevant logics. *Journal of Symbolic Logic*, 37, 159–69. Reprinted in Anderson, Belnap, and Dunn (1992).

Further Reading

A longer, more technical but very readable introduction is J. M. Dunn (1986). Dunn and Greg Restall have rewritten this article for a new addition of the *Handbook of Philosophical Logic* which is forthcoming. Greg Restall has also written a textbook on substructural logics (2000), among which are relevant logics. Mares and Meyer (forthcoming) is a longer introduction than the current piece and somewhat more detailed (although it does not deal with all the topics treated here).

For the more advanced reader, Anderson and Belnap (1975) and Anderson *et al.* (1992), as well as Routley *et al.* (1982) contain most of the formal material on these logics. Of course, there have been technical developments since these books were published.

On the more philosophical side, Read (1988) is an interesting attempt to interpret and defend relevant logic. His work is idiosyncratic, but this is necessarily the case in relevant logic. Relevant logic, unlike intuitionist logic for example, was not developed with the aim of articulating a philosophical position. And relevant logic fits with many philosophical perspectives. It is my experience that there are as many interpretations of relevant logic as there are relevant logicians. For a very different philosophical outlook, see Routley *et al.* (1982).

39

On Paraconsistency

BRYSON BROWN

1 What is Paraconsistency?

The term paraconsistent logic is due to F. Miro Quesada. 'Para' can mean any of 'against,' 'near,' or 'beyond.' It's not clear which of these Quesada had specifically in mind, or whether he deliberately embraced the ambiguities inherent in the term. But each of these meanings suits the programs of at least some who have worked on paraconsistent logic. The most radical paraconsistentists, the dialetheists, are indeed against consistency, at least as a global constraint on our metaphysics. They hold that the world is inconsistent, and aim at a general logic that goes beyond all the consistency constraints of classical logic. More modest practitioners of paraconsistent logic aim to give us a logic suitable for treating nearly (but not quite) consistent sets. But the central logical problem is something both the moderates and the radicals share: the classical trivialization of inconsistent and unsatisfiable sets.

A set Γ is inconsistent iff its closure under deduction includes both α and $\neg\alpha$ for some sentence α; it is unsatisfiable if there is no admissible valuation that satisfies all members of Γ. Classical logic trivializes all such premise sets. That is, from any inconsistent premise set, we can derive any sentence in the language, and from any unsatisfiable premise set, every sentence in the language follows. Each term when I teach introductory logic I have to explain this intuitively odd fact, which we can express roughly as:[1]

Triv: $\{A,\neg A\} \vdash_c B$; $\{A,\neg A\} \models_c B$

In that logic class I sometimes try to motivate this oddity by locating it in the context of persuading others to accept various sentences. Of course we know the premise set $\{A,\neg A\}$ cannot be true. So, if someone grants you (or anyone) those premises, they should be prepared to grant you anything at all (how could they object to B, having already accepted A and $\neg$A?). But this defense is just a rhetorical dodge. It assumes something that's clearly false, that is that we can never have good reasons to accept both A and $\neg$A, while not having any good reason to accept B.

More famously, C. I. Lewis argued for Triv by presenting a proof of it, along these lines:

1.	A	premise
2.	¬A	premise
3.	(A ∨ B)	1, ∨-intro
4.	B	2,3, disjunctive syllogism

The proof produces a challenge for those who reject Triv: Which of these two rules, ∨-intro or disjunctive syllogism, will they give up? Both seem pretty obviously sound. How can we deny that A ∨ B follows from A, given the standard understanding of '∨' as inclusive disjunction? Similarly, how can we deny that B follows from A ∨ B and ¬A, given that '¬' is negation and '∨' is inclusive disjunction? The usual answer (first proposed by Anderson and Belnap[2]) is that disjunctive syllogism must fail in such cases. After all, we obtained A ∨ B by inferring it from our premise A. How can we then justify turning around and inferring B with the help of our other premise, ¬A? This inference depends on the assumption that the assertion of ¬A can be used to rule out that it's due to having asserted A that we are entitled to assert that A ∨ B. But to assume this is to deny that we ever reason with sets like {A,¬A}, not to infer something we would want to infer from such sets.

There are other reasons to adopt paraconsistent logic, reasons that are telling even for those who accept the argument for trivialization above. For example, a deontic logic that does not trivialize conflicting obligations, or an epistemic logic that does not trivialize inconsistent beliefs, will require a paraconsistent consequence relation. And many philosophers who accept Lewis' trivialization argument are still prepared to accept non-trivial but inconsistent obligations and/or beliefs. Inconsistent theories, such as Bohr's theory of the hydrogen atom, and the early calculus, provide another application that calls for consequence relations that don't trivialize inconsistent sets of sentences. We'll examine some logics specifically designed for these applications later; for now, we will focus on Triv and the view of consequence relations that it derives from. Our first aim will be a general taxonomy of paraconsistent logic based on alternative approaches to avoiding Triv.

2 Motives for Paraconsistency

Many philosophers follow Lewis, and respond to Triv by accepting and defending it. One of the principal arguments in favour of Triv is to say that Triv is just classical logic's way of telling you to get yourself a new set of premises. If your premises are unsatisfiable or inconsistent, then they cannot all be true, and you cannot accept them all without at least implicitly contradiction yourself. But if you know that you are reasoning from a starting point that already contains an error, you should clear that error up, not compound it by drawing further conclusions from your already erroneous premises.

However, as many have pointed out, eliminating inconsistency from our premises is easier said than done. There is a clear technical sense in which we can say that eliminating inconsistency is difficult. Of course, before we can even try to eliminate an inconsistency from our premises, we must find the inconsistency. But the consistency of a set of sentences is not decidable, that is there is no procedure which will show

whether or not an arbitrary set of sentences in a first-order language is consistent.[3] So inconsistencies may lie hidden in our set of accepted claims.

More importantly, there may be strong practical reasons for not eliminating an inconsistency, even when we can see that one is present. For example, consider Bohr's model of the hydrogen atom. While Bohr used classical electrodynamics to model the radiation the atom absorbs and emits, he also allowed the electron to orbit the nucleus in a 'stationary state' without emitting any radiation and without the radiation it emits in shifting between stationary states having a frequency related to the frequency of the electron's periodic motion around the nucleus. But classical electrodynamics requires that all accelerated charges radiate, and that the radiation emitted by an accelerated charge have a frequency related to the frequency of the accelerated motion. Bohr's model was clearly inconsistent – but these inconsistent features were indispensable to capturing the phenomena Bohr was trying to understand. The upshot was an extended period in which our best physical theories were clearly inconsistent. A similar situation exists today, since our best theory of the structure of space–time, Einstein's general theory of relativity, is inconsistent with our quantum-mechanical view of microphysics.

Of course the importance of Triv depends on what we take to be the aim of logic, and in particular that of consequence relations. One influential view is that consequence relations are models of inference. By inference, we mean the process of adding new sentences to those we have already accepted by reasoning (rather than by observation). But it's clear that we would never accept every sentence. So if we ever accept inconsistent premises as a starting point for inference, we must use a paraconsistent logic in our inferences. Of course a defender of Lewis' position might argue that we never really accept inconsistent premises. But this seems untenable. After all, we are finite thinkers who do not always see the consequences of everything we accept. In particular, as we saw above, we may well not know our premises are inconsistent.

However, the view that consequence relations aim to model inference is hard to defend. First, as Gilbert Harman has argued, inference (in the sense of reasoning-driven change of belief) does not always involve adding to our set of accepted sentences. Sometimes, when we discover that some new sentence follows from the sentences we have already accepted, our response is to reject one or more of our premises rather than accept the new sentence. Second, real inference is constrained in ways that logical consequence relations are not. Many inferences that are 'logical' in the sense that they preserve truth and consistency are nevertheless a pointless waste of time for anyone to carry out. For example, disjoining a sentence with itself, and conjoining a sentence to itself, are perfectly correct from a merely logical point of view. But it would be ridiculous to waste our time idly inferring such consequences of the sentences we accept.

This leads to an alternative take on what the role of logic is. Inference is a highly pragmatic process involving both logical considerations and practical constraints of salience, along with rich evaluations of how best to respond both to our observations and to the consequences of what we have already accepted. So the role of logical consequence relations is clearly not to tell us what we may or should infer from the sentences we accept. But consequence relations can still tell us what is involved in accepting those sentences, that is they can operate as closure conditions on our com-

mitments. What I mean by this is that, in accepting a set of sentences, we can fairly be said to be *committed* to the consequences of those sentences – including the consequences we would never infer, whether because inferring them would be a waste of time, or because, rather than infer them, we would reject something that we now accept. In fact, I believe it's because we regard our commitments as closed under logical consequence that we are inclined to either accept sentences we find 'follow' logically from our commitments, or to reject some of the commitments they follow from.

This fits both our limits and our intuitions quite well. On the one hand, it does not assume inference is monotonic, or that a proper logic must shape its consequence relation to fit the exigencies of actual inferential practice. But on the other hand, it allows a substantial role for logical consequences in determining the content of our commitments. Without some such closure operation on commitment, our commitments would be too closely tied to the actual sentences we were willing to utter. And the point of assertion is not just to utter a sentence, but to express a commitment to what that sentence says.

This view of consequence relations provides a conclusive argument for paraconsistency. If we ever need to model commitments to inconsistent (and/or unsatisfiable) sets of sentences, then we will need a paraconsistent logic to do it. Otherwise, such commitments will be trivial, and indistinguishable from each other. For example, consider

1. Leibniz' calculus. This mathematical theory is inconsistent because it treats infinitesimals as both equal to zero and not equal to zero. From an addition equation in which an infinitesimal is an addend we can infer the corresponding equation without the infinitesimal, but division by infinitesimals is still well defined.
2. Naïve set theory. This mathematical theory is inconsistent because it claims that there exists a set corresponding to each condition on membership we can state. As a result, it claims that the Russell set exists, *viz.* the set of all sets that are not members of themselves. But if this set is a member of itself, then by its condition of membership, it is not a member of itself. And if it is not a member of itself, then by its condition of membership, it is a member of itself. So it is a member of itself if and only if it is not a member of itself, that it is an inconsistent object.

If we use classical logic to model commitment to these theories, they are equivalent (assuming we employ the same language to present them), and commitment to either is trivial. Avoiding triviality and distinguishing commitment to Leibniz' calculus from commitment to naïve set theory will require a paraconsistent logic.

In a closely connected use of logic, it is standard to call a set of sentences closed under a consequence relation a *theory*. If we are to cope constructively with inconsistent theories, the consequence relation they are closed under will have to be a paraconsistent one. Further applications for paraconsistent logic arise in epistemic logic, when we aim to model inconsistent beliefs, and in deontic logic, when we aim to model inconsistent obligations or rules. Standard modal techniques in epistemic logic represent a belief state in terms of the sentences true at all worlds at which the subject's beliefs are all true. Similarly, standard modal techniques in deontic logic represent obligations by appeal to the sentences true at all the worlds at which those obligations are met, and the demands of rules in a situation by appeal to the sentences true at all (later)

situations at which the rules have been obeyed. The result, if these worlds and situations provide classical valuations of the sentences of our language, is that beliefs, obligations, and the demands of rules are represented by classical theories. Coping with inconsistent beliefs, obligations, and rules will require instead that we represent them by paraconsistent theories, that is as sets of sentences closed under a paraconsistent consequence relation.

A more radical motivation for paraconsistency can be found in the work of the Australian dialetheists. The motives we have considered so far are to provide a coherent account of how to reason from inconsistent premises, or how to describe the content of an inconsistent theory. But the dialetheists aim at a logic that will allow that such theories are *true*. In particular, the dialetheists take paradoxes such as the liar and the paradoxes of naïve set theory at face value. That is, they view these paradoxes as proofs that certain inconsistencies are true. Consider the liar paradox:

L: This sentence is false.

The dialetheists claim that L is both true and false. We can prove this by *reductio* – suppose, for *reductio*, that L is true. Then what it says must be true. But it says that L is false. Therefore, if L is true, L must be false. Now suppose for *reductio* that L is false. But L says that L is false. Thus what L says is true, so L is true. Thus if L is false, then L is true. Therefore, L is true if and only if L is false. But if we accept the (semantic) law of excluded middle, that is that every sentence is either true or false, it follows that L must be both.[4]

Whether we want to reason in a non-trivializing way from inconsistent premises, or to model inconsistent beliefs, obligations, or theories, or to reason about an inconsistent metaphysics in which the liar sentence is both true and not true (and the Russell set both is and is not a member of itself), we will need a paraconsistent logic. Of course the particular kind of paraconsistent logic we will choose may depend on which of these applications we have in mind. But we will address the problem of choosing between various paraconsistent logics below.

3 The Sources of Trivialization

The usual definition of validity says that a sentence α can be validly inferred from a set of sentences Γ when and only when α is a semantic consequence of Γ, that is iff $\Gamma \models \alpha$, where:

(1) $\Gamma \models \alpha$ iff the truth of all Γ's members guarantees the truth of α.

This definition provides us with a straightforward explanation of the importance of deductive validity in evaluating arguments. Ideally, an argument should give conclusive reasons in support of its conclusion. And part of what we ask of conclusive reasons is that if the premises (the 'starting points' of reasoning) all hold, then the conclusion will be guaranteed to hold as well. (Of course 'holding' here is just being true.) But as introductory logic courses point out very early on, it is difficult, if not impossible, to say

in general when an argument in a natural language is valid. To give such a general account of validity, we would need a theory that identifies all the various sentences of a natural language and then tells us, in a logically revealing way, what it takes for each sentence to be true. Only then could we give general rules for determining whether meeting the truth conditions for an arbitrary set of sentences will guarantee that the truth condition for some other sentence is also met. Such a theory is far beyond us.

Logic courses that seek, nevertheless, to move towards such a theory quickly turn away from arguments expressed in English, French, Japanese, etc., to arguments expressed using formal languages with formal semantics. Such languages provide a formal syntax telling us what strings of symbols count as sentences, and a formal semantics specifying the truth (or 'satisfaction') conditions for sentences in the language. With these in hand, we can say clearly what it would take for a set's members to be true, and then use formal methods to show whether arranging things so that they are true will guarantee a conclusion sentence is true as well.

But there is another approach to capturing the notion of a valid argument, founded in what we regard as simple rules of good reasoning. On this approach, we are given a set of rules to follow in reasoning with sentences. These rules are based on the syntactic structure of the sentences, that is the symbols and how they are arranged in each sentence, rather than on an account of their truth conditions. A sequence of sentences (sometimes accompanied by other book-keeping devices) that follows the rules is called a *derivation*. The rules determine what sentences we may write down where in a derivation, and when a derivation is complete. The aim of this approach is to provide a set of simple, obviously correct steps that are collectively sufficient to capture all the consequences that follow from any premise set. This approach gives rise to the following account of a consequence relation:

(2) $\Gamma \vdash \alpha$ iff α can be derived from Γ.

One advantage of this approach to consequence relations is that it focuses our attention on the process of reasoning, rather than on 'meanings' that are taken to lie behind that process. Either way, as we will see, such consequence relations suffer from a serious limitation. In both cases, a tacit assumption is made about the premise sets we are thinking of – an assumption whose failure must trivialize these consequence relations.

We say that Γ is consistent if and only if it is impossible to derive a sentence and its negation from Γ,[5] and that Γ is satisfiable if and only if some valuation assigns a designated value to every sentence in Γ. A set is *maximally* consistent if it is consistent and adding any sentence to it would render it inconsistent. Similarly, a set is maximally satisfiable if it is satisfiable and no proper superset is satisfiable. The standard account of consequence relations results from a certain picture of what our consequence relations ($\models$ and $\vdash$) are supposed to preserve. Crudely, $\models$ is said to preserve truth while $\vdash$ preserves consistency. More carefully put, $\models$ preserves the satisfiability of all satisfiable extensions of the premise set, and $\vdash$ preserves the consistency of all consistent extensions, in the following senses:

(3) $\Gamma \models \alpha$ iff $\forall\Gamma'[(\Gamma' \supseteq \Gamma$ & Γ' is satisfiable$) \rightarrow \Gamma', \alpha$ is satisfiable$]$.
(4) $\Gamma \vdash \alpha$ iff $\forall\Gamma'[(\Gamma' \supseteq \Gamma$ & Γ' is consistent$) \rightarrow \Gamma', \alpha$ is consistent$]$.[6]

That is, if Γ' extends Γ satisfiably or consistently, then every consequence of Γ must satisfiably or consistently extend Γ'. We can say that the consequence relations $\models$ and $\vdash$ *beg no questions*, in the sense that closing Γ under $\models$ or $\vdash$adds nothing to Γ that is incompatible with any semantically or syntactically acceptable extensions of Γ. And of course, given the soundness and completeness of our system of derivation (3) and (4) are simply alternative definitions of the same consequence relation.

The principal point of this section can now be set out. If Γ is unsatisfiable (or inconsistent), then Γ has no satisfiable (consistent) extensions. You can never make a set that is already unsatisfiable or inconsistent into a satisfiable or consistent set by adding sentences to it. Thus every sentence trivially 'preserves' the satisfiability or consistency of all satisfiable or consistent extensions of such sets. So clauses (3) and (4) are satisfied for every sentence α, that is, every sentence is a consequence of an unsatisfiable or inconsistent set. The trivialization of unsatisfiable/inconsistent sets of sentences is deeply embedded in these standard accounts of consequence relations.

A logic is (minimally) paraconsistent iff it resists this trivialization, that is iff for some *classically* unsatisfiable or inconsistent set of sentences, the closure of the set under the logic's consequence relation is not the set of all sentences. But there are different ways to go about producing such a logic, rooted in different choices about what we choose to change in (3) and (4).

4 A Natural Taxonomy for Paraconsistent Logics

For a variety of reasons, including a focus on axiomatic presentations of logical systems and the vivid appeal of citing non-Euclidean geometries as a precedent, Jan Łukasiewicz, Jaskowski, and some other early figures saw paraconsistent logic (and nonclassical logic in general) on analogy with non-Euclidean geometry. Certain 'logical laws,' they suggested, could be treated as analogous to Euclid's parallel postulate in geometry, that is as characteristic of a particular *kind* of logic, but not essential for all logic.[7] This approach suggests a taxonomy of paraconsistent logics based on the classical theorems and rules that they retain, and those they give up.

In accord with this, the best known contemporary taxonomy of paraconsistent logic focuses on the tactics by which various paraconsistent logics avoid Triv, that is, the changes that are made to classical axioms and/or derivation rules. In Priest et al. (1989), three principal groups of paraconsistent logic are distinguished. First, Jaskowski's discursive logics are grouped with Rescher and Brandom's semantics for truth-value gaps and gluts and Schotch and Jennings' weakly aggregative logics, under the heading 'non-adjunctive' logics (Priest et al. 1989, p. 57). These logics block the inference from $\{p, \neg p\}$ to $p \wedge \neg p$. While classical contradictions, such as $p \wedge \neg p$, continue to be trivial for these logics, inconsistent sets that contain no contradictions can have non-trivial consequences. Second, the C-systems of Da Costa, among others, are labeled the 'positive-plus' logics. These logics lay the blame for trivialization on the classical theory of negation. They begin with an axiomatization of the positive (negation-free) fragment of classical logic, and then add to it a weakened account of negation, which blocks the derivation of arbitrary conclusions from $\{p,\neg p\}$, and even from $\{p \wedge \neg p\}$. The final (and preferred) place in the Priest/Routley/Norman taxonomy is reserved

for relevance-based approaches. These include the more conservative American school of non-dialetheic relevance logic, and the relevance-based dialetheic school originating in Australia. These logics block trivialization by means of a substantial departure from classical logic, both with regard to how they treat negation and with regard to the conditional.

But the taxonomy I will apply here focuses instead on the various ways in which the classical consequence relations can be modified to avoid Triv. This *strategic* choice seems to me a better basis for distinguishing various approaches to paraconsistency. After all (as we shall see) giving a fair interpretation of the inferential tactics employed will depend on understanding the strategic maneuvers that lie behind those tactics. And the result is both comprehensive and finer grained. Any paraconsistent logic must change some part(s) of (3) and (4), just as every paraconsistent logic must make some tactical modification(s) of the axioms and rules of deduction. And the very same change in the tactical rules can be arrived at by quite different strategies for changing (3) and (4); a tactical taxonomy must simply pass over these differences.

The trivialization of inconsistency that arises from clauses (3) and (4) above has its roots in three distinguishable sources:

(A) The classical accounts of satisfiability and consistency
If we propose a logic according to which (some) classically unsatisfiable/inconsistent sets turn out to be satisfiable/consistent after all, then of course preserving this new form of satisfiability or consistency can be the basis of a consequence relation that does not trivialize all classically unsatisfiable or inconsistent sets of sentences. This raises a kind of puzzle, though not a terribly deep one, regarding whether such a logic is *really* paraconsistent, since it simply aims to preserve a new form of consistency, and attributes this new form of consistency to sets of sentences that classical logic regards as inconsistent.

(B) An unsatisfiable or inconsistent set lacks the only *property that the consequence relation seeks to preserve. As a result, there are no grounds for constraining the consequences of such sets*
Thus the fact that the consequence relation is defined in terms of preserving consistency or satisfiability, rather than some other (desirable) feature of our premise set is essential to Triv. If we chose instead to preserve some new (desirable) feature of our premise sets, then the fact that a set lacks consistency or satisfiability need not imply that all constraints on the consequence relation are removed.

(C) The fact that the metalinguistic '$\rightarrow$' holds whenever the antecedent (i.e. that Γ' is a consistent or satisfiable extension of Γ) is false
If we were to alter our reading of this connective, we might create room to deny that clauses (3) and (4) hold trivially whenever Γ is inconsistent. This third strategy has a relevance flavor about it, though the paraconsistent relevance logics we will consider here all locate the problem with (3) and (4) in the classical account of consistency and satisfiability.

The roles of A, B and C in producing Triv lead to three basic strategies for avoiding it:

Strategy A: New accounts of 'truth' and 'consistency'

This is the road most traveled in paraconsistent logic. Its practitioners include those who propose a dialetheic account of truth. But they also include more conservative figures who take the new semantic values they propose for sentences to express epistemic commitment, or some other more metaphysically modest status than truth, *tout court*. On our taxonomy, any paraconsistent semantics that operates by non-trivially assigning *designated values* to all members of some classically unsatisfiable sets of sentences falls into this group. That is, this approach assumes that whenever α is not a (semantic) consequence of Γ in some logic, this must be because the semantics provides an acceptable valuation V such that for all $\gamma \in \Gamma$, $V(\gamma) \in \{v: v \text{ is designated}\}$ and $V(\alpha) \notin \{v: v \text{ is designated}\}$. Though truth is the standard example of a designated value, it's not necessary to interpret all, or even any, designated value as a formal theory of truth. N. D. Belnap, for instance, reads the values of Dunn's four-valued logic epistemically, as "told true," "told false," "told both" and "told neither." But even when we keep this interpretational latitude in mind, this account of the consequence relation is very constraining. It focuses all our attention on the assignment of values to sentences, the distinction between designated and undesignated values, and the consequence relation we get when $\Gamma \models \alpha$ is said to hold if and only if α is assigned a designated value whenever all the members of Γ are assigned designated values.

Strategy B: Preservationism

This approach has been less widely pursued. But it has, as we will see, some clear advantages over the first. The general idea has been put in various ways –

> Don't make things worse. (P. K. Schotch)
> Find something you like about your premises, and preserve it. (R. E. Jennings)

As we have seen, from the classical point of view, there is nothing worse than an inconsistent, unsatisfiable set of sentences. Classical logic aims only to preserve consistency and satifiability; once these are lost, there is nothing left that a classical logician cares to preserve. But there are other features of premise sets that are worth preserving. Non-triviality is the most obvious example, but we can (and will) be more specific. Going from a set of sentences that merely includes p and $\neg$p for some sentence p, to the set of all sentences clearly does make things worse. It takes us from a set that we could use to represent someone's commitments or the contents of an inconsistent theory in a non-trivial way, to a trivial representation of the commitments or the theory. And this suggests that we pursue precisely the constructive project that Jennings proposes.

This proposal creates a significant widening of the options before us. To propose another slogan, it liberates us from the tyranny of designated values. That is to say, unlike the first approach, it does not demand that we find a way to assign a designated value to the premises of a rejected consequence while assigning a non-designated value to the conclusion. In fact, the whole business of assigning values to sentences can be left just as classical logic has it, while we concern ourselves with features of sets that some inconsistent sets possess, and that are worth preserving.

Strategy C: A new metalinguistic '→'

Classically, the '→' in clauses (3) and (4) holds if either the antecedent is false, or the consequent is true. But if we demand some relevant connection between the antecedent and the consequent before declaring that conditions (3) and (4) are met, then the mere fact that the antecedent is false does not imply that the whole statement is true. Oddly, to the best of my knowledge those who have worked to apply relevance logic to these issues have all focused on the first approach to paraconsistency, rather than this one.

But there is an explanation of this. The tradition of relevance logic has endorsed the 'preserving designated values' account of consequence relations, while insisting that the consequence relation (and its object language reflection, the conditional) must respect considerations of relevance. Thus relevance logic demands that we reject both $\{p \wedge \neg p\} \models q$ as a consequence, and $(p \wedge \neg p) \rightarrow q$ as a conditional theorem. But if we continue to view consequences as a matter of preserving designated values, then to reject $\{p \wedge \neg p\} \models q$, we must be able to assign a designated value to $p \wedge \neg p$ while not assigning a designated value to q. And having done so, we will obviously be able to assign a designated value to the antecedent of our conditional, while not assigning a designated value to the consequent. So long as we treat the consequence relation as merely preserving designated values, there is no need to think of the corresponding conditional as preserving something more than designated values either. So from the perspective of traditional relevance logics, changing the metalinguistic → in the sort of way required here would involve doing just what the first approach to paraconsistency demands, namely producing a valuation which designates all the premises while failing to designate the conclusion.

However, a broader perspective might hold that → must preserve something other than designated values. This third approach is clearly distinct from the first, and worth pursuing independently. I recommend it to anyone interested in such preservationist conditionals.[8]

5 Paraconsistent Logics

Non-adjunctive logics

S. Jaskowski proposed the first formal paraconsistent logic in 1948 (reprinted in English in 1969). His approach was motivated by the idea of a discussion involving more than one participant, each contributing a consistent set of assertions. If we treat the assertions of each participant as 'holding' for the discussion, we end up with a potentially inconsistent set of sentences representing the overall product of the discussion. If we want a consequence relation under which we can reasonably close such sets of sentences, it must be a paraconsistent one. Moreover, the consequence relation should reject adjunction, that is the principle that $p,q \models p \wedge q$. After all, the fact that someone has contributed p to the discussion and someone has contributed q to the discussion in no way implies that anyone has contributed $p \wedge q$. Neither contributor need be in any way committed to $p \wedge q$. In fact, both contributors may regard p and q as incompatible

with each other. Jaskowski's main concrete proposal for a logic that would respect these constraints is his D2.

To produce D2, Jaskowski appeals to the possibility operator of the strong modal logic S5. He lays his proposal out in terms of a correspondence between a discussive logic and an underlying modal logic. For each participant in the discussion, we consider the worlds satisfying all the claims they contribute to the discussion. Then we say that 'p' is *discussively* true if and only if '$\Diamond$p' is true at a world to which all and only these worlds are accessible. We say, further, that p discussively implies q if and only if $\Diamond p \rightarrow q$ is true in the underying modal logic, and (in an oddly asymmetrical definition) p and q are discussively equivalent if and only if $(\Diamond p \rightarrow q) \wedge (\Diamond q \rightarrow \Diamond p)$ holds in the underlying logic.

In S5, as in all standard modal logics, $\{\Diamond p, \Diamond q\} \not\models \Diamond(p \wedge q)$. Thus in D2, the corresponding discussive logic, $p,q \not\models (p \wedge q)$. So the rejection of adjunction is supported by this modal reading. Of course, the simplest view[9] today of how to represent a few people's commitments would be to have a separate accessibility relation for each individual, such that all and only the worlds at which someone's commitments are true are accessible to the actual world for that individual. The result, from the point of view of modal logic, would be a set of necessity operators, one for each individual, with the truth condition that p is discussively true if and only if $\Box_i p$ is true for some $\Box_i$. This approach to discussive logic would make it clear that (assuming for now the consistency of each individual's contribution to the discussion) each individual's contributions to the discussion should be closed under adjunction (in fact, will constitute a classical theory), even though the sum total of those contributions should not. Retaining some aggregative force here (particularly just how much we can or should retain) is an issue we will return to when we discuss the weakly aggregative logics of Schotch and Jennings.

This debate over the best means to arrive at a discussive logic aside, I'll finish here by making two points. First, Jaskowski's systems clearly fall within our first class of paraconsistent logics. p is 'true' in this logic if and only if '$\Diamond$p' true by the standards of the corresponding modal logic. The consequences of a set of sentences, Γ, are precisely those sentences α such that whenever $\Diamond\gamma$ is true for each γ in Γ, $\Diamond\alpha$ will also be true in the corresponding modal logic. Symbolically,

$\Gamma \models_d \alpha$ iff $\Diamond(\Gamma) \models_L \Diamond\alpha$, where $\models_L$ is the modal logic in question.

So preservation of truth for the admissible evaluations is the criterion of consequence for discussive logics. Second, the consequence relation here is the classical singleton consequence relation. A holds discussively if and only if some B such that {B} has A as a classical consequence has been added to the discussion by some participant. And the (propositional) theorems of the logic are simply the classical tautologies.

A different, very straightforward approach to non-adjunctive logic is due to Rescher and Brandom (1980). Beginning with the set of classical valuations, they propose two semantic operations: superposition and schematization. Applying superposition to two valuations produces a valuation assigning t to every sentence assigned t by either of the input valuations. Applying schematization produces a valuation assigning t to only those sentences assigned t by both of the input valuations. The full set of

Rescher–Brandom valuations results from closing the set of classical valuations under superposition and schematization.

The upshot is very similar to Jaskowski's D2: a logic whose theorems are just the classical tautologies, and whose consequence relation (as defined over the class of all such valuations) is the classical singleton consequence relation.

C-systems and weakened negation

Newton da Costa proposed his C-systems in the 1960s; a semantics for these logics did not emerge until later. But we will focus on the semantics here, since they are quite intuitive, and make it clear that da Costa's approach belongs solidly in the first category of our taxonomy. A da Costa evaluation maps every formula to t (the designated or 'true' value) or f, given an assignment to the sentence letters, as follows:

1. $v(A \wedge B) = t$ if and only if $v(A) = t$ and $v(q) = t$
2. $v(A \vee B) = t$ if and only if $v(A) = t$ or $v(B) = t$
3. $v(\sim A) = t$ if $v(A) = f$
4. $v(A) = t$ if $v(\sim\sim A) = t$

This class of valuations gives us the non-implicational fragment of the weakest C-system, C_ω. The C-systems are another example of our first class of paraconsistent logics: $\Gamma \models_{C\omega} \alpha$ if and only if every valuation assigning 1 to all members of Γ also assigns 1 to α. As is clear from the clauses for valuations, da Costa's logic is classical except in terms of how it treats the negation '~'. But (in part as an inevitable result of this choice) the negation is very nonclassical. In fact, Priest et al. (1989) argue that it is not a negation at all, but rather a "sub-contrary forming functor," that is, a functor *f* such that while $(p \wedge fp)$ can be true, $(p \vee fp)$ must be true. They point out as well that many very basic consequences involving classical negation fail for this negation. And while things get (for those wffs that behave 'consistently') more classical in the stronger C-systems, this does not help with the basic difficulties described by Priest, Routley, and Norman.

Further details of the C-systems and other systems that have emerged in the research programs of da Costa and his colleagues are left aside here for want of space.[10] The main point I want to emphasize for now is that the issue of how to tell a real negation from a pseudo-negation is a persistent problem for paraconsistent logic. It is by no means an easy question to answer. The initial assumption tends to be that classical negation is the paradigm case of a 'real' negation. But while this may represent the position a paraconsistent logician must *respond* to when arguing with critics of paraconsistency, it is unfair to begin by assuming that all the features of the classical negation are features that a good negation should have. Paraconsistent logic must insist that some, at least, of what classical logic does with negation is mistaken, at least in some applications.

Relevance logics

These logics have their roots in the program of relevance logic pioneered by Ackermann, and then developed and greatly extended by Anderson, Belnap, and their students. In

these logics a tight correspondence is assumed to hold between conditional sentences and the consequence relation. But both are taken to be subject to strong constraints of relevance. In particular, a variable-sharing constraint is urged. $\{p, \neg p\} \models q$ is rejected on the grounds that there is no relevant connection of meaning between the premise set and the conclusion. Such a connection (at the propositional level) would demand at least that some variable appearing in the conclusion also appear in the premises. For reasons of space, we'll confine ourselves to examining three consequence relations that have their roots in the relevance program, avoiding the issue of conditional sentences whose logic and semantics would add quite a bit of complexity to the picture.

LP, the logic of paradox, is dramatically different from the C-systems in two respects. First, its semantics allows for sentences to be assigned both true and false at one and the same time. The values assigned to sentences in LP can be described as the nonempty subsets of the set {true, false}. Thus an assignment will assign one of the values {t}, {f}, or {t,f} to each sentence. We begin with an assignment of values to the sentence letters, and then extend it to the rest of the sentences following the clauses:[11]

1. (a) $t \in v(\neg A)$ iff $f \in v(A)$ (b) $f \in v(\neg A)$ iff $t \in v(A)$
2. (a) $t \in v(A \wedge B)$ iff $t \in v(A)$ and $t \in v(B)$ (b) $f \in v(A \wedge B)$ iff $f \in v(A)$ or $f \in v(B)$
3. (a) $t \in v(A \vee B)$ iff $t \in v(A)$ or $t \in v(B)$ (b) $f \in v(A \vee B)$ iff $f \in v(A)$ and $f \in v(B)$

The consequence relation is again defined in the standard way

$\Gamma \models_{LP} \alpha$ iff every such valuation making $t \in v(\gamma)$ for all $\gamma \in \Gamma$ also makes $t \in v(\alpha)$.

So once again we have an example of our first class of paraconsistent logics. The upshot here is quite clean and straightforward. The theorems (i.e. the consequences of the null set) are just the classical tautologies. And unlike the C-systems, the negation here looks very much like classical negation. The usual equivalences all hold:

$\{\neg A \wedge \neg B\} \models_{LP} \neg (A \vee B)$; $\{\neg(A \vee B)\} \models_{LP} \neg A \wedge \neg B$;
$\{\neg A \vee \neg B\} \models_{LP} \neg (A \wedge B)$; $\{\neg(A \wedge B)\} \models_{LP} \neg A \vee \neg B$
$\{A\} \models_{LP} \neg\neg A$; $\{\neg\neg A)\} \models_{LP} A$

Transitivity, as well as introduction and elimination inferences for $\wedge$ and $\vee$ are all retained.

The main classical principle that fails here (thereby preventing the trivialization of inconsistent premise sets) is disjunctive syllogism:

$A, (\neg A \vee B) \not\models B$

The failure of this principle is easy to see. If $v(A) = \{t,f\}$, while $v(B) = f$, then A has a designated value, as does $(\neg A \vee B)$, but B does not.

Having introduced truth-value 'gluts' here, that is sentences assigned both true and false, it might well seem natural to consider the possibility of truth-value gaps, that is

sentences assigned neither true nor false. But in fact we can achieve the same effect with gluts alone, if we recognize an important fact about them: Sentences that are both true and false are both correctly assertable and correctly deniable.

Before we can apply this recognition to arrive at a different consequence relation in the relevance family, we will have to take a short detour back into classical logic, to bring clearly to mind some fundamental symmetries that apply there, and that have been lost in the transition to LP. Just as (for classical logic) 'truth' is the value that sustains the assertion of a sentence, that is that makes its assertion correct, 'false' is the value that sustains the denial of a sentence, that is that makes it correct to deny that sentence.

The notion that there are two distinct attitudes we can take with regard to declarative statements, assertion and denial, has often been rejected in favor of the view that we can make do with assertion alone. And in the context of classical logic there seems to be little reason to object to this. The position I take by denying that I am hungry seems equivalent in every important respect to the one I take by asserting that I am not hungry. However, a certain amount of expressive power is lost when we dispense with denial as a separate attitude, and with it we lose the capacity to express an important constraint on the consequence relation.

So far we have represented the consequence relation as a relation between sets of sentences on the left and individual sentences on the right. This is pretty standard practice, but it has some drawbacks. It makes the consequence relation something asymmetrical from the outset. But there are important and illuminating symmetries hidden behind this asymmetrical veil. We can reveal them by adopting a picture of the consequence relation that puts sets of sentences on both sides. We will say that $\Gamma \models_c \Delta$ iff every classical valuation satisfying all of Γ's members also satisfies some member of Δ. But now we can also say, equivalently, that this holds iff every classical valuation making all of Δ's members false also makes some member of Γ false.

If we have the notion of denial in hand, as well as the notion of assertion, then we can describe the condition under which $\Gamma \models_c \Delta$ holds in a different, but equivalent way. We can require that if every member of Δ is correctly denied in some valuation, then some member of Γ must also be correctly denied on that valuation. Of course, this is just the contrapositive of the truth-preserving account (i.e. correct assertability) of validity. However, preserving correct deniability from the right side of $\models$ to the left, as well as truth from the left to the right, can impose a real additional constraint on the consequence relation when our logic is not classical.

In particular, consider LP. What values shall we preserve from right to left, given that we preserve {t,f} and {t} from left to right? If the value {t,f} (often called 'both') is truly paradoxical,[12] one way to understand what we mean by that is to say that it sustains both the correctness of asserting a sentence that has it, and the correctness of denying that sentence. A sentence that is true and false both is both correctly assertable and correctly deniable. LP is the logic we get when we take the first point (that such sentences are correctly assertable) and ignore the second. As a result, LP preserves only {f} from right to left. But what we really should demand is that LP preserve both {f} and {t,f} from right to left:

$\Gamma \models_{LP^*} \Delta$ iff every LP valuation making $t \in v(\gamma)$ for all $\gamma \in \Gamma$ also makes $t \in v(\delta)$ for some $\delta \in \Delta$,

and every LP valuation making $f \in v(\delta)$ for all $\delta \in \Delta$ also makes $f \in v(\gamma)$ for some $\gamma \in \Gamma$.

Such a logic preserves both LP-correct assertability from left to right, and LP-correct deniability from right to left.

What is SLP, this symmetrical version of the LP logic? The answer is, it's a step along the way to a logic familiar to students of relevance logic, namely first degree entailment (FDE).[13] The symmetrical form of LP behaves just like FDE except when the premise set cannot be consistently asserted *and* the conclusion set cannot be consistently denied. In these doubly (classically) trivial cases, this symmetrical version of LP trivializes just as classical logic does. One further step is required to arrive at FDE. We must coordinate our use of LP assignments to render the premises consistently assertable and the conclusions consistently deniable. If and only if there is an LP valuation 'satisfying' the premise set and an LP valuation 'falsifying' the conclusion set such that the two agree on their classical sub-valuations, and don't overlap on the sentence letters assigned 'both,' then the FDE consequence relation fails to hold between the premises and the conclusion.

The main point that I want to make about SLP and FDE here is that they very simply restore some fundamental symmetries present in classical logic, and given up in LP. For instance, no sentence is trivial on the left in LP, that is, there is no sentence A such that $\{A\} \models_{LP} \Delta$, for all sets Δ. But every sentence that is trivial on the right in classical logic is also trivial on the right in LP. These right-trivial sentences are, of course, the classical tautologies. That is, A is a classical tautology iff $\Gamma \models_{LP} A$, for all Γ. But in classical logic there is a perfect symmetry (duality) between the sentences that are trivial on the left and the sentences that are trivial on the right. For instance, $(p \wedge \neg p)$ is trivial on the left, and $(p \vee \neg p)$ is trivial on the right. LP forces us to surrender this symmetry; SLP and FDE restore it.

One final point is in order here. It is possible to get exactly the same consequence relations we have arrived at with the help of LP valuations while retaining a *fully classical* semantics. The trick is to change what the consequence relation is required to preserve. Rather than preserve truth from left to right (and falsehood from right to left), we can preserve a class of projections capable of producing a consistent image of our inconsistent set. We omit the details for reasons of space – but the general lesson is well worth drawing: one and the same consequence relation can be underwritten by quite different semantics. As a result, a paraconsistent consequence relation that is arrived at by one sort of change to our clauses (3) and (4) can (at least sometimes) also be achieved by a different sort of change.

Adaptive logics

Diderik Batens, inspired initially by L. Apostel, is the central figure in a research program working on a range of paraconsistent logics at the University of Ghent. Together with students and colleagues he has focused on a class of logics which he calls adaptive logics. The motives that lie behind these logics are a good fit with the preservationist approach to paraconsistent logic: Batens and his co-workers have been con-

cerned with not making things worse when inconsistency rears its ugly head. However, the means by which they achieve these goals still focus, in the conventional way, on the preservation of designated values. Thus Batens says,

> An adaptive logic La localizes the abnormal properties of Γ, safeguards the theory from triviality by preventing specific rules of L (the initial, non-paraconsistent logic) from being applied to abnormal consequences of Γ, but behaves exactly like L for all other consequences of Γ . . . The (dynamic) proof theory of adaptive logics is based on the idea that a formula is considered to behave normally 'unless and until proved otherwise'. The semantics is better understood by another metaphor: La interprets Γ by eliminating its unnecessarily inconsistent L-models. For ACLuN2, e.g., the La-semantic consequences of Γ are the formulas true in the Lf-models of Γ that are minimally abnormal (not more inconsistent than required to make Γ true).[14]

Both the idea that we should not "make things worse," as Peter Schotch urges, and the idea that to constrain the consequences of a set of sentences we must find a way to make the set 'true' in some sense, are in the air here. As a result, these systems constitute a borderline case for this taxonomy. For reasons of space we will focus on the propositional fragment **PI** of the base paraconsistent logic, CLuN, and then explain briefly how the adaptive logic based on CLuN works.

PI includes sentence letters, and the usual truth-functional connectives. Their treatment is modified, however, from the familiar classical one, by the explicit surrender of the consistency assumption:

If, on some admissible valuation v, $v(A) = t$, then $v(\neg A) = f$.

This assumption is avoided by the simple expedient of making an assignment to the negations (the formulae that have '$\neg$' as their main connective) a separate part of producing a valuation:

So we take as the base of our evaluation both an assignment v_b to the sentence letters and an assigment v_b to the negations:

$S \rightarrow t,f$
$N \rightarrow t,f$

The valuation v that results from this assignment is arrived at by the following rules

$v(S) = t$ iff $v_b(S) = t$ where S is a sentence letter
$v(\neg A) = t$ iff $v(A) = f$ or $v_b(\neg A) = t$
$v(A \wedge B) = t$ iff $v(A) = t$ and $v(B) = t$
$v(A \vee B) = t$ iff $v(A) = t$ or $v(B) = t$

This makes the main features of how this logic copes with inconsistency pretty clear. Note in particular that the effect of assigning values directly to the negations is only to make some negations true which would otherwise be false. Whenever a negation would be true given the initial assignment to the sentence letters alone, it remains true after

the effect of the v_b assignment to the negations is factored in, since *either* the usual truth condition for '¬' or setting $v_b(\neg A) = t$ is sufficient to make $v(\neg A) = t$. So we can arbitrarily force any negation we like to receive the value true simply by assigning it true in v_b, but we cannot arbitrarily force negations to be false. Any negation whose truth follows from the classical components of the PI valuation will be true in the PI valuation. But in general many other negations will also be true.

It is dead simple, of course, to construct valuations that make an inconsistent set such as {p,¬p} true, while avoiding trivialization for some sets of sentences. The resulting logic is clearly paraconsistent.

PI and CLuN are clearly paraconsistent logics in the traditional, 'truth'-preserving mode. But the adaptive logics based on CLuN have something of the spirit of preservationism about them. A central idea for the propositional adaptive logic is the following theorem:

$$\vdash_c A \text{ iff, for some } C_1, \ldots C_n\ (n \geq 0),\ \vdash_{\mathbf{PI}} ((C_1 \wedge \neg C_1) \vee \ldots \vee (C_n \wedge \neg C_n) \vee A$$

That is, if A is a theorem of classical logic, then A will hold for **PI** as well, *unless* some of the C_i behave inconsistently. This theorem suggests a plan for the adaptive logic.[15] If we begin with the assumption that our premises are consistent, we can prove anything that classical logic allows us to prove. Suppose that we have proved A from our premises, using classical logic. Then by the theorem, there is some set $\{C_i\}$ of sentences whose consistent behavior is sufficient to assure us that A really does follow (in the **PI** sense) from our premises. But of course the assumption that the members of $\{C_i\}$ do behave consistently may turn out to be wrong. So the proof is *tentative*. It depends on the consistent behavior of the $\{C_i\}$ associated with A, which are kept track of at each step of the proof. If we should find, in the course of proving further consequences of our premises, that some C_i behaves inconsistently, we would have to *withdraw* our earlier proof of A. So proofs within an adaptive logic can require the deletion of earlier lines from the proof, based on what is shown later in the proof.

The result is a system of proof that allows us to derive the consequences that follow from our premises in the *minimally* inconsistent **PI** models of our premises. So these adaptive logics do involve a modification of the notion of a consequence. They make consequences turn on preserving something other than satisfiability. More inconsistent sets of sentences *are*, after all, still satisfiable, but we confine our attention to the minimally inconsistent sub-class of the models satisfying our premises, and define as a consequence what holds in these. From the point of view of our semantic clause (3), we have gone from requiring that a consequence α preserve the satisfiability of every satisfiable extension of our premises Γ to requiring that α preserve the minimally-inconsistent satisfiability of every minimally-inconsistent satisfiable extension of Γ. This is clearly a preservationist maneuver.

Weakly aggregative logics

With the work of P. K. Schotch and R. E. Jennings we finally arrive at a logic that is clearly and self-consciously preservationist. Rather than find a way of making premises true and conclusions false in some valuation in order to defeat an undesirable conse-

quence, Schotch and Jennings identify a set of desirable properties that some inconsistent sets have, and propose a logic that preserves these properties.

To begin, we need the notion of a *level*, a generalization of consistency that our consequence relation will be required to preserve. The level of a set of formulae, Γ, is the minimum number, n, such that Γ can be partitioned into n consistent subsets. A formula α is a level preserving consequence (LPC) of Γ (Γ forces α, or, more formally, $\Gamma \vdash_{LPC} \alpha$), iff α is a classical consequence of some cell in every partition of Γ amongst n sets. Three important consequences of this are immediately apparent. First, any set not including a contradiction will have some well-defined level (possibly an infinite level, if Γ is an infinite set). Second, no set with a well-defined level will have any contradiction as an LPC. Third, no set including a contradiction has a well-defined level, since no partition of such a set will have only consistent cells. We assign the 'level' ∞ to such sets. Since such sets lack the property (having a well-defined level) that these logics aim to preserve, their consequences are trivial for this logic.

This non-adjunctive system goes beyond the completely non-adjunctive approaches of Jaskowski, and of Brandom and Rescher, allowing a weakened form of aggregation.[16] Given a level of 2, a set closed under LPC will include the disjunction of pairwise conjunctions of all triples in the set; given a level of 3, the set will include the disjunction of the pairwise conjunctions of all quadruples in the set, and so on. These disjunctions of pairwise conjunctions capture fully the adjunctive strength of forcing: Where n is the level of Γ, closing under the classical consequences of singleton subsets of Γ together with the rule $2/n + 1$, which allows us to infer from any $n + 1$ formulae the disjunction of all their pairwise conjunctions, is consistent and complete with respect to forcing.[17]

Unlike the relevant systems[18] Schotch and Jenning's non-adjunctive logic agrees with some important intuitions concerning inconsistent input and conjunctions: if a set of claims including 'p,' '~p' is closed under forcing – the Schotch/Jennings inference relation – ~(p ∧ ~p) is forced but (p ∧ ~p) is not. So a computer using forcing would not regard itself as having been told the conjunction was true in such a case. This seems the right answer for an appropriately conservative inference engine to give.

Priest and Sylvan have objected to this, claiming that the computer has indeed been told the conjunction is true by implication.[19] Priest and Sylvan's argument assumes that the fact that conjunction just *is* the connective which gives a truth when the two things it joins are true implies that if adjunction fails as a rule of inference, then the connective it fails for can't be conjunction. But a crucial assumption about inference underlies this objection, *viz.* that truth-preservation is sufficient for the correctness of an inference. If correctness requires more than just truth preservation, the failure of {A,B} to imply (A ∧ B) does not show the logic has a non-standard truth condition for '∧.' And this is precisely the case for LPC. LPC requires preservation of level as well as truth. Adjunction *is* truth preserving, on Schotch and Jennings' view – it fails to be a rule of the system only because it fails to preserve level.

An interesting pragmatic element in inference emerges for non-adjunctive logics. Some individual conjunction-introductions are guaranteed to be level-preserving, but allowing conjunction-introduction in general leads to explosion. We must *decide* which conjunctions to adopt (if any). Our reasons for choosing some rather than others will normally derive from our epistemic aims. These often make some conjunctions

indispensable; but they usually leave us also with a wide field of potentially interesting or desirable conjunctions whose value remains to be determined. On the question of which non-level increasing conjunctions to adopt and which to avoid, the logic is silent: adding them is just like extending any classical theory by adding further sentences consistent with, but not implied by, the theory. Which non-level increasing conjunctions we will accept depends on which are valuable – which seem required for effective application of the theory, which promise to produce interesting predictions without absurdity, and so on.

With regard to applications, this feature of LPC suits Bohr's theory of the hydrogen atom very nicely. The development of old quantum theory (OQT) involved a gradual clarification of which classical results can be applied in the quantum domain, and when. These results emerged from a process of investigating which classical results can be conjoined with quantum principles to good effect, in effect a choice of level-preserving conjunctions from among many candidates. Thus one prediction a forcing-based model makes concerning OQT is confirmed by the history of OQT: the addition of conjunctions of classical principles with quantum principles to the theory is ampliative, and requires independent theoretical and/or empirical justification. Conjunction introduction is not a trivial inference, but a substantial step that carries both risks and potential rewards.

6 Current Issues

Paraconsistent logic remains controversial. Many logicians still defend Triv and reject the entire field of paraconsistency as misguided. And even within the paraconsistent camp, the different approaches involve very different views of negation, consequence relations, and the nature of logic in general. As a result, proponents of one approach are often very critical of others. The breadth and range of positions in this field, only briefly and partially outlined here, makes giving a simultaneously brief and fair summary of the state of the field an impossible task. In this last section I want to briefly touch on two issues in paraconsistency that are at the center of my present work, and indicate how I hope, in further work, to provide some insight into them. The positions I will be sketching here are my own, and others will have their own responses (and, no doubt, trenchant criticisms of mine). So at this point I surrender all pretence of giving a balanced discussion.

Negation

As we have seen already, debates about negation, and especially about how to tell whether an operator in some logic is *really* a negation, have played a central role in the development of paraconsistent logic. On this issue I have a modest proposal: negation is *denial* in the object language. This is, I think, at least as credible a position as the standard relevance view that the object language conditional must express the (metalinguistic) consequence relation. On this view, a satisfactory paraconsistent logic needs an extended understanding of deniability that corresponds to the extended notion of assertability proposed by the logic. Only then can the symmetries between assertion

and denial, negated and un-negated sentences, and the duality of $\wedge$ and $\vee$ be preserved. This leads us to a requirement on a satisfactory paraconsistent logic that some present systems meet and some fail: contraposition for the consequence relation:

$\Gamma \vdash \Delta$ iff $\neg(\Delta) \vdash \neg(\Gamma)$ $\quad (\vdash/\neg)$
($\neg(\Delta)$ is the set of sentences that results by negating each element of Δ.)

The idea here is that when we put a set of sentences in premise position, we are treating them as *in some sense assertable*. We take $\Gamma \vdash \Delta$ to say that if Γ is, in that sense, assertable, then so is some member of Δ. More explicitly, for every assertable extension of Γ, some element of Δ will be an assertable extension of the extension. So committing ourselves to Δ (in the sense of committing ourselves to accepting some member of Δ) begs no questions. It does not in any way extend the commitment we have already made in accepting Γ. But symmetrically, Γ should preserve the *deniability* of Δ, that is some member of Γ must be an acceptably deniable extension of every acceptably deniable extension of Δ. So if '$\neg$' really is the object language 'image' of denial, then from the right-to-left preservation of deniability must follow the left-to-right preservation of the assertability of the negations.

Interpretation

B. H. Slater has offered a Quinean objection to paraconsistent logic.[20] His claim is that if, in some logic, $\{A, \neg A\} \nvdash B$, this is just evidence (and well-nigh conclusive evidence) that '$\neg$' just isn't negation. In reply to this, I have pointed out that some preservationist logics, such as Schotch and Jennings' forcing, retain a fully classical semantics. They obtain $\{A, \neg A\} \nvdash B$ not by making A and $\neg$A somehow 'satisfiable' or 'consistent', but by finding another desirable property that $\{A,\neg A\}$ has, and that $\{A,\neg A,B\}$ lacks. But, taking this reply a step further, I have also shown that LP can be given a preservationist semantics based on using ambiguity to project consistent images of inconsistent sets of sentences. Details aside, the first lesson I think we should draw from this is that we should keep a healthy distance between our interest in various paraconsistent consequence relations and the particular semantics and definitions of consequence that we have used to arrive at them. The two are, in general, separable.

This also raises a more general question: can all non-preservationist paraconsistent logics be reinterpreted in preservationist terms? At least in some cases, such reinterpretations seem to be illuminating. And they do have the rhetorical advantage, when they begin from classical semantics, of saying nothing about the premise sets and the new consequence relation that classical logicians are inclined to deny. Preserving something other than satisfiability or consistency allows us not to argue with each other about what satisfiability or consistency *really* mean. Of course, this is not meant to cut off such discussions – as we remarked above, there is no reason to assume the classical account of these things must be the right one. But if our concern is to arrive at consequence relations for paraconsistent applications such as non-trivial inconsistent theories, we may do well to demonstrate the tenability and usefulness of such relations within a classical framework before (or even rather than) taking on the job of replacing it.

Notes

1 Strictly speaking, this should be set out more generally:

$$\frac{\Gamma \vdash \alpha,\ \Gamma \vdash \neg\alpha}{\Gamma \vdash \beta} \qquad \frac{\Gamma \models \alpha,\ \Gamma \models \neg\alpha}{\Gamma \models \beta}$$

2 Woods (1975: 165–7).

3 We can show that a set is inconsistent, by showing that some contradiction follows from it. And we can show that a set is consistent by presenting a model of it. But the process of finding a model of a set of sentences is not something we can go through systematically in such a way as to be sure that we will find a model at some finite point along the way if one is to be found. And when we fail to show a contradiction follows from some set of premises, our failure does not show that no contradiction follows.

4 Of course many have suggested that we reject excluded middle to avoid this unwelcome and radical conclusion. For example, we can add a 'gap' value to our semantics, which is neither true nor false, and assign 'gap' as the value of L. But as Graham Priest has often argued, this is not enough to resolve the problem. We can replace L with L′, the 'strengthened liar':

L′: This sentence is not true.

Now we can argue much as we did before. Suppose L′ is true. Then what it says must be true, so L′ is not true. Suppose L is not true. But this is precisely what L′ says of itself. So what L′ says is true, so L′ is true. Thus L′ is true if and only if L′ is not true. But now adding the value 'gap' (or some other, non-true value) to the usual values true and false is no help. See Priest (1995).

5 In fact, there are several notions of consistency, all closely related. A set that does not include both a sentence and its negation is called negation consistent; any set that doesn't include every sentence in the language is said to be absolutely consistent. In classical logic, the closure under deduction of any negation inconsistent set is the set of all sentences, i.e. absolutely inconsistent. In fact, the closure under deduction of any set whose closure under deduction is negation inconsistent is also absolutely inconsistent. The notion of consistency I am using here is that of any set whose closure under deduction is negation consistent. If we were sticking with classical logic, this would be equivalent to speaking of any set whose closure under deduction is not absolutely inconsistent.

6 Here we're using the notational convention that $\Gamma, \alpha = \Gamma \cup \{\alpha\}$. We could present (3) and (4) in slightly different form:

(3′) $\Gamma \models \alpha$ iff $\forall\Gamma'[(\Gamma'$ is a maximal satisfiable extension of $\Gamma) \rightarrow \alpha \in \Gamma']$.
(4′) $\Gamma \vdash \alpha$ iff $\forall\Gamma'[(\Gamma'$ is a maximal consistent extension of $\Gamma) \rightarrow \alpha \in \Gamma']$.

7 See Arruda (1989) for further details.

8 See Jennings and Johnston (1983), and D. Sarenac (2000) for work on such conditionals at the object-language level.

9 Based on the simplest approaches to epistemic logic.

10 See Arruda (1989) for more information on these systems and their applications.

11 See Kleene (1952) for an equivalent set of valuations, though Kleene does not treat the third ("paradoxical") value as designated.

12 See Priest (1995).
13 See Dunn (1986).
14 Batens (1998: 447).
15 For reasons of space the details must go unexplored here; see Batens (1998).
16 See Kyburg (1970).
17 See Schotch and Jennings (1989); Apostoli and Brown (1995).
18 See Belnap (1977) and Priest (1988).
19 Priest et al. (1989: 158).
20 Slater (1995).

References

Anderson, A. and Belnap, N. (ed.) (1975) *Entailment: The Logic of Relevance and Necessity*. Princeton, NJ: Princeton University Press.

Apostoli, L. and Brown, B. (1995) A solution to the completeness problem for weakly aggregative modal logics. *Journal of Symbolic Logic*, 60, 832–42.

Arruda, M. (1989) Aspects of the historical development of paraconsistent Logic. In G. Priest, R. Routley and J. Norman (eds.), *Paraconsistent Logic* (pp. 99–130). Munich: Philosophia Verlag.

Batens, D. (2000) A survey of inconsistency-adaptive logics. In Batens, Mortenson, Priest and Van Bendegem (eds.), *Frontiers of Paraconsistent Logic*. Baldock, Hertfordshire: Research Studies Press.

Batens, D. (1998) Inconsistency adaptive logics. In E. Orlowska (ed.), *Logic at Work: Essays Dedicated to the Memory of Helena Rasiowa*. Heidelberg: Springer.

Batens, D. (1989) Dynamic dialectical logics. In G. Priest, R. Routley and J. Norman (eds.), *Paraconsistent Logic*, Munich: Philosophia Verlag.

Batens, D., Mortenson, C., Priest, G., and Van Bendegem (eds.) (2000) *Frontiers of Paraconsistent Logic*. Baldock, Hertfordshire: Research Studies Press.

Belnap, N. D. (1977) How a computer should think. In Ryle (ed.), *Contemporary Aspects of Philosophy*. Oxford: Oriel Press; and also in Priest (1988) *Beyond Consistency*. Munich: Philosophia Verlag.

Béziau (2000) What is paraconsistent logic? In Batens, Mortenson, Priest and Van Bendegem (eds.), *Frontiers of Paraconsistent Logic*. Baldock, Hertfordshire: Research Studies Press.

Bohr (1913) On the constitution of atoms and molecules. *Philosophical Magazine*, 6, 1–25.

Brandom, D. and Rescher, N. (1980) *The Logic of Inconsistency*. Oxford: Basil Blackwell.

Braybrooke, P., Brown, B., and Schotch, P. K. (1995) *Logic on the Track of Social Change*. Oxford: Oxford University Press.

Brown, B. (1989) How to be realistic about inconsistency in science. *Studies in the History and Philosophy of Science*, 21, 2.

Brown, B. (1992) Rational inconsistency and reasoning. *Informal Logic*, XIV, 5–10.

Brown, B. (1993) Old quantum theory: a paraconsistent approach. *PSA 1992*, vol. 2, 397–411.

Brown, B. (1999) Yes, Virginia, there really are paraconsistent logics. *Journal of Philosophical Logic*, 28, 489–500.

Brown, B. (2000) Simple natural deduction for weakly aggregative paraconsistent logics. In Batens, Mortenson, Priest and Van Bendegem (eds.), *Frontiers of Paraconsistent Logic*. Baldock, Hertfordshire: Research Studies Press.

Brown, B. and Schotch, P. K. (1999) Logic and aggregation. *Journal of Philosophical Logic*, 28, 265–87.

da Costa, N. (1974) On the theory of inconsistent formal systems. *Notre Dame Journal of Formal Logic*, XV, 497–510.

da Costa, N. and Alves (1977) A semantical analysis of the calculi C_n. *Notre Dame Journal of Formal Logic*, XVIII, 621–30.

Dunn, J. M. (1986) Relevance logic. In Gabbay (ed.), *Handbook of Philosophical Logic*. Dordrecht: Reidel.

Jaskowski, M. (1969) Propositional calculus for contradictory deductive systems. *Studia Logica*, 24, 143–57 (originally published in Polish in 1948, *Studia Scientarium Torunensis*, Sec. A II, 55–77).

Jennings, R. E. and Schotch, P. K. (1981) Some remarks on (weakly) weak modal logics. *Notre Dame Journal of Formal Logic*, 22, 309–14.

Jennings, R. E. and Johnston, D. (1983) Paradox-tolerant logic. *Logique et Analyse*, 26, 291–308.

Jennings, R. E. and Schotch, P. K. (1984) The preservation of coherence. *Studia Logica*, 43, 89–106.

Jennings, R. E., Schotch, P. K. and Johnston, D. (1980) Universal first order definability in modal logic. *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik*, 26, 327–30.

Jennings, R. E., Schotch, P. K. and Johnston, D. (1981) The n-adic first order undefinability of the Geach formula. *Notre Dame Journal of Formal Logic*, 22, 375–78.

Johnston, D. (1978) A generalized relational semantics for modal logic, MA thesis, Department of Philosophy, Simon Fraser University, Barnaby, British Columbia.

Kleene, S. C. (1952) *Introduction to Metamathematics*. New York: Van Nostrand.

Kyburg, H. (1961) *Probability and the Logic of Rational Belief*. Middleton, CT: Wesleyan.

Kyburg, H. (1970) Conjunctivitis. In Marshall Swain (ed.), *Induction, Acceptance, and Rational Belief* (pp. 55–82). Dordrecht: Reidel.

Kyburg, H. (1997) The rule of adjunction and reasonable inference. *Journal of Philosophy*, XCIV, 109–25.

Kyburg, H. (1988) Full belief. *Theory and Decision*, 25, 137–62.

Priest, G. (2000) Motivations for paraconsistency: the slippery slope from classical logic to dialetheism. In Batens, Mortenson, Priest and Van Bendegem (eds.), *Frontiers of Paraconsistent Logic*. Baldock, Hertfordshire: Research Studies Press.

Priest, G. (1995) Reply to Parsons. *Canadian Journal of Philosophy*, 25.

Priest, G. (1988) *Beyond Consistency*. Munich: Philosophia Verlag.

Priest, G., Routley, R., and Norman, J. (eds.) (1989) *Paraconsistent Logic: Essays on the Inconsistent*. Munich: Philosophia Verlag.

Priest, G. and Tanaka (1998) Paraconsistent logic. In Stanford's online philosophy encyclopedia, http://plato.stanford.edu/entries/logic-paraconsistent/logic-paraconsistent.html.

Sarenac, D. (2000) The preservation of meta-valuational properties and the meta-valuational properties of implication. In Woods and Brown (eds.), *Logical Consequence: Rival Approaches*. London: Hermes.

Schotch, P. K. and Jennings, R. E. (1980a) Inference and necessity. *Journal of Philosophical Logic*, 9, 327–40.

Schotch, P. K. and Jennings, R. E. (1980b) Modal logic and the theory of modal aggregation. *Philosophia*, 9, 265–78.

Schotch, P. K. and Jennings, R. E. (1989) On Detonting. In G. Priest, R. Routley and J. Norman (eds.), *Paraconsistent Logic* (pp. 306–27). Munich: Philosophia Verlag.

Slater, B. H. (1995) Paraconsistent logics? *Journal of Philosophical Logic*, 24, 451–4.

Woods, J. (2000) Pluralism about logical consequence: resolving conflict in logical theory. In Woods and Brown (eds.), *Logical Consequence: Rival Approaches*. London: Hermes.

Woods, J. and Brown, B. (eds.) (2000) *Logical Consequence: Rival Approaches*. London: Hermes.

40

Logicians Setting Together Contradictories: A Perspective on Relevance, Paraconsistency, and Dialetheism

GRAHAM PRIEST

> You shall never be good logician, that would set together two contradictories: for that, the schoolmen say, God cannot do.
>
> (Thomas Cranmer, cited in the entry for "contradictory" in Little et al., 1973)

1 Introduction

There were doubtlessly many notable features of philosophy in the twentieth century. Perhaps we will have to wait for the perspective afforded by the passage of time to see clearly what they all were. But I think it true to say that one very notable feature is already visible. This is the final breaking of the taboo against inconsistency – the "superstitious dread and veneration in face of contradiction" as Wittgenstein put it (1978: 122). In Western philosophy, since Aristotle onwards, as the quotation from Archbishop Cranmer illustrates, inconsistency has been the ultimate 'no-no.' Accounts of truth, validity, rationality, have all taken it for granted. True, a few enterprising philosophical spirits, notably Hegel, have challenged the orthodoxy. But this was secure whilst its heartland in formal logic lay unchallenged. It is precisely this heartland that was challenged in the twentieth century, and which allowed the unthinkable to become thinkable. The challenge was laid down by paraconsistent formal logics. These logics allow for a discriminating handling of inconsistencies, not the crude 'contradictions entail everything' beloved by the latter-day friends of consistency.

This chapter is not about paraconsistent logics as such. There are many places where readers may go to find out technical details of these logics if they are not already familiar with them. The aim here is to provide a perspective on issues in the philosophy of logic that arise in connection with paraconsistency. This terrain is itself large, though. There is no hope of drawing a comprehensive map – even one of small scale. Rather, readers should regard this essay as a geographical sampler which will (with a bit of luck) encourage them to go and explore the terrain for themselves. Suggestions for further reading are scattered through the chapter at appropriate places. (Short accounts of paraconsistent logics can be found in the articles on paraconsistency in Craig (1999) and Zalta (1999). A much more thorough account can be found in Priest

(2000a), which may be consulted for nearly all the formal details alluded to in this chapter. Another good source of papers on paraconsistency in general is Priest et al. (1989).)

The perspective of the terrain that I will offer here turns around the notion of worlds, actual, possible, and impossible. This will put some order into affairs concerning paraconsistency and two closely connected, but distinct, notions: relevance and dialetheism. I cannot claim that this perspective is a neutral one. On the other hand, I hope that it is a bit more engaging than an account of the kind 'x says this, and y says that.'

Before we start with matters of more substance, let me define some of the crucial notions, so that we know what we are taking about. A propositional logic is *relevant* if, whenever $A \rightarrow B$ is a logical truth, A and B share a propositional parameter, where $\rightarrow$ is the conditional operator. A consequence relation, $\vdash$, is *paraconsistent* if the inference $A, \neg A \vdash B$ (for all A and B) fails. *Dialetheias* are truths of the form $A \wedge \neg A$; and *dialetheism* is the view that there are such things. Let us start with the first of these notions, relevance.

2 Relevant Logic

The thought that for a conditional, $A \rightarrow B$ ('if A then B'), to be true there must be some connection between antecedent and consequent is a very natural one. That is, the antecedent must, in some sense, be relevant to the consequent. The condition is not, of course, satisfied by the material conditionals of classical or intuitionist logic; nor does it appear to be satisfied by the strict conditionals of standard modal logics. In particular, let L be any logical falsehood; then the conditional $L \rightarrow B$ is both materially and strictly valid. Yet, for an arbitrary B, there would seem to be no connection between antecedent and consequent.

Providing an adequate analysis of the notion of the connection is another matter. Even for propositional logics, this is not straightforward. One well-known approach insists that for a conditional to be *really* logically valid it must be logically valid in a truth-preservational sense, and must also satisfy some extra condition of relevance. Thus, we might suggest, $A \rightarrow B$ is logically valid iff $A \rightarrow B$ is a classical tautology (that is, in every interpretation in which A is true, so is B) and, further, $R(A, B)$. Here, $R(A, B)$ is some suitable relationship; for example, that A and B share a propositional parameter. (For explorations of this idea, see the essays in Lehrer and Pollock 1979.)

The notion of conditionality that arises from this approach is a very tractable one, but the approach raises an obvious question. If a conditional is truth-preserving, why is it necessary to add some *extra* condition as well? After all, the whole *point* of a conditional is that its truth provides a guarantee that we can proceed from antecedent to consequent at will. What more than truth-preservation do you need?

A very different approach to relevant logic, that normally associated with the world-semantics of standard relevant logics, regards relevance not as something that should be tacked on to truth-preservation, but as something that falls out of a more adequate notion of truth-preservation. What is wrong with the conditional $L \rightarrow B$, for arbitrary B, is precisely that there are situations in which L holds, but where B does not. For example, let L be the claim that the Peano Arithmetic is complete. This is a logical

falsehood. Let B be the claim that Gödel proved that Peano Arithmetic is incomplete. Then the conditional $L \rightarrow B$ is false precisely because there are situations in which Peano Arithmetic is complete and (because of this, indeed) Gödel did not prove its incompleteness.

Situations like this are not logically possible situations. They are logically impossible: logic (and arithmetic) must be different at these worlds. The notion of a physically impossible situation will not raise an eyebrow in these enlightened times. We can all imagine situations where things can accelerate through the speed of light; Newton taught us what such situations might be like. But similarly, we can all imagine situations were the laws of logic are different. We all know what a situation would be like where the law of double negation fails; Brouwer taught us what such situations might be like.

In the simplified world-semantics of relevant logics, logically impossible worlds are normally called *non-normal*, or *irregular*. Their salient feature is that at such worlds conditionals have truth conditions different from those that they have at normal worlds. If w is a normal world, $A \rightarrow B$ is true at w if at all worlds where A is true, so is B. The simplest policy at non-normal worlds is to assign $A \rightarrow B$ an arbitrary truth-value. The rationale for this procedure is straightforward. It is precisely conditionals (or at least, conditionals of this kind) that represent laws of logic. Hence, they should behave differently at worlds where logic is different. How differently? There would seem to be no *a priori* bound on what is logically impossible. Hence a conditional might take on any value. Validity is defined in terms of truth-preservation at *normal* worlds. After all, we want to know what follows from what where logic *isn't* different. (For further details, see Priest 2000b: chapter 9.)

The semantical procedure just described gives a relevant logic. The truth conditions of conditionals at normal worlds are given in terms of truth preservation, but logically valid conditionals are relevant: if $A \rightarrow B$ is logically valid, A and B share a propositional parameter. And this arises because we take into our sweep logically impossible worlds.

The logic obtained in the way that I have described is, in fact, weaker than the logics in the standard family of relevant logics. The stronger logics of the standard family are obtained by evaluating conditionals at non-normal worlds slightly differently. Specifically, an interpretation is furnished with a ternary relation, R; and $A \rightarrow B$ is true at w iff for all x and y such that $Rwxy$, if A is true at x, B is true at y. (See Restall 1993 and Priest 2000b: chapter 10.) What the ternary relation *means* and why one might employ it in this way, is another matter, and one which is still philosophically *sub judice*. (See chapter 38, "Relevance Logic," of this volume for discussion, and for further references to relevant logic.)

Of course, the interpretations of a formal semantics are just abstract sets of certain kinds. They are not themselves the situations about which we reason. (Though we certainly can reason about situations concerning sets.) The sets *represent* situations. What, then, ontologically speaking, are the situations that they represent?

This is a thorny issue, but of a very familiar kind. There are many views concerning what possible worlds are. (See, e.g., the essays in the anthology of Loux 1979.) Some people, such as David Lewis, are realists about them: the worlds are exactly like the one in which we live, but with their own space, time, and causation. For others, such as Stalnaker, they are abstract objects of a certain kind, for example sets of propositions.

For yet others, such as Routley or Sylvan, they are nonexistent objects of a certain kind. The question of impossible worlds adds little, I think, to this debate. Whatever one takes possible worlds to be, impossible worlds are exactly the same kind of thing. Even if one is a realist about worlds, there is no reason, as far as I can see, why impossible worlds could not be of the same kind – worlds just like ours, with concrete individuals in a reality structured by its own space, time, causation, and now we add: logic. It is not even difficult to draw a picture of what such worlds may be like, the art of Maurits Esher often depicts situations where the logically impossible happens (such as geometric objects assuming configurations impossible in Euclidean space). (For a discussion of impossible worlds, see the essays in Detlefsen 1997.)

3 Paraconsistent Logic

The notion of negation, $\neg$, is an important one, and features in many important laws of logic. Negation is a contradictory-forming operator. That is, for any A, one of A and $\neg A$ must be true, and they cannot both be: $\Box(A \vee \neg A)$ and $\Box\neg(A \wedge \neg A)$. These are the logical laws of excluded middle and non-contradiction. Given these laws, in every possible world, $A \vee \neg A$ and $\neg(A \wedge \neg A)$ hold. There will be impossible worlds where, for any given A, $A \vee \neg A$ fails, or $A \wedge \neg A$ holds, though. For exactly this reason the conditionals $B \rightarrow (A \vee \neg A)$ and $(A \wedge \neg A) \rightarrow B$ may fail in relevant logics.

But let us look a little more closely at possible worlds. Given that disjunction behaves normally, the fact that $A \vee \neg A$ holds at such a world entails that either A or $\neg A$ holds. It might be thought that the fact that $\neg(A \wedge \neg A)$ holds at a world entails that one or other of A and $\neg A$ fails; but this does not necessarily follow, even given that conjunction behaves normally. Whether it does depends very much on the truth(-at-a-possible-world) conditions of negation. How negation functions is not at all obvious. In the history of philosophy, many such accounts have been given. According to some, contradictions entail nothing; according to others, contradictions entail everything; and according to yet others, contradictions entail some things but not others. Even in the 20th century, many different formal semantics for negation have been offered.

To see how it may be possible to have all of A, $\neg A$ and $\neg(A \wedge \neg A)$ holding at a world, consider the following very simple semantics. At every world, w:

$\neg A$ is true at w iff A is false at w
$\neg A$ is false at w iff A is true at w

Now, suppose that it is possible for A to be both true and false at a world. Then at that world, both A and $\neg A$ are true. Moreover, given the law of excluded middle, one of A and $\neg A$ is true; so one of A and $\neg A$ is false. Given that conjunction behaves normally, it follows that $A \wedge \neg A$ is false; and so $\neg(A \wedge \neg A)$ is true at the world as well.

Formal semantics where A may be both true and false are not difficult to construct. But it is natural to ask whether there really are possible worlds at which something may be both true and false. This is a fair question. I think it is also a fair answer that the best reasons for thinking this to be possible are also reasons for thinking it to be actual. So let us shelve this question for a moment.

If there are possible worlds at which A and $\neg A$ are true, and validity is defined in terms of truth-preservation at all normal worlds, then the inference $A \wedge \neg A \vdash B$ (*Explosion*) will fail. The notion of consequence delivered will therefore be paraconsistent. Relevant logics are not necessarily paraconsistent. For example, Ackermann's original relevant logic Π' was not. But relevant logics in the standard Anderson–Belnap family are. Conversely, many paraconsistent logics are not relevant (and may also employ a quite different treatment of negation); for example, the da Costa logic C_ω and its like are not.

Given that the inference *Explosion* fails in a logic, it follows that there may be inconsistent but non-trivial theories – that is, sets of sentences closed under logical consequence, which contain A and $\neg A$, for some A, but not every B. Such theories may not be candidates for the truth in any serious sense. They may, as it were, be descriptions of worlds that, though they are possible in a *logical* sense, are clearly very far from the actual world. Recall, after all, that even consistent worlds where frogs turn into people, and rich capitalists all give their money to the poor, are logically possible.

For all that, these theories may yet be important and interesting; and this is so for many reasons. For a start, such theories can be *mathematically* interesting. They may have a significant abstract structure which demands mathematical investigation, just as much as consistent ones do. (After all, one does not have to be an intuitionist to find intuitionist structures mathematically interesting.) Thus we have the rapidly developing study of inconsistent mathematical structures, a notable example of which are inconsistent arithmetics. (For an introduction to the whole area of inconsistent mathematics, see Mortensen 1995.)

Inconsistent theories may have physical importance too. An inconsistent theory, if the inconsistencies are quarantined, may yet have accurate empirical consequences in some domain. That is, its predictions in some observable realm may be highly accurate. If one is an instrumentalist, one needs no other justification for using the theory. And even if one is a realist, one may take the theory, though false, to be a significant *approximation* to the truth. This would seem to be how those who worked on early quantum mechanical models of the atom regarded the Bohr theory, for example. The theory was certainly inconsistent, as all agreed; yet its empirical predictions were spectacularly successful.

Finally, inconsistent theories may have practical importance too. This would be the case if our best understanding of how a piece of technology functions were provided by an inconsistent physical or mathematical theory of the kind we have just considered. Perhaps more importantly at the present, in information-processing of a kind that is now essential to everyone's life, there is always the possibility, indeed the high probability, of information that is inaccurate; inaccurate to the point of inconsistency. Where we discover that our information is inaccurate we will, of course, want to correct it. But on many occasions we may not know that it is inaccurate; nor may there even be a practical way of finding out. There is no algorithm, after all, for determining when information expressed in the language of first-order logic is inconsistent. In such circumstances, employing a paraconsistent logic is the only sensible strategy. We do not want our information-processor to tell us that the quickest way from Brisbane to Sydney is via New York, just because it has corrupt information about bus times in Moscow.

Before we leave the issue of paraconsistency as such, let us return to the Bohr Theory of the atom. A major reason why this was never regarded as a serious candidate for the truth was not so much that it was inconsistent as that it refused to allow inferences that were obviously truth-preserving, on pain of empirical inadequacy. In particular, it refused to allow the inference of adjunction: $A, B \vdash A \wedge B$. This was because the theory was chunked in a certain sense. The theoretical postulates were formed into certain groups (not necessarily disjoint). In computing the stationary states of the atom the quantum postulate was employed, but not Maxwell's electrodynamic axioms. In computing the results of transitions between the stationary states, Maxwell's axioms were employed. Within each chunk inference was allowed free reign. There was also a limited amount of information which was allowed to permeate between the chunks; but what one was not allowed to do was to take arbitrary information, A, from one chunk, and add it to another, containing the information B, and so infer the conjunction $A \wedge B$. (See Brown 1993.)

The chunking strategy is one that is employed in certain kinds of paraconsistent logics of the non-adjunctive variety. Specifically, given inconsistent premises including A and $\neg A$, one is not allowed to put these together in the same chunk to infer $A \wedge \neg A$, and so an arbitrary B, classical logic being the logic standardly in force in each chunk.

There are many ways of enforcing the chunking strategy, but the various details need not concern us here. I want merely to note that the strategy has no intrinsic connection with paraconsistency. For a start, there may be reasons for chunking information that have nothing, as such, to do with inconsistency. For example, one might chunk, not because failure to do so would lead to contradiction, but simply because failure to do so would lead to empirical inadequacy: false observational predictions. Or one may want to keep the information obtained from different sources in different chunks, not because the chunks may be mutually inconsistent (though they may be); but because information sources, such as witnesses, are notoriously unreliable. The fact that the same information occurs in different chunks speaks to its reliability, and is therefore itself a significant piece of information.

Moreover, and most importantly, there is no reason why the logic in force in each chunk must be classical logic. It could itself be a paraconsistent logic. For example, suppose that one of the sources of information was dialetheic, endorsing certain contradictions (though not all). In this case, to determine the proper content of that chunk, one would need a paraconsistent logic. Chunking strategies can, in fact, be employed with any kind of logic within the chunks – even with different logics within different chunks.

4 Dialetheism

Let us come back to worlds again. Someone may well hold that there are possible worlds that are inconsistent without holding that the *actual* world is. After all, the actual world is special. Truth at that world coincides with truth *simpliciter*. And truth has special properties all of its own. For example, one might well hold that for any A, $\neg A$ is true iff *A fails to be true*, whilst this is not true of worlds in general. The claim that the actual

world is inconsistent, though, is dialetheism. What reasons, then, are there for supposing that some contradictions are true?

There are many such reasons. (A number are discussed in Priest 1987.) Perhaps the best concerns the paradoxes of self-reference. One of the oldest, and most notorious, of these is the liar. This is a sentence, L, of the form $\neg T\langle L\rangle$, where T is the truth predicate, and angle brackets represent some naming device. The T-schema, $T\langle A\rangle \leftrightarrow A$ (for any sentence A), is an intuitively correct principle about truth. Substituting L in this gives $T\langle L\rangle \leftrightarrow \neg T\langle L\rangle$; and contradiction is but a few logical moves away.

The liar paradox and self-referential arguments of its kind, like Russell's paradox, are apparently sound arguments ending in contradiction. Of course, many other paradoxes are this too. But it is a striking fact about the paradoxes of self-reference that, though they have been the centre of so much philosophical attention for over 2000 years (at least the older ones), there is no consensus as to what, if anything, is wrong with them.

There are also reasons for supposing that the failure to solve the paradoxes is not simply a matter of lack of skill on the part of logicians. The paradoxes seem enormously robust. When steps are put forward to solve them, the contradictions concerned just seem to move elsewhere (in the shape of so called 'strengthened paradoxes'). It seems that contradiction is inherent in the various set-ups, and that all we can do is juggle it around. It is like those old-fashioned children's puzzles where one moves around pieces inside a frame, to try to achieve some predetermined pattern. Given a space in the frame, any adjacent piece may be moved into it. In this way, one can fill any given space; but filling it always creates another. There is always a space somewhere.

The appearance of the inevitability of contradictions is, I think, correct. The contradictions involved in the paradoxes of self-reference are, in a sense, inherent in thought. Our conceptual structures give us, at once, mechanisms for totalization and mechanisms that provide the ability to break out of any totality, such as diagonalization. The two mechanisms together produce contradiction. (This theme is explored at length in Priest 1995.) If this is the case, then certain contradictions are not only *actually* true, but, being inherent in thought, are *necessarily* true.

Of the other *prima facie* examples of dialetheias that one might cite, let us look at just one more. Boundaries are very puzzling things. They are almost contradictory objects by definition. For they both separate and join the areas of which they are the boundary. It is not, perhaps, surprising, then, that various kinds of boundaries seem to realize contradictions. Consider, for example, the boundary between the interior of a room (that which is in it) and the exterior (that which is not in it). If something is located on that boundary, is it in the room or not in it? Or suppose that a radioactive atom instantaneously and spontaneously decays. At the instant of decay, is the atom integral or is it not? In both of these cases, and others like them, the law of excluded middle tells us that it is one or the other. Yet the boundary is symmetrically placed with respect to each of its sides; so the only possibility that Reason countenances is a symmetric one. Thus, the object on the boundary of the room is both in it and not in it; and the atom at the point of decay is both integral and non-integral.

We see, then, there are reasons, at least *prima facie* reasons, for supposing that there are dialetheias. What reasons are there for holding such conclusions to be mistaken; that is, for holding that for no A are A and $\neg A$ both true?

The classical defense of this view is to be found in Aristotle's *Metaphysics*, Γ, 4; but this is hardly very successful. The major argument in the chapter is tangled and convoluted. It is not clear *how* it is meant to work, let alone *that* it works. The other arguments are short and similarly unsuccessful. Many of them do not even get to first base, since their conclusion is patently that it is not the case that *every* contradiction is true – or even that it is not possible to believe that every contradiction is true – things which are quite compatible with some contradictions being true. Moreover, I know of no way of reworking any of these arguments which makes them successful. (For a discussion of all of this, see Priest 1998.)

It is a singular fact that no philosopher since Aristotle has attempted a sustained defence of the view. What arguments are there? Here are a couple of notable ones. The first starts from the claim that for any statement to be meaningful, it must exclude something: it must say that we are in *this* situation, rather than *that*. But, the argument continues, the negation of a sentence holds in exactly those situations that the sentence does not hold in. Hence, we cannot have both A and $\neg A$ holding at any situation, and in particular, in the actual situation.

The argument appeals to a contentious theory of negation, one that a paraconsistent logician is likely to dispute. But let us suppose, for the sake of the argument, that the theory can be substantiated. The argument still fails. The claim that a meaningful sentence must exclude something, the other of its major premises, is precisely not available to classical logicians. For according to them, all necessary truths hold at *all* worlds. In particular, given the account of negation, the claim that $\forall A \neg(A \wedge \neg A)$, since it holds in all worlds, is itself meaningless! Ironically, it is the broader, relevant/paraconsistent, perspective that can accommodate the view about meaning in question. For given that there are impossible worlds, all claims, even logical truths, fail at some world.

A second argument appeals to the fact that we never observe contradictory situations: we never see a person both sitting and not sitting; we never see a group of people in which there are both three and not three. (Even if contradictions arise at instantaneous transition states, being instantaneous, these are not observable.) So there is good reason to believe that contradictions are never true. The argument is an inductive one, which might be thought strange, since the conclusion is supposed to be a logical truth; but one can collect *a posteriori* evidence for *a priori* truths: for example, it is *a priori* true that if a is taller than b, and b is taller than c, then a is taller than c; and we can collect evidence for this by going around measuring lots of as, bs, and cs.

The argument is not just an inductive one, though: it is not a very good inductive one. For the crucial question is whether the sample from which we are inducing is, in fact, a typical one; and the observable realm is not very typical in many ways. This is one of the lessons of modern science. Unobservable realms, particularly the micro-realm, behave in a very strange way, events at one place instantaneously affecting events at others in remote locations. Indeed, it would sometimes (in the well-known two slit experiment) appear to be the case that particles behave in a contradictory fashion, going through two distinct slits simultaneously. The micro-realm is so different from the macro-realm that there is no reason to suppose that what holds of the second will hold of the first. *A fortiori* when we move away from empirical realms altogether, the realm of sets appears to be inconsistent. Why should the way that observable things behave tell us anything about this?

Giving arguments to the effect that A and $\neg A$ are *never* true together is clearly a difficult matter. Some have concluded that it is impossible: this fact is so basic that there is no way that one can argue for it at all – at least, without begging the question. Despite the fact that Aristotle did give arguments, this was, in fact, his view of the matter. Only the 'uneducated' would ask for a proof (*Metaphysics*, 1006^a5–7). Whether his own views were consistent on this matter we will leave Aristotle scholars to argue about! The two arguments we have just looked at show at least the *possibility* of mounting sensible arguments for the claim. And though the arguments do not work, they do not fail simply because they beg the question.

5 Boolean Negation

At this point, let us look at a more subtle objection to dialetheism. This starts by conceding that the truth-in-a-world conditions of negation may well be what they are claimed to be by a paraconsistent logician. But, it continues, we can characterize a connective, call it ∇, by *giving* it the classical truth conditions. For every world, w:

BN ∇A is true at w iff A is **not** true at w

(And maybe giving it appropriate falsity conditions too.) I have boldfaced the negation in the conditions so that we can keep track of it in what follows. Since ∇ has the truth conditions of classical negation, it satisfies all the inferential principles we associate with that connective. In this context, ∇ is usually called *Boolean negation*, and contrasted with some relevant/paraconsistent negation (*RP negation*). Whether or not it is $\neg$ or ∇ that expresses vernacular negation is now largely irrelevant. For what the classical logician *wished* to express by negation can be expressed by ∇.

Now it would certainly appear to be the case that we can characterize a connective with the truth conditions BN. The problem is in establishing that this connective really does have all the properties of classical negation. To establish, for example, that $A, \nabla A \vdash B$ we have to reason: (1) for any w, it is **not** the case that both A and ∇A hold at w; hence (2) for any world, w, if A and ∇A hold at w, so does B. But what is this **not**? If it is $\neg$, the last inference is clearly invalid. ($\neg C \vdash C \rightarrow D$ is not valid in any paraconsistent logic.) Suppose, then, that it is ∇. If ∇ satisfies all the properties of classical negation, then (2) is acceptable. But recall that we were precisely in the process of mounting an *argument* that it does have these properties. Such a claim therefore simply begs the question.

It is sometimes suggested that metatheoretic truth-conditions of the kind BN are always given employing classical logic – in which case the inference in question is valid. But metatheory is not necessarily classical. For example, intuitionistic metatheory of intuitionistic logic is well-known. (See, e.g. Dummett 1977: chapter 5.) And why, in the last instance, if you think that one particular logic is correct, should there be any significance to a metatheory for it couched in a different, and incorrect, logic?

For a paraconsistent logician, the connective whose truth conditions are given by BN is a perfectly sensible connective. It just doesn't satisfy the classical advertising hype that goes with it. Could we not, though, simply stipulate that ∇ is a connective whose

meaning is determined by the proof-theoretic rules of classical negation? In a gem of an article, Prior (1960) pointed out that one cannot simply lay down a set of rules and expect it to characterize a meaningful connective. Suppose that we try to extend our set of logical operators by adding a new binary connective, * (*tonk*), satisfying the rules $A \vdash A * B$ and $A * B \vdash B$. Then all hell breaks loose: we can infer everything. If * were a meaningful connective, its addition would not interfere with the pre-existing machinery. In particular, then, its addition would not allow us to infer any sentence not containing * that was not inferable before. In technical jargon, the extension by *tonk* would be *conservative*. The fact that the extension is not conservative shows, therefore, that *tonk* is not meaningful.

Now, in a similar way, suppose that ∇ were stipulated to satisfy all the inferential principles of classical logic. Then given machinery which includes the *T*-schema and self-reference, we could construct a sentence, L, of the form $\nabla T\langle L\rangle$, and all hell would break loose in the same way: everything could be inferred. Hence, its addition is not conservative; so no meaningful connective can satisfy all the principles of classical negation. (It does not follow that there are not operators that behave as classical negation does in limited contexts. In situations that are consistent $\neg$ behaves in exactly that way.)

Of course, the question of conservative extension is relative to what one is extending. In the argument concerning ∇, one is extending machinery that is broader than propositional or first-order logic. But to restrict one's logical machinery to just this is somewhat arbitrary. The truth predicate, governed by the *T*-schema, would seem to be just as much a logical constant as the identity predicate, governed by its usual axioms.

It may be something of a shock that Boolean negation is meaningless. But what is, and what is not, a meaningful specification is not a matter of self-evidence. Such questions are highly theory-laden. And a dialetheist about the paradoxes of self-reference lines up with an intuitionist on this front. For the intuitionist, too, Boolean negation is meaningless, though for quite different reasons. (For an intuitionist, it must be possible, in principle, to recognize the truth of any sentence. Sentences starting with a Boolean negation do not have this property.)

There is an illuminating argument to the effect that Boolean negation is indeed meaningful, which goes as follows. (A version of this can be found in Batens 1990.) It must be possible to deny something, that is, to indicate that one does not accept it. Even dialetheists, after all, need to show that they don't accept that $1 = 0$. Now, if $\neg A$ is compatible with A, then asserting $\neg A$ cannot constitute a denial. To deny A one must assert something that is incompatible with it; so Boolean negation must make sense. We need to assert something with this force in denying.

Now, denial is a certain kind of illocutory act, an act with a certain linguistic force. It conveys the information that the utterer does not accept the thing denied. Other kinds of linguistic force include: asserting, questioning, commanding. Since Frege, it has been common to hold that denying is not an act *sui generis*. To deny A is simply to assert its negation. But this cannot be right. For example, we all, from time to time, discover that our views are, unwittingly, inconsistent. A series of questions prompts us to assert both A and $\neg A$ for some A. Is the second assertion a denial of A? Not at all; it is conveying the information that one accepts that $\neg A$, not that one does not accept A. One does this as well.

Denial, then is a linguistic act *sui generis*. Moreover, from the fact that one can deny A, it does not follow that there is some operator on content, ∇, such that to deny A is to assert ∇A, any more than from the fact that one can command that A it follows that there is some operator on content, !, such that to command A is to assert $!A$. Linguistic force is an element of communication *over and above* content. Suppose I utter 'The door is open'; then depending on the context, this could be an assertion, a question, a command. Similarly, if I utter 'It is not the case that A', this could be an assertion of $\neg A$, a denial of A – or even a command, or an act with some other linguistic force. The question is simply one of whether the act is intended to convey the information that the speaker does not accept A, or something else. Denial, then, is a linguistic act, performed by dialetheist and non-dialetheist alike, which in no way presupposes the meaningfulness of Boolean negation. (For further discussion of the material in this section, and negation in general, see Priest 1999.)

6 The Logical Choice

The issues that we have been dealing with concern, either implicitly or explicitly, the question of what the correct logic is. And this raises the question: how do you decide this matter? How, for example, does one determine the correct truth-at-a-world-conditions for negation?

Some have thought that such questions are silly. Logical principles are *a priori* obvious. Those who deny them are uneducated or insane. Such a view could be held, however, only by someone largely ignorant of the history of logic. In the history of logic there are dozens of different accounts of how negation functions, of when a conditional is true, of what inferences are valid – and corresponding disputes. (For a good discussion, see Sylvan 2000.) Moreover, views that have been well-entrenched for centuries have been overturned. For hundreds of years, 'All As are Bs' was held to entail 'Some A's are B's,' though it is not now. It may well have been the case that some of these principles were thought to be obvious. What was obviously true to one person, may be obviously false to another.

Such questions must, then, be taken seriously. But how do you resolve disputes about the correctness of logical principles themselves? Such disputes are liable to invoke arguments of a form whose very validity is itself disputed.

In disputes that involve high-level and very abstract principles, such as disputes about logic, it is not to be expected that any individual and simple argument, even if its validity is agreed upon by both parties, will be decisive. Arguments of any complexity invoke sundry 'auxiliary assumptions,' which may always be questioned. One is always, therefore, looking at package deals – theoretical complexes that have to be evaluated as a whole. In the case of logic, the package is liable to spread beyond principles simply about validity. There is such an intimate connection between truth and validity, for example, that questions about the nature of truth are likely to be embroiled in the debate as well.

How does one assess such a complex, then? First of all, theories are always proposed to account for some phenomenon, to explain some data; and the first consideration is always how adequate an explanation is provided. In the case of logic, we have

intuitions about which inferences are valid and which aren't; which conditionals are true and which aren't; and so on. We must look to see how well the theory accounts for the data. If a theory gives a result that is at variance with them, this is not fatal, but at least we must be able to explain the incongruity. For example, in virtually all relevant and paraconsistent logics, the disjunctive syllogism ($A, \neg A \vee B \vdash B$) is invalid. If we have an intuition that the Syllogism is valid, or at least that it is correct to use it on certain occasions, we must explain why this is so. We may say, as many have said, that the Syllogism is acceptable provided that we are reasoning about a consistent domain – just as an intuitionist may apply the law of excluded middle provided that they are reasoning about finite domains.

Adequacy to the data is not, therefore, likely to be a definitive factor. We have to invoke other criteria. The question of what these criteria are leads to well-known debates in the philosophy of science. Possible candidates include the following: the less a theory invokes *ad hoc* hypotheses the better it is; the more it gives a unified account of its subject matter, the better it is; the more a theory leads to new conceptual developments (fruitfulness), the better it is. There may be many other criteria too. For example, the first two criteria I just mentioned fall under the banner of simplicity; there may be other criteria that fall under this banner too.

Is inconsistency a negative criterion? If the logic of the theoretical complex is explosive, then everything will follow, and this is going to play havoc with the adequacy of the theory to handle the data. So inconsistency is highly relevant. If a paraconsistent logic is used, though, this is not necessarily going to be the case. Is consistency, in this case, a *sui generis* criterion? Is it the case that a theory that is more consistent than another is *ipso facto* a better theory? This is a question that cannot be divorced from the *rationale* for epistemic criteria; and this is a notoriously difficult question. Why, for example should simplicity of any given kind be a positive criterion? If there is some reason for supposing that reality is, quite generally, very consistent – say some sort of transcendental argument – then inconsistency is clearly a negative criterion. If not, then perhaps not.

Let me illustrate some of the preceding points concerning theory-choice. Suppose, for example, that one is comparing classical logic and a paraconsistent logic, as providing accounts of validity for sentences concerning truth-functional operators. As I noted, one cannot simply close ones eyes to other things. The *T*-schema and the inferences that this permits also strike us as valid. If classical logic is correct (and self-reference is legitimate), then this cannot be so: triviality is only a few steps away. Hence, some account of truth must be given which explains away the *T*-schema. If one accepts an appropriate paraconsistent logic, however, one can endorse a natural and simple account of truth: truth just is that notion characterized by the *T*-schema. We must compare, therefore, a package deal concerning (at least) Logic + Truth. Now, most paraconsistent logics are more complex than classical logic – though perhaps not much more so in the simplest cases. But all consistent accounts of truth are enormously more complex than the natural account, involving infinite hierarchies, epicycles to avoid strengthened paradoxes, and so on. What, then, is the simplest overall package? I leave you to judge.

The preceding discussion of theory-choice is, of course, quite general. Though I have couched it in terms of choice of logic, it applies just as much to a choice of any other

kind of theory. In particular, it shows how it may be rational to accept an inconsistent theory. (Paraconsistent logic plus the *T*-schema and self-reference is, indeed, inconsistent.) Even if inconsistency is a negative criterion, simplicity and consistency may well pull in opposite directions; a high degree of simplicity may outweigh a low degree of inconsistency. The discussion also shows something else. It is often claimed that if it could be rational to accept a contradiction, a person could never be forced, rationally, to give up any view. For there is nothing to stop the person accepting both their original view and the objection put to it, which is inconsistent with it. It is clear now that this objection fails. It is rational to give up a theory if there is a better one. And even if one can rationally accept an inconsistent theory (or theory plus objection) this may be trumped by a position that is simpler or has greater epistemic virtue of some other kind.

7 Conclusion

God, according to Cranmer in the quote with which we started, cannot set two contradictories together. Cranmer, Archbishop though he was, sold God short (though it was not this for which he was burned at the stake): contradictories can be set together by much lesser creatures. In the last 60 years, logicians have been setting them together in many ways. They may set them together in impossible worlds, to give relevant logics, logics which provide accounts of the conditional which make other accounts look crude and undiscriminating. They may set them together in possible worlds, to provide paraconsistent logics, logics which allow for the sensible handling of inconsistent information and theories. Or if they are daring, they may set contradictories together in the actual world, to allow for things such as a simple and natural theory of truth. These developments in logic, like all interesting new developments, are contentious. And no doubt the issues flagged in this essay will continue to be debated in the foreseeable future. So will many related questions: for the logical views that we have been discussing have implications that spread through metaphysics, epistemology, and many other areas of philosophy. One may presently only speculate as to what lands there are on the far side of the terrain I have been mapping.

References

Batens, D. (1990) Against global paraconsistency. *Studies in Soviet Thought*, 39, 209–29.

Brown, B. (1993) Old quantum theory: a paraconsistent approach. *Proceedings of the Philosophy of Science Association*, 2, 397–441.

Craig, E. (ed.) (1999) *Routledge Encyclopedia of Philosophy*. London: Routledge.

Detlefsen, M. (ed.) (1997) *Notre Dame Journal of Formal Logic*. 38, no. 4 (a special issue on impossible worlds).

Dummett, M. (1977) *Elements of Intuitionism*. Oxford: Oxford University Press.

Lehrer, K. and Pollock, J. (1979) *Philosophical Studies*, 26, no. 2 (a special issue on relatedness logics).

Little, W., et al. (1973) *The Shorter Oxford English Dictionary*, 3rd edn. Oxford: Oxford University Press.

Loux, M. J. (1979) *The Possible and the Actual: Readings in the Metaphysics of Modality*. Ithaca, NY: Cornell University Press.
Mortensen, C. (1995) *Inconsistent Mathematics*. Dordrecht: Kluwer Academic.
Priest, G. (1987) *In Contradiction: A Study of the Transconsistent*. The Hague: Martinus Nijhoff.
Priest, G. (1995) *Beyond the Limits of Thought*. Cambridge: Cambridge University Press.
Priest, G. (1998) To be and not to be – that is the answer. On Aristotle on the law of non-contradiction. *Philosophiegeschichte und Logische Analyse*, 1, 91–130.
Priest, G. (1999) What not? A defence of a dialetheic theory of negation. In D. Gabbay and H. Wansing (eds.), *What is Negation?* (pp. 101–20). Dordrecht: Kluwer Academic.
Priest, G. (2000a) Paraconsistent logic. In D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, 2nd edn. Dordrecht: Kluwer Academic.
Priest, G. (2000b) *Introduction to Non-classical Logic*. Cambridge: Cambridge University Press.
Priest, G., Routley, R. and Norman, J. (eds.) (1989) *Paraconsistent Logic: Essays on the Inconsistent*. Munich: Philosophia Verlag.
Prior, A. (1960) The runabout inference ticket. *Analysis*, 21, 38–9.
Restall, G. (1993) Simplified semantics for relevant logics (and some of their rivals). *Journal of Philosophical Logic*, 22, 481–511.
Sylvan, R. (2000) A preliminary western history of sociative logics. In D. Hyde and G. Priest (eds.), *Sociative Logics and their Applications* (ch. 5). Aldershot: Ashgate Publishers.
Wittgenstein, L. (1978) *Remarks on the Foundations of Mathematics*, 3rd edn. Oxford: Basil Blackwell.
Zalta, E. (ed.) (1999) Stanford internet encyclopedia of philosophy. http://plato.stanford.edu/contents.html.

Part XIII

LOGIC, MACHINE THEORY, AND COGNITIVE SCIENCE

This page intentionally left blank

41

The Logical and the Physical

ANDREW W. HODGES

This chapter on logic, machine theory and cognitive science will focus on the work of A. M. Turing (1912–54), who first related these topics, and whose ideas still dominate the interconnections after 50 years. In particular Turing pioneered the links between abstract logic and physical mechanisms. I shall here deal largely with a misconception which has recently found a wide circulation, but in the process I hope to shed light on Turing's foundations and also suggest a more positive point of new interest.

I will start with the scene in Bletchley Park, Buckinghamshire, England, in 1941. There Alan Turing is in charge of breaking the U-boat messages, as enciphered by the now-famous Enigma in its most fiendish variation. One of the many riddles wrapped inside the Enigma was that nine different code books were used to encipher what would now be called the session key. As revealed by his colleague I. J. Good (2000), Turing had devised a sophisticated statistical method for guessing which book had been used. The results of this work were then fed into the "Banburismus" method (Good, in Hinsley and Stripp 1993), based on giving numerical scores to the weight of evidence from various coincidences of letters, giving a scientific replacement to intuitive guessing. The actual counting of letter coincidences was done by WRNS servicewomen holding up punched paper tapes.

This scene, reminiscent of Searle's Chinese Room where people act out algorithms, is where Alan Turing was at a time when (as described in Hodges 1983: 214), he was first pondering questions of artificial intelligence through chess-playing algorithms. In 1941 he was also reading a book (Sayers 1941) by the detective novelist and religious writer Dorothy L. Sayers with reflections on determinism and creativity which, it seems, struck a peculiar chord in him. The question he apparently asked, judging by the comments on this book in Turing (1948), was whether these concepts were in fact exclusive as Dorothy L. Sayers supposed. Could not something purely *mechanical* nevertheless exhibit the features of *originality*? Perhaps there was a special meaning for Turing, whose mathematical methods were replacing the inspired guesswork of pre-scientific codebreaking. But to see the full point of this question, we must return to an earlier year. Alan Turing had made his name in 1936 with the concept of 'a machine', and we must see what he meant by it.

Turing's great paper "On Computable Numbers, with an application to the *Entscheidungsproblem*" (Turing 1936–7) was, as its awkward title said, intended to

give a definite meaning to, and hence derive a definite conclusion to, Hilbert's *Entscheidungsproblem*. To do this, Turing had to characterize the most general process that a human mathematician might carry out as a definite method. This the Turing machine definition supplied. Church (1937a) characterized it thus: "a computing machine, occupying a finite space and with working parts of finite size." More technically, a Turing machine is limited to finitely many configurations (or states) and finitely many types of symbols, with only finitely many symbols written on a tape at any time. The extent of the tape used however, is unlimited. Thus the Turing machine embodies finite means, but unlimited time and space to work out their consequences.

This finiteness deserves attention, and an interesting aspect of Turing (1936–7) is his claim that we may assume only finitely many states of mind. Indeed it is a striking claim that unobservable 'states of mind' should be countable, let alone finite. The philosopher B. J. Copeland has contested the signficance of this finiteness restriction (Copeland 1992: 280), suggesting that Turing said a Turing machine could simulate a device with infinitely many states. But a 'machine' with infinitely many states could encode the answers to every mathematical question, thus rendering trivial the very problem that the 'machine' concept was intended to settle. Thus the finite-state restriction is *crucial*, as is the restriction to a finite alphabet of symbols. (The same goes for allowing infinitely many symbols to be printed on the tape, or roll of toilet paper as Copeland (1992) vividly describes it, at a finite time.)

Did Turing ever consider infinite-state machines? 'Calculable by finite means' is how Turing characterizes the mechanical, and this rules out an infinite-state machine, with infinite means. The *very significance* of algorithms is that they encode potentially *infinite* outputs by *finite* specifications. Extending them to allow drawing on an infinite store of data would miss the point of what 'calculation' involves.

In mathematics, Turing's non-trivial discovery was that defining a real number in a finite number of words is not the same as being able to calculate it effectively. Turing's work has great significance outside mathematics. In computer science, the vital thing is that Turing's universal machine and its mode of operation can be implemented in electronics. In cognitive science, Turing's interpretation of states of mind developed into a thesis of the computability of mental operations. But in 1936–9, Turing expressed his work mainly in terms of how it affected Church's thesis, which as a result is now often described as the Church–Turing thesis: that anything effectively calculable is computable by Turing machines.

Turing gave this statement a definitive form in his Ph.D. thesis, submitted at Princeton in 1938, his supervisor being Church. This was later published as Turing (1939). Turing's formulation was:

> A function is said to be 'effectively calculable' if its values can be found by some purely mechanical process. Although it is fairly easy to get an intuitive grasp of this idea, it is nevertheless desirable to have some more definite, mathematically expressible definition. Such a definition was first given by Gödel at Princeton in 1934. . . . These functions were described as 'general recursive' by Gödel. . . . Another definition of effective calculability has been given by Church . . . who identifies it with λ-definability. The author [i.e. Turing] has recently suggested a definition corresponding more closely to the intuitive idea. . . . It was stated above that 'a function is effectively calculable if its values can be found by a purely mechanical process.' We may take this statement literally, understanding by a

> purely mechanical process one which could be carried out by a machine. It is possible to give a mathematical description, in a certain normal form, of the structures of these machines. The development of these ideas leads to the author's definition of a computable function, and to an identification of computability with effective calculability. It is not difficult, though somewhat laborious, to prove that these three definitions are equivalent. (166)

Thus, as Church and Gödel endorsed, the Turing machine definition, while equivalent in mathematical scope to lambda-calculus and recursive function theory, offers a convincing argument for why Church's thesis should be accepted. Turing's paper-tape definition also suggests, in a manner hard to define precisely, operations that can physically be *done*. The very word 'effect' means to *do* as opposed to postulate. Hence people sometimes distinguish Turing's thesis from Church's, though Turing himself never did this, and I shall continue to refer to this 1938 statement as the Church–Turing thesis.

But what did Turing mean by 'mechanical' or 'machine'? It is noteworthy that Turing does not make any qualifications; he does not say "carried out by a machine of a certain type"; he says "carried out by a machine." Nor did he ever devote any paper to the subject of "what is a machine." Newman (1955) said Turing came to his definition by analyzing the notion of a computing machine. Clearly Turing did indeed seize upon the concept of 'acting mechanically' implicit in the axiomatic program, turning it into the more definite 'machine,' and finding fascination in the mechanical thereafter – indeed developing it into our dominant technology. But as (Gandy 1988) put it, Newman was "subtly misleading"; Turing's 1936–7 analysis was of a *human being* computing in a mechanical way.

At this point I turn to the widely published and vividly expressed view of B. J. Copeland on just this question of what Turing meant by 'a machine.' It is my duty to point out a difficulty in Copeland's position. First, it is surprising that in his exposition (Copeland 1997) of the Thesis, Copeland omits the definitive 1938–9 statement, instead citing an informal version from Turing (1948). However that is not the most serious point, which is that Copeland (1997) holds that Turing's use of 'machine' was always made in a carefully restricted manner, and judiciously so because of the possibility of machines with *greater power* than Turing machines. Copeland explains that "For among a machine's repertoire of atomic operations there may be those that no human being unaided by machinery can perform." Copeland does not actually assert this was Turing's reasoning, in 1936 or subsequently, but his reader might well assume that this is why Turing made the definition he did. In fact Turing never suggested anything of the sort. Turing's thought stands within the natural and classical position, which is the other way round: it is to investigate whether and how a machine could possibly encompass the apparently greater faculties of human minds.

Copeland's suggestion of Turing having superhuman machines in mind clearly derives from his reading of Turing (1939), which is not mentioned in Copeland (1997), but is much discussed elsewhere (Copeland 1998, 1999; Copeland and Proudfoot 1999). In this paper Turing, after stating the Church–Turing thesis as quoted earlier, defined "a new type of machine." Copeland and Proudfoot (1999) quote this phrase as one of "Turing's Forgotten Ideas in Computer Science," and make it their mission to rescue it from the 'obscurity' of mathematical logic. Before succumbing to their excitement, however, we must analyze what Turing meant.

To understand Turing's "new type of machine" one must see what Turing had already defined. A vital point is that in (Turing 1936–7) the word 'machine' appeared in the definition of *choice-machine*. Such 'machines' are by definition *not purely mechanical*, being defined so as to require the decisions of 'an external operator.' Machines working without human choices, 'purely mechanically,' Turing called *automatic* or *a*-machines. Automatic machines are what (following Church's lead) we call Turing machines, and which Turing himself later (1948) called Logical Computing Machines. Such *a-machines* define the scope of computability and are the subject of the Church–Turing thesis, for that is concerned with whether a calculation can be carried out 'purely mechanically.' (Likewise, in Artificial Intelligence, attention is focused on *a*-machines. We should not be impressed by an AI program that relied upon human input!) Thus Turing used the word 'machine' for entities which are only *partially mechanical*. Perhaps this was courting confusion, but he was, of course, in accord with common usage.

An *oracle-machine* is likewise a machine only in the sense of being *partially mechanical*. Its specification, as a mathematical definition, is given as something merely postulated, *not* as an effective procedure. It differs from an *a*-machine in that it has one state in which it does *not* behave mechanically. Instead, the *o*-machine's next step then depends on an imaginary 'oracle' which has the power of giving the answer to any number-theoretic problem; as Turing shows this is equivalent to being able to answer the halting question for any Turing machine. Of the oracle Turing says, crucially, *"We shall not go any further into the nature of this oracle except that it cannot be a machine."* Turing describes the oracle as performing 'non-mechanical' steps. Thus an *o*-machine has a non-mechanical element, just as a choice-machine does.

Unfortunately Copeland, on the strength of the phrase "new type of machine," gives the vivid impression that the *o*-machine should be conceived as a *purely mechanical* device. Copeland and Proudfoot (1999) have written of how "the impact on the field of computer science could be enormous" if there were found "some practicable way of implementing an oracle." But the oracle is not a machine, so the question of 'implementing' it arises no more in Turing's exposition than does the question of 'implementing' the external operator's choices called for by a choice-machine.

The reader may suspect that further discussion is needed of what Turing meant by "cannot be a machine." Did he not merely mean, that it cannot be a *Turing machine*. Indeed he surely did: for in its context, the word 'machine' should mean the same as in the preceding statement of the Church–Turing Thesis, quoted above. There, it means what we call a 'Turing machine.' But this is no restriction on the force of Turing's remark, since he gave no indication that there could be any *other* types of purely mechanical machine.

If Turing had in mind the possibility of more general types of pure machine, he would have written that the oracle "cannot be an *a*-machine" or "cannot be a machine of any kind so far considered," or some such. Far from it: making himself even more categorical, he wrote that the *nature* of the oracle is that it cannot be a machine. Had Turing written that the *oracle* was a new kind of machine, Copeland would have his case. But by saying that the *o-machine* is a new kind of machine, Turing meant merely that it is a new type of 'not purely mechanical' machine.

As corroboration, note that if Turing had in mind the possibility of purely-mechanical machines other than *a*-machines, his 1938–9 statement would have required words like: '*understanding by a purely mechanical process one which could be carried out by an automatic machine of the type defined*'. In contrast, his actual statements, like Church's, identify the *a*-machine with the *most general* process that could be called 'purely mechanical.' Thus if he was saying the oracle was not an *a*-machine, he was saying it was not mechanical in *any* sense.

As already noted, Copeland (1997) chooses not to cite the Church–Turing thesis from Turing (1939), and so omits to analyze the use of 'machine' in this definitive version of it. Unfortunately an error in a later paper has denied Copeland the opportunity to show how he reconciles his concept of the 'new type of machine' with Turing's statement about the oracle. For Copeland (1999) quotes Turing as saying: "Of the nature of [an] oracle we shall say nothing." This truncation omits the essential substance. Further, Copeland uses the expression 'black box' to introduce the oracle, and says it could be conceptualized as having a tape with an infinite amount of information on it, giving the misleading impression that such physical images are Turing's.

Copeland also notes a difficulty in reconciling his standpoint with the endorsement in Church (1937b) of Turing's definition: "To define effectiveness as computability by an arbitrary machine, subject to restrictions of finiteness, would seem to be an adequate representation of the ordinary notion." Copeland (1997) states that 'arbitrary' refers to the arbitrary technical aspects of the way Turing machines or equivalent definitions are made. But no one could have read Church's sentence in so contrived a sense. Copeland asserts also that Church meant 'machine' to refer only to a machine mimicking the human calculator. In fact, Church (1937a) characterized the Turing machine in notably more general terms than this, and his words were that "in particular" a human calculator working to explicit instructions "can be regarded as a kind of Turing machine."

A more convincing claim would be that 'finiteness' restrictions *ipso facto* assert the logical possibility of machines with *infinitely* many states or symbols. But as indicated above, Turing never thought of such entities as 'machines,' and indeed the question of 'oracles' gives further evidence for this. For if Turing *had* wished to contemplate the oracle as a new kind of 'machine,' he could readily have done so by allowing a 'machine' to have infinitely many states. But Turing gave no such interpretation or definition, then or later.

Summarizing, Copeland misleads through his description of the *o*-machine as a new kind of machine. The oracle is not a machine. So *o*-machines are not purely mechanical. In fact they are *not machines*, if by 'machine' we mean something that works *independently*. Indeed the *whole point* of an oracle-machine is that it models non-mechanical steps. Unfortunately Copeland's exposition of the *o*-machine all rests upon this elementary confusion.

The reader may now wonder what possible reason Turing had to introduce a non-mechanical 'oracle.' One answer lies in some very interesting pure mathematics. I owe Feferman (1988) for an authoritative review. First, it is important to note that Turing (1939) is not focused on the *o*-machine definition. The paper is about a deep mathematical question, the fact that knowledge of one uncomputable number only provides

one step in an infinite journey. The countable set of uncomputable numbers derived (computably) from this given one does not even scratch the surface of the (uncountable) totality of uncomputable numbers. This journey into the transfinite, subsequently formalized as 'relative computability,' is the subject of Turing's paper and its difficulty is what needed his genius; the oracle itself is trivial. Turing used the oracle concept to give a simple proof relating to the first step in this journey, but as Feferman points out, he could have proved it using even simpler cardinality arguments.

The very fact that the oracle is *not* necessary to his mathematics suggests that Turing may have had some extra-mathematical idea in mind when introducing it. There is indeed an interpretation which can be made from the context. The 1936–7 definition of computability arises from a human mind carrying out a rule; so one may well ask what Turing thought the mind was doing when *not* following a rule. The section in Turing (1939) headed "Interpretation of Ordinal Logics," tends to confirm this line of thought. It is about 'intuition' which as Turing explains, is how he considers the step of seeing the truth of a formally unprovable Gödel sentence; the whole point is that as Gödel showed, this step cannot be made 'mechanical.' Newman (1955) gives an interpretation of the oracle as the mathematician "having an idea" as distinct from "making mechanical use of a method."

This carries weight in view of Newman's unique cooperation with Turing in mathematical logic, and their wartime discussion of 'intuition,' but the interpretation needs care. The interpretative section of Turing's paper does not mention the 'oracle,' let alone *identify* the oracle with intuition, and this is probably for a good reason. The oracle does *far more* than any human being could: it knows the answer to all number-theoretic questions. And yet, as Turing must have known and Penrose (1994: 380) has emphasized, intuition, using a diagonal argument, can outdo an oracle-machine. Thus in another sense, when set against the mind, the oracle is *too weak*. We can only safely say that Turing's *general setting* for the introduction of uncomputable elements has to do with the role of the mind in apparently outdoing the mechanical.

Turing's one actual mention of an oracle in the substance of his paper refers to seeing "whether a given formula is a ordinal formula": the essential non-mechanical step that explains why repeated addition of new axioms cannot overcome the limitations of formal proof. I suspect that Turing's use of the lambda-calculus formalism, developed in his collaboration with and respect for Church, obscured what Turing worked out for himself more in the language of mechanical and non-mechanical steps, the latter corresponding to a mental act of 'seeing' intuitively.

Although Copeland (1999) correctly offers a mental context for the oracle, Copeland and Proudfoot (1999) suggest a very different technological picture to their popular readership. In a graphic illustration of a 'black box,' they suggest that "what Turing imagined" could be implemented as something like an electrical capacitor measured to infinite precision, and that one could solve a halting problem by reading off (say) the 8735439th binary place. Oracles, if discovered, might revolutionize computer science; and modern theorists of uncomputable physical effects are chided for not recognizing Turing's anticipation of their ideas. The words 'notional' and 'abstract,' are used to describe 'what Turing imagined,' but it is said that the oracle is abstract only in the same sense as is the Turing machine operation of scanning symbols, which clearly *can* be implemented.

Turing's oracle only amounts to a few lines of mathematical definition, so those expecting blueprints will be disappointed. It could be suggested, however, that physical black boxes with oracular properties might exist, even if Turing never aroused such a prospect himself. We must indeed distinguish between the historical question of what Turing entertained, and the scientific question of what actually is the case, and on the latter question research has certainly turned to more complex issues than Turing considered. The oracle, something 'fictional' in Turing's thought, as Penrose has described it, would be factual if in some way infinite data could be stored in a finite system. This question has stimulated investigations into the relationship between computability, continuous systems, and physical properties of universes real and imaginary. However, there is nothing in modern physics to suggest the crude 'black box' in Copeland and Proudfoot's illustration of 'what Turing imagined,' requiring measurements of unlimited precision! And nothing could be further from Turing's logical calculus, in which the whole point of the oracle is to investigate the structure of what can *not* be done mechanically. (One should note also that even if a 'black box' were found to emit an uncomputable sequence, in fact rather than in fiction, there is no reason to expect it to solve any identifiable logical problem.)

There is one physical context for discussing 'oracles,' which has both some historical pertinence and modern interest. This is not mentioned by Copeland and Proudfoot (1999) amidst their technological imagery, but it is alluded to by Copeland (1998, 1999): it is the physics of the *brain.* This is at least consistent with Turing's setting, that of the human mind appearing to outdo machines. This is also the context in which Penrose (1994: 379) introduces a discussion of oracles, in his discussion of how uncomputability might feature in an as yet unknown physical law. (See also Penrose (1996) for an introduction to Penrose's arguments about computability and Mind.)

Penrose might well ask how in 1938 Turing would have answered the question of how the brain performed these acts of intuition. This is not an ahistorical speculation because, although Turing never used the word 'brain' in his 1936–39 writing, in 1930–32 he had thought seriously of its physics, stimulated by A. S. Eddington's view that quantum mechanical physics removed the classic conflict of free-will with physical determinism (Hodges 1983: 63).

But lacking any trace of such a discussion in 1936–39, it appears that Turing then simply left open the question of how it can be that a physically embodied mind appears to do the non-mechanical 'intuition' involved in seeing the truth of Gödel statements. He would not have been alone; Gödel himself seems to have taken a view of mind as non-mechanical without trying to reconcile this with its embodiment in brains.

But now let us return to the Turing of 1941, who unlike Gödel has linked logic with the physical to astonishing effect. Turing's Enigma-breaking machines are demonstrating the power of algorithms, by following through logical implications. Machines have become practical, and aspects of guessing have become mechanical. And while defending human freedom with the machines his mind had made, Turing is reading *The Mind of the Maker* by Dorothy L. Sayers.

In Hodges (1997) I suggested that it was at this 1941 period that Turing concluded that the scope of the computable was *not* limited to processes where the mind follows a definite rule. Machines which modified their own rules of behavior would show features which had not been foreseen by anyone designing them. Such machines might

be said to learn, and perhaps to act with the appearance of intelligence. From 1941 onwards Turing began to speak of such ideas to his Bletchley Park colleagues, and also to use the word *brain*. Again, my guess is that having confronted the problem of how the physical brain could support the appearances of non-mechanical 'intuition,' Turing concluded that the function of the brain *was* that of a machine, but one so complex that it could have the appearance of not following any rule. From this point Turing apparently became gripped by the potential of computability, and of his own discovery that all computable operations can be implemented on a single, universal, machine.

Thus, in this view it was during the war that Turing formulated both his central contribution to cognitive science and the practical universal machine. In his 1945 vision, algorithms are enough to account for all mental activity, including the kind previously thought of as non-mechanical intuition; the universal machine is enough for all algorithms, and electronics make practical a universal machine. In 1945 Turing embarked on his own independent design of what he called a "practical universal computing machine." (For a new account of the origin of the modern electronic stored-program computer, which credits Turing with giving von Neumann the central idea, see Davis 2000.)

Although Turing promoted the practical benefits of a computer, it was more the prospect of using it to simulate the brain that engaged him from the start. Thus even in Turing (1946), his technical report, a statement about the prospect of a machine showing intelligence in chess-playing appears. It makes a reference to *mistakes* in chess-playing, which, as expanded in later papers, betrays Turing's concern for answering the problem of how minds can see the truth of Gödel sentences. His postwar argument is that humans make mistakes, machines make mistakes, they are on a par, and that once infallibility is off the agenda, the Gödel argument does not apply. (This is the same argument that 50 years later Davis (2000: 197) upholds, as against Penrose's argument that a non-mechanical 'intuition' of truth is inescapable.) These remarks are to my mind evidence of how the postwar Turing had to respond to the implications of Gödel's work and his own 1938 discussion of 'intuition.'

It as not long before Turing (1948) described the problem of 'intelligent machinery' as that of how to create machines with 'initiative.' This is not the same word as the 'intuition' of 1938 but has the same role as describing that what the mind does when not apparently following a rule. But in Turing's postwar thought, initiative does not need uncomputable steps; it is as computable as the 'mechanical processes' even though this is against one's expectations of 'machines' (it is in this paper that he quotes from Dorothy L. Sayers). However it is necessary to depart from computations that follow the programmer's explicit plan. To this purpose Turing sketched nets of logical elements which, as Ince (1993) put it, can be said to predict the neural network approach. The paper of Copeland and Proudfoot (1996) on this subject is perhaps their greatest achievement since it has stimulated new scientific work (Teuscher 2001; Webster 1999). The modern climate is more favorable than were the 1970s and 1980s to Turing's viewpoint, in which advanced programming and evolutionary networks were not perceived as distinct alternatives, both being avenues for research in machine intelligence.

We have now reached the question of Turing's *postwar* writing on the concept of 'machine,' for Turing gave a classification of machines in that same paper (Turing

1948). Reflecting his greater acquaintance with physical machinery, Turing widened the scope of 'machine' and distinguished Logical Computing Machines from *continuous* and from *active* machines (e.g. 'a bulldozer'). Naturally, oracles do not feature in this discussion, since oracles are not machines. And Turing suggests nothing about the possibility of superhuman machine operations.

Copeland (1997) does not refer to this 1948 discussion, surprisingly as it is the closest Turing came to an essay on 'what is a machine.' But Copeland has usefully cited a later discussion by Turing's student Robin Gandy, who undertook the kind of formal analysis of computing machines that Newman (1955) attributed to Turing. Gandy (1980) introduced Thesis M: *"What can be calculated by a machine is computable."* Gandy showed that computability followed under very general assumptions about a mechanical system; and that if these conditions were weakened, *anything* could be calculable. 'Thesis M' allows Copeland to formulate what he claims of Turing: the essence of Copeland (1997) is that there is no evidence that Turing subscribed to Thesis M. It is true, as Copeland points out, that many writers have given versions of the Church–Turing thesis varying from what Church or Turing actually said, particularly in making sweeping statements about physical systems. Nor can we say quite how Turing would have responded to Gandy's formulation. But Copeland (1999) errs in claiming that Turing's definition of oracles precluded him from believing Thesis M. The reason is simple: Thesis M concerns machines, and Turing's oracle is *not* a machine.

The discussion by Copeland (1997) also neglects to engage with the fact that the general force of Turing's postwar arguments is that computable operations always suffice. How could he have made this thrust if, all the time, he had secretly in mind the potential of 'machines' to perform uncomputable operations? Copeland cites the careful definition of computability from Turing (1950) but ignores the central claim in that paper that Mind, including its appreciation of sonnets and the rest, can be imitated by a computer, that is computable. Copeland's position would be consistent with Turing's assertion of the computability of Mind only if Turing had believed that whereas minds were limited to the computable, there was a possibility of machines *not* so limited. But there is no such component detectable in Turing's thought. Despite this, Turing's later papers are searched for clues that he was leaving room for *o*-machines. For example, Copeland (1997) warns that in Turing (1947), a statement about 'machine process' should be read carefully to refer only to *a*-machines. It should indeed, but this is simply because Turing was here distinguishing Turing machines from the differential analyser, a *continuous* machine. (Morever he was in this passage advocating the the digital computer as *superior* in performance to the analog machine.)

In Copeland (1999) another claim is made relating 'oracles' to *randomness.* This is done through a paper of Church (1940), which gives a careful definition of an infinite random sequence, entailing that it must be an uncomputable sequence. Copeland goes on to describe a random sequence as an oracle-machine output, then to claim that Turing himself made this identification. But Turing did not allude to any connection between randomness and the infinite data stores implicit in an 'oracle.' On the contrary, Turing (1948, 1950) asserted that pseudo-random (i.e. computable) sequences would suffice for random effects, a statement he could never have made if he thought the uncomputable played a role. These were hardly idle comments; he as well as anyone on the planet in 1950 knew the significance of pseudo-random sequences, having in

wartime extracted machine patterns from apparently random ciphertext; he had also addressed himself to the engineering of randomness in the Manchester computer. Besides, oracles were introduced by Turing to model not *randomness*, but a kind of infinite *knowledge*.

Turing gave scant analysis of 'randomness', but this very brevity and the rather cavalier treatment of underlying physics in Turing (1948, 1950) is telling. Had he ever had uncomputable effects in mind, he could not have been so terse. He treated the relationship of discrete to continuous systems in a similar way: in Turing (1948) merely asserting that the brain though continuous is effectively discrete, returning to this point again briefly in Turing (1950) to counter "the Argument from Continuity of the Nervous System."

After scouring Turing's later works in the hope of glimpsing an allusion to oracle-machines, Copeland (1999) concedes, "if Turing did think that *o*-machines other than partially random machines are physically possible, then perhaps he would have said as much, and he does not appear to have done so." Quite so. Copeland has not shown how he reconciles this conclusion with the statement in Copeland and Proudfoot (1998) that:

> Taking their cue from Turing's 1939 paper, a small but growing international group of researchers is interested in the possibility of constructing machines capable of computing more than the universal Turing machine . . . research in this direction could lead to the biggest change computing has seen since 1948. Hodges's Turing would regard their work as a search for the impossible. We suspect that the real Turing would think differently. (Copeland and Proudfoot 1998: 6)

In this account, Turing's imagined disposition towards 'constructing' oracle-machines is pronounced to be *so definite* that it is required to inform the general public that the standard view, as conveyed in Hodges (1997), is not "the real Turing."

Yet if Turing's use of 'machine' for the only partially mechanical has confused these writers, it surely must have perplexed others less eminent, and so perhaps it is as well that this confusion has emerged so publicly through Copeland's work. Copeland (1999) has also recently done the service of editing the radio broadcast made by Turing (1951), and in a preface, amidst a fruitless hunt for oracles, he correctly draws attention to a comment of Turing about the question of whether brains can be seen as machines. It is a remarkable comment which suggests that he had given more thought to physics since writing the 1948 and 1950 papers. Turing here describes the universal machine property, applying it to the brain, but says its applicability requires that the machine whose behavior is to be imitated "should be of the sort whose behaviour is in principle predictable by calculation. We certainly do not know how any such calculation should be done, and it was even argued by Sir Arthur Eddington that on account of the indeterminacy principle in quantum mechanics no such prediction is even theoretically possible."

I described in Hodges (1983: 441) how Turing in this passage harked back to his 1930–2 wonder about the physics of the brain. Now I would consider fuller analysis needed. It was Penrose's question about how Turing saw computability after 1936 that encouraged me to take Turing's 1938–9 reference to 'intuition' more seriously and suggest (Hodges 1997) that his thesis of computable mind probably came later, in about

1941. Now, again, prompted by Penrose's arguments, I take Turing's reference to quantum mechanics more seriously than in 1983, and see it as a link between his work on computability and the burst of work in physics just before his death.

There is nothing about oracles in Turing's 1951 sentences. (Note also that even though now expanding the concept of 'machine,' Turing still addresses the question of whether a machine can do as much as the mind, not Copeland's question of super-human machines.) The point is that Turing here characterized the nature of quantum physics as possibly *unpredictable in principle*. Now, much is (rightly) made of Turing's contact with von Neumann in connection with the origin of the digital computer, but it is less well-known that Turing's first contact with von Neumann's work came in 1933 from studying the mathematical foundations of quantum mechanics (Hodges 1983: 79). So, as Turing knew so early, it is the *reduction or measurement process* for which there is no prediction even in principle; the evolution of the wave-function by Schrödinger's equation is predictable.

It is unlikely that Turing here was suggesting Penrose's view of quantum mechanics. More probably, he was seeking to reformulate quantum mechanics as a predictable theory when in 1953–4 he pursued this interest in physics. He wrote to Gandy, partly perhaps in jest, "I'm trying to invent a new Quantum Mechanics but it won't really work. How about coming here next week and making it work for me?" (Turing 1953–4). He was apparently focusing on the problem of the 'watched pot paradox' of wave function reduction (Gandy 1954; Hodges 1983: 495). Nevertheless Turing did acknowledge that here lay a fundamental problem for anyone assuming the computability of brain function.

Turing's interest in state-reduction, and the lack of a rule for it, should not be confused with prospect of *quantum computation* as developed since the 1980s. Quantum computation does not cross the boundary of computability, and moreover depends on the *predictability* of unitary evolution. Yet the elementary applications of quantum computation, as applied in quantum cryptography, have already led to procedures depending on non-local effects which cannot usefully be formulated as classical algorithms. This is enough to show that logic and physics can no longer be kept apart. The interpretation of the Church–Turing Thesis must necessarily be influenced by this development. In 1983 I used *the logical and the physical* as an organizing principle for the life and work of Alan Turing. But perhaps I did not follow this principle far enough. If we look to Turing for a prophecy of developments beyond the Turing machine, our best bet lies in his hint that the full discussion of computability requires the as yet incompletely known laws of quantum mechanics.

It is notable that in his 1951 talk Turing also raised the question of interpreting Gödel's theorem, and with less assertiveness than in 1950 that the problem went away through 'mistakes.' Thus although the philosophical detective, B. J. Copeland, has handcuffed the wrong suspect, through mistaking the identity of the oracle, I agree with him that the case as regards Turing's thought in his last period is not entirely closed. Turing did not draw a connection between Gödel's theorem and quantum mechanics, as Penrose does, but he did point to just these areas as leaving open and awkward questions.

Turing probably took the name 'oracle' from Shaw's *Back to Methuselah* which he enjoyed seeing performed at Princeton. Shaw doubted that our current human span is

long enough to gain sufficient wisdom. Turing said, perhaps echoing this pessimism, that no individual can do very much in a life. But the collective mind allows us greater optimism. In 1900, Planck's quantum, and Hilbert's problem of the consistency of arithmetic, soon joined by Russell's paradox, were unrelated. A century of investigations into the logical and physical have led to quantum computing, a connection unimaginable in 1900. The next century may see more unexpected developments, though perhaps no individual (or machine) more surprising than Alan Turing.

References

Church, A. (1937a) Review of Turing, 1936–7. *Journal of Symbolic Logic*, 2, 42.

Church, A. (1937b) Review of Post 1936. *Journal of Symbolic Logic*, 2, 43.

Church, A. (1940) On the concept of a random sequence. *Bull. Amer. Math. Soc.*, 46, 130–5.

Copeland, B. J. (1992) *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

Copeland, B. J. (1997) The Church–Turing thesis, in E. N. Zalta (ed.), *Stanford Encyclopaedia of Philosophy*, http://plato.stanford.edu

Copeland, B. J. (1998) Turing's o-machines, Searle, Penrose and the brain. *Analysis*, 58.2, 128–38.

Copeland, B. J. (1999) A lecture and two radio broadcasts on machine intelligence by Alan Turing. In K. Furukawa, D. Michie and S. Muggleton (eds.), *Machine Intelligence*, 15. Oxford: Oxford University Press.

Copeland, B. J. and Proudfoot, D. (1996) On Alan Turing's anticipation of connectionism. *Synthese*, 108, 361–77.

Copeland, B. J. and Proudfoot, D. (1998) Enigma variations. *Times Literary Supplement* (London), July 3, 1998.

Copeland, B. J. and Proudfoot, D. (1999) Alan Turing's forgotten ideas in computer science. *Scientific American*, April.

Davis, M. (2000) *The Universal Computer*. New York: W. W. Norton.

Feferman, S. (1988) Turing in the Land of O(Z). In Herken (1988).

Gandy, R. O. (1954) Letter to M. H. A. Newman, in King's College archive, to appear in the remaining volume of *The Collected Works of A. M. Turing*, ed. R. O. Gandy and C. E. M. Yates.

Gandy, R. O. (1980) Principles of mechanisms. In J. Barwise, H. J. Keisler and K. Kunen (eds.), *The Kleene Symposium*. Amsterdam: North-Holland.

Gandy, R. O. (1988) The confluence of ideas in 1936. In Herken (1988).

Good, I. J. (2000) Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the Naval Enigma. *J. Stats. Comp. & Simul.*, 66: 101–11; and in the remaining volume of *The Collected Works of A. M. Turing*, ed. R. O. Gandy and C. E. M. Yates.

Herken, R. (ed.) (1988) *The Universal Turing Machine: A Half-Century Survey*. Berlin: Kammerer und Unverzagt; Oxford: Oxford University Press.

Hinsley, F. and Stripp, A. (eds.) (1993) *Codebreakers*. Oxford: Oxford University Press.

Hodges, A. (1983) *Alan Turing: The Enigma*. London: Burnett; New York: Simon & Schuster; new editions London: Vintage 1992; New York: Walker, 2000.

Hodges, A. (1997) *Turing: A Natural Philosopher*. London: Phoenix; New York: Routledge, 1999. Included in *The Great Philosophers* (eds. R. Monk and F. Raphael) London: Weidenfeld and Nicolson, 2000.

Ince, D. C. (1993) Preface to Turing (1948) in D. C. Ince (ed.), *Mechanical Intelligence*, vol. III of *The Collected Works of A. M. Turing*. Amsterdam: North-Holland, 1993.

Newman, M. H. A. (1955) *Alan M. Turing*, Biographical memoirs of the Royal Society.

Penrose, R. (1989) *The Emperor's New Mind*. Oxford, New York: Oxford University Press.
Penrose, R. (1994) *Shadows of the Mind*. Oxford, New York: Oxford University Press.
Penrose, R. (1996) Beyond the doubting of a shadow. *Psyche Electronic Journal*, http://psyche.csse.monash.edu.au/v2/psyche-2–23-penrose.html
Sayers, D. (1941) *The Mind of the Maker*. London: Methuen.
Teuscher, C. (2001) *Turing's Connectionism*. London: Springer-Verlag.
Turing, A. M. (1936–7) On computable numbers with an application to the *Entscheidungsproblem*. *Proc. London Maths. Soc.*, ser. 2, 42, 230–65; also in M. Davis, (ed.), *The Undecidable*. New York: Raven, 1965.
Turing, A. M. (1939) Systems of logic defined by ordinals. In M. Davis (ed.), *The Undecidable*. New York: Raven, 1965. (This was also Turing's 1938 Ph.D. thesis, Princeton University.)
Turing, A. M. (1946) Proposed Electronic Calculator, report for National Physical Laboratory, Teddington; published in B. E. Carpenter and R. W. Doran (eds.), *A. M. Turing's ACE Report of 1946 and Other Papers*, Cambridge, MA: MIT Press, 1986; and in D. C. Ince (ed.), *Mechanical Intelligence*, vol. III of *The Collected Works of A. M. Turing*. Amsterdam: North-Holland, 1993.
Turing, A. M. (1947) Lecture to the London Mathematical Society, typescript in King's College, Cambridge, available in Turing Digital Archive, http://www.turingarchive.org. Published in B. E. Carpenter and R. W. Doran (eds.), *A. M. Turing's ACE Report of 1946 and Other Papers*, Cambridge, MA: MIT Press, 1986; and in D. C. Ince (ed.), *Mechanical Intelligence*, vol. III of *The Collected Works of A. M. Turing*. Amsterdam: North-Holland, 1993.
Turing, A. M. (1948) Intelligent machinery, unpublished NPL report, typescript in Turing Archive, King's College, Cambridge, available in Turing Digital Archive, http://www.turingarchive.org. Published (ed. D Michie) in *Machine Intelligence* 7 (1969), and in D. C. Ince (ed.), *Mechanical Intelligence*, vol. III of *The Collected Works of A. M. Turing*. Amsterdam: North-Holland, 1993.
Turing, A. M. (1950) Computing machinery and intelligence. *Mind*, 49, 433–60.
Turing, A. M. (1951) BBC Radio talk, typescript in King's College, Cambridge, available in Turing Digital Archive, http://www.turingarchive.org. Published (ed. B. J. Copeland) in K. Furukawa, D. Michie and S. Muggleton (eds.), *Machine Intelligence* 15. Oxford: Oxford University Press, 1999.
Turing, A. M. (1953/4) Letter to R. O. Gandy, undated, marked by Gandy "1953 or 54." In Turing Archive, King's College, Cambridge, section D/4, available in Turing Digital Archive, http://www.turingarchive.org.
Webster, C. S. (1999) Unorganized machines and the brain. Available from http://home.clear.net.nz/pages/cw/unorganized.htm.

42

Modern Logic and its Role in the Study of Knowledge

PETER A. FLACH

Knowledge is at the heart of intelligent behavior. The ability to obtain, manipulate, and communicate knowledge, in explicit form, is what distinguishes humans from other animals. This suggests that any study of intelligent behavior, theoretical or experimental, would have the same starting point, namely a Science of Knowledge, which studies the basic forms of knowledge, its acquisition, and its processing.

Yet there does not seem to exist such a unified and mutually agreed science of knowledge. In ancient times philosophy, the 'love of knowledge,' would aim to fulfill this role of the Mother of all Sciences, but philosophy has since long lost its central place and has mostly fragmented into specialized sciences such as physics, biology, and mathematics. Computer science, a relatively young branch on the tree of knowledge, has some aspirations to be the science of knowledge, but is currently at best a loosely connected collection of engineering technologies and abstract mathematical theory. (In fact, scholars of more established disciplines such as physics or chemistry often hesitate to call computer science a science at all, because its design-oriented approach does not fit in well with the doctrines of experimental sciences.) Artificial intelligence – the discipline studying fruitful connections between intelligent behavior and computers – would be another contender, but has been accused of overstating its claims, having unclear goals, and applying sloppy methodology.

In this chapter I argue that logic, in its widest sense, is – or at least, should be perceived as – the science of knowledge. This would be an unsurprising statement for a nineteenth-century logician, who would study the kind of inductive reasoning involved in experimental sciences as eagerly as he would investigate the kind of reasoning that is employed in mathematical proofs. However, in the last century logic seems to have developed into a relatively specialized and not seldomly obscure branch of mathematics. This is all the more paradoxical since the first half of the twentieth century has often been called 'the Golden Age of logic.' Following the pioneering work of Gottlob Frege, who developed a forerunner of predicate logic called *Begriffsschrift* ('concept language') in 1893, Russell and Whitehead published their three-volume *Principia Mathematica* between 1910 and 1913, in which they re-established the foundations of pure mathematics in logical terms. Whereas Kurt Gödel dealt a severe blow to the ambitions of logicians when he demonstrated that any logical system powerful enough to include natural numbers is also necessarily incomplete (i.e. the logical system allows

the formulation of true statements which are demonstrably unprovable within the system), this didn't stop logicians like Alonzo Church to develop ever more powerful logical systems (e.g. combinator logic and higher-order logic). Furthermore, Alfred Tarski invented what I consider one of the most important contributions of modern logic, namely the notion of an independent semantics.

1 The Key Ingredients of Logic

The duality between syntax and semantics is not only central in logic, it is also ubiquitous in linguistics and computer science. Syntax deals with the structure of a logical or linguistic expression in terms of its constituent symbols; semantics deals with mappings to objects capturing the meaning of expressions. Both are essentially about relationships between expressions, rather than about individual expressions. For instance, certain syntactic logical transformations produce new expressions from given ones by, for example, renaming or unifying variables. Semantics tells us under which conditions the syntactically modified expressions are equivalent to the original ones. Syntactic transformations can be chained together to form *derivations*, chains of expressions each of which is obtained from the previous one by one of the possible transformations. Semantics, on the other hand, is mostly concerned with the relation between the initial and final expressions. Syntax is more concerned with *how* to compute the final expression from the initial one, while semantics is more concerned with *what* the relation between them is. This what–how duality permeates all of computer science: from specification-design, via grammar-parser, to declarative-imperative programming.

A semantic relation with particular significance in mathematics and computer programming is the relation of *logical equivalence*, requiring that under no circumstance should a syntactic operation remove or add meaning. For instance, logically equivalent statements of a theorem (such as 'there exists no largest prime number' and 'there are infinitely many primes') are essentially seen as one and the same theorem. A related notion, and in fact far more useful, is the notion of *entailment* or *logical implication*. Two expressions are logically equivalent if, and only if, each entails the other. Many syntactic transformations produce weaker expressions that are entailed by, but not logically equivalent with, the original expression. For instance, we can *specialize* the expression 'there exists no largest prime number' to '4,220,851 is not the largest prime number.' Syntactic transformations which specialize expressions into weaker entailed expressions are called *sound* transformations. It may seem wasteful to throw away knowledge in this way, but logicians are often interested in *complete* sets of syntactic transformations which, when applied in every possible way, generate all possible implied expressions.

Soundness and completeness constitute the canon of mathematical logic. They allow us to reformulate mathematical knowledge into more manageable specializations about the particular topic we are interested in. They also allow us to combine several pieces of knowledge: for instance, from '4,220,851 is not the largest prime number' and '4,220,851 is a prime number' we can infer '4,220,851 is not the largest natural number.' Sound and complete transformations, or *inference rules* as they are often called,

are also central in many areas of computer science, for instance when we want to prove that a particular computer program meets its specification. In all these cases the starting point (the mathematical axioms, the grammar, or the program specification) is already, in an abstract sense, complete. If a mathematical theorem embodies knowledge that was not already present in the axioms we started from, the theorem is simply wrong. In mathematics the only allowed form of reasoning is sound reasoning or *deduction*.

2 Non-Deductive Reasoning Forms

In experimental sciences, and indeed in everyday life, the overwhelming majority of inferences is not deductive. Any physical theory that is to be of any use is expected to *generalize* the observations, in the sense that it makes predictions about as yet unobserved phenomena. If inference of such a theory from observations were required to be sound, no such predictions would be possible. Similarly, if our observations are insufficient to warrant a certain conclusion, we are usually happy to make educated guesses about the missing knowledge, even if this renders our inference, strictly speaking, unsound. The good news about giving up soundness is that our inferences may become much more useful; the bad news is that they may turn out to be wrong.

The fact that in science and everyday life non-deductive reasoning is ubiquitous suggests that we humans are relatively successful in avoiding most of the pitfalls of unsound reasoning, and that our non-deductive inferences are none the less correct most of the time. It follows that unsound reasoning comes in kinds – for instance, there is a trivial distinction between incorrect reasoning (such as inferring that all swans are black after observing 10 white swans) from unsound but potentially correct reasoning (such as inferring that all swans are white from the same observations). More interestingly, we would expect there to be different forms of unsound reasoning: one to deal with missing premises, one to propose a theory generalizing given observations, one for performing what-if analysis, one to explain observed behavior of a particular object, and so on. We would also expect to have some way to assess the reliability of an unsound inference, expressed in terms of, for example, the predictions it makes, the explanations it provides, the assumptions it requires, and the observations on which it was based.

There is a plethora of interesting research questions to explore. Which different kinds of unsound reasoning can be meaningfully distinguished? How different is each of them from deduction? Can we draw up a list of necessary and sufficient conditions for any kind of reasoning to be called deductive? Can we remove conditions from this list, and still obtain sensible but unsound forms of reasoning? Are soundness and completeness relative notions, for example does it make sense to talk about *inductive* soundness as distinct from deductive soundness? All these are issues one would expect to be central on most logicians' agendas. Yet, they seem to have fallen off during the 'Golden Age':

> The central process of reasoning studied by modern logicians is the accumulative deduction, usually explained semantically, as taking us from truths to further truths. But

> actually, this emphasis is the result of a historical contraction of the agenda for the field. Up to the 1930s, many logic textbooks still treated deduction, induction, confirmation, and various further forms of reasoning in a broader sense as part of the logical core curriculum. And moving back to the 19th century, authors like Mill or Peirce included various non-deductive modes of reasoning (induction, abduction) on a par with material that we would recognize at once as 'modern' concerns. Since these non-deductive styles of reasoning seemed irrelevant to foundational research in mathematics, they moved out quietly in the Golden Age of mathematical logic. But they do remain central to a logical understanding of ordinary human cognition. These days, this older broader agenda is coming back to life, mostly under the influence of Artificial Intelligence, but now pursued by more sophisticated techniques – made available, incidentally, by advances in mathematical logic. (van Benthem 2000)

To be sure: I am not arguing that logicians stopped investigating the research issues I indicated above – on the contrary, there have been many exciting developments regarding these questions, some of which will be covered in this chapter. However, they do seem to have disappeared from the main logical agenda. I believe it is important to revive the broader logical agenda, on which mathematical logic is an important subtopic but not the only one. If anything, such a broader agenda would stimulate cross-fertilization among subtopics, something which happens too seldom nowadays:

> Some members of the traditional logic community are still very conservative in the sense that they have not even accepted non-monotonic reasoning systems as logics yet. They believe that all this excitement is transient, temporarily generated by computer science and that it will fizzle out sooner or later. They believe that we will soon be back to the old research problems, such as how many non-isomorphic models does a theory have in some inaccessible cardinal or what is the ordinal of yet another subsystem of analysis. I think this is fine for mathematical logic but not for the logic of human reasoning. There is no conflict here between the new and the old, just further evolution of the subject. (Gabbay 1994: 368, note 7)

In the remainder of this chapter I will be considering the following fundamental question: which are the main forms of reasoning that make up the logical agenda, and what are their key characteristics? Informally, *reasoning* is the process of forming arguments, that is drawing conclusions from premises. By fixing the relation between premises and acceptable conclusions we can obtain various *reasoning forms*. For instance, an argument is *deductive* if the conclusion cannot be contradicted (or *defeated*) by new knowledge without contradicting the premises also; a form of reasoning is deductive if it only allows deductive arguments. We also say that deductive reasoning is *non-defeasible*. A *logical system*, or *logic* for short, is a particular formalization of a reasoning form. There may exist several logics formalizing a particular reasoning form; for instance, there is a range of deductive logics, such as modal, temporal, relevance, and intuitionistic logics, each formalizing certain aspects of deductive reasoning. These deductive logics do not necessarily agree on which arguments are deductively valid and which are not. For example, the argument 'two plus two equals four; therefore, if the moon is made of green cheese, then two plus two equals four' will be rejected by those who favor a causal or relevance interpretation of if–then rather than a truth-functional interpretation.

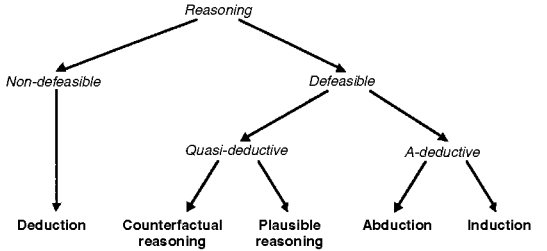


Figure 42.1 A classification of reasoning forms

However, as soon as such an argument is accepted as deductively valid, the only way to defeat the conclusion is by denying that two plus two equals four, and this defeats the premises also.

Non-deductive reasoning forms, on the other hand, are defeasible: a conclusion may be defeated by new knowledge, even if the premises on which the conclusion was based are not defeated. For instance, the argument 'birds typically fly; Tweety is a bird; therefore, Tweety flies' is non-deductive, since Tweety might be an ostrich, hence non-typical. The argument 'every day during my life the sun rose; I don't know of any trustworthy report of the sun not rising one day in the past; therefore, the sun will rise every future day' is non-deductive, since if the sun would not rise tomorrow, this would invalidate the conclusion but not the premises. The Tweety-argument is a well-known example of what I call *plausible reasoning*: reasoning with general cases and exceptions. An important observation is that plausible reasoning encompasses deductive reasoning: if we know that Tweety is a typical bird, the argument will be deductively valid. In this sense plausible reasoning is 'supra-deductive' or, as I will call it, *quasi-deductive*. Another example of quasi-deductive reasoning is so-called *counterfactual reasoning*, or 'what-if' analysis, starting from premises known to be false. For instance, the argument 'if you hadn't called me this morning, I would surely have missed my train' is a counterfactual argument, as both premise and conclusion are false in the intended interpretation. The point of such an argument is to investigate what would change if certain circumstances in the world had been different.

Other reasoning forms do not aim at approximating deduction, hence do not include deduction as a special case. I will call such reasoning forms *a-deductive*. The sunrise argument is an example of *induction*, an a-deductive reasoning form aimed at generalizing specific *observations* (also called evidence) into general rules or *hypotheses*. Note that I do not yet claim to have *defined* plausible or inductive reasoning in any way. Like with all forms of reasoning, this requires a formal definition of the consequence relation between premises and acceptable conclusions, analogous to deductive entailment. (The general term I will use for such a relation is *consequence*: thus, we will speak about

'inductive consequence' or 'plausible consequence,' and avoid potentially confusing terms like 'inductive validity' or 'plausible soundness').

Another form of a-deductive reasoning is *abduction*, a term originally introduced by C. S. Peirce to denote the process of forming an explanatory hypothesis given some observations (a hypothesis from which the observations can be deduced). For instance, the argument 'All the beans from this bag are white; these beans are white; therefore, these beans are from this bag' is an abductive argument. In recent years, abduction has become popular in the logic programming field, where it denotes a form of reasoning where the general explanation is known, but one of its premises is not known to be true; abduction is then seen as hypothesizing this missing premise. As a consequence, abduction and induction are viewed as complementary: induction infers the general rule, given that its premises and its conclusion hold in specific cases; abduction infers specific premises, given the general rule, and specific instances of its conclusion and some of its premises. Also, there are strong links between abduction and plausible reasoning: abduction can answer the question 'what do I need to assume about the bird Tweety if I want to infer that it flies' (answer: that it is a typical bird). I will expand on some of these issues below – the reader interested in finding out more about the relation between abduction and induction is referred to (Flach and Kakas 2000a).

The classification of reasoning forms I am advocating is depicted in Figure 42.1. While the justification for some of the distinctions made here have been admittedly sketchy, they will be elaborated in the rest of the chapter. The reader should also be aware that this classification should be taken as a starting point and is not intended to be set in stone. The main point is that on the map of logic, deduction occupies but a small part. I will now proceed to discuss some of these reasoning forms in more detail.

3 Plausible Reasoning

I should start by stressing that the term 'plausible reasoning' is not generally accepted – reasoning with exceptions is normally referred to as non-monotonic reasoning. *Monotonicity* is a technical term denoting that the set of conclusions grows (monotonically) with the set of premises. In other words, addition of a premise to a given argument never invalidates the conclusion – the same property as what I called non-defeasibility above. Since any non-deductive reasoning form is defeasible, it follows that any non-deductive reasoning form is non-monotonic. Thus, the property of non-monotonicity is of limited use in singling out a particular non-deductive reasoning form; for this reason I prefer a different (and more meaningful) term for reasoning with general rules and exceptions. (*Default reasoning* would be a good term, but this seems too strongly connected to a particular logic, i.e. default logic.)

Plausible reasoning is the process of 'tentatively inferring from given information rather more than is deductively implied' (Makinson 1994). It can thus be said to be more liberal or more credulous than deductive reasoning. Correspondingly, the set of arguments accepted by a plausible reasoning agent (also called a *consequence relation*, and defined as a subset of $L \times L$, where L is the language) can be divided into a deductive part and a plausible part. The deductive part corresponds to arguments not

involving any rules which have exceptions. (Alternatively, one can deductively extend a set of plausible arguments by treating all exceptions to rules as inconsistencies from which everything could be inferred, although this would be rather less interesting.)

The non-monotonicy of plausible reasoning can be demonstrated as follows: from *bird* one would infer *flies*, but from *bird and penguin* one wouldn't infer *flies*. That is, the rule *if bird then flies* is a *default rule* which tolerates exceptions, and the formula *bird and not flies* is not treated as an unsatisfiable formula from which anything can be inferred. The question then arises as to what other properties of deductive reasoning, besides monotonicity, are affected by allowing exceptions to rules. This is the main question addressed in a seminal paper by Kraus et al. (1990). In general, propositional deductive reasoning can be characterized by the following rules:

Reflexivity: $\alpha \vdash \alpha$ for all α;
Monotonicity: if $\alpha \vdash \beta$ and $\gamma \models \alpha$, then $\gamma \vdash \beta$;
Right Weakening: if $\alpha \vdash \beta$ and $\beta \models \gamma$, then $\alpha \vdash \gamma$;
Cut: if $\alpha \vdash \beta$ and $\alpha \wedge \beta \vdash \gamma$, then $\alpha \vdash \gamma$;
Left Or: if $\alpha \vdash \gamma$ and $\beta \vdash \gamma$, then $\alpha \vee \beta \vdash \gamma$.

In these rules, $\alpha \vdash \beta$ indicates that the reasoner in question accepts the inference from α to β, possibly with respect to an implicit body of background knowledge. $\models$, on the other hand, stands for classical deductive consequence (with respect to the same background knowledge). These rules can be combined: for instance, Reflexivity and Right Weakening together imply that $\alpha \vdash \gamma$ whenever $\alpha \models \gamma$, that is the consequence relation $\vdash$ is *supra-classical*.

Kraus et al. prove that the above five rules characterize deductive reasoning. Notice that equivalent rule sets exist: for instance, Cut could be replaced by Right And, and Left Or could be replaced by Right Implication:

Right And: if $\alpha \vdash \beta$ and $\alpha \vdash \gamma$, then $\alpha \vdash \beta \wedge \gamma$;
Right Implication: if $\alpha \wedge \beta \vdash \gamma$, then $\alpha \vdash \beta \rightarrow \gamma$.

Furthermore, they study the kinds of reasoning that result from weakening some of these rules. One variant they consider is obtained by replacing Monotonicity with the following two rules:

Left Logical Equivalence: if $\alpha \vdash \beta$ and $\models \alpha \leftrightarrow \gamma$, then $\gamma \vdash \beta$;
Cautious Monotonicity: if $\alpha \vdash \beta$ and $\alpha \vdash \gamma$, then $\alpha \wedge \beta \vdash \gamma$.

Both rules are clearly entailed by Monotonicity – Cautious Monotonicity, in particular, states that premises can be strengthened with their plausible consequences. This kind of plausible reasoning is called *preferential* reasoning, because it can be semantically modeled by assuming a (partial) preference order between states, where a state is a set of models, and stipulating that $\alpha \vdash \beta$ if and only if every *most preferred* state satisfying α also satisfies β (a state satisfies a formula iff all its models satisfy the formula). Preferential reasoning can be further weakened by dropping the condition that the preference relation between states be a partial order; this invalidates Left Or but none

of the other rules. This kind of reasoning is called *cumulative* reasoning because Cut and Cautious Monotonicity together imply that if $\alpha \vdash \beta$, then $\alpha \vdash \gamma$ if and only if $\alpha \wedge \beta \vdash \gamma$, that is plausible consequences can be accumulated in the premises.

From the foregoing it follows that deductive reasoners are preferential (with empty preference relation), and preferential reasoners are cumulative. However, a more meaningful comparison between reasoning forms X and Y would be obtained if we could establish, for each X-reasoner, a unique maximal subset of arguments that satisfy the rules of Y. Such a *reduction* from preferential to deductive reasoning was given in (Flach 1998). Basically, it involves using Monotonicity in the opposite direction (if $\gamma \models \alpha$ and $\gamma \not\models \beta$, then $\alpha \not\models \beta$) to remove arguments that are not deductively justified. Semantically, this amounts to ignoring the preference relation. (As stated before, we can also use Monotonicity in the forward direction to turn all plausible arguments into deductive ones, amounting to removing all states that satisfy exceptions to rules; however, this would rather endorse the less natural view that plausible reasoning is the process of inferring *less* than deductively implied.)

There is an interesting analogy between non-monotonic reasoning and non-Euclidean geometry. For many centuries it was assumed that Euclid's fifth axiom (parallel lines don't intersect) was self-evident, and that denying it would lead to inconsistencies. However, non-Euclidean geometry was proved to be consistent in the early nineteenth century. Similarly, many logicians argued that logic was necessarily monotonic, and that the concept of a non-monotonic logic was a contradiction in terms. However, there is a difference between monotonicity as a property of mathematical reasoning, and monotonicity of the logic under study. Kraus et al. used the deductive metalogic of consequence relations to formalize various forms of non-deductive reasoning. Rules such as Cautious Monotonicity are in fact *rationality postulates* that need to be satisfied by any rational reasoning agent of the class under study. This is a crucial insight, and their approach establishes a methodology that can be applied to analyze other forms of reasoning as well. This will be explored in the next section.

4 Induction and Abduction

Induction is the process of generalizing specific evidence into general rules. A simple form of induction is the following sample-to-population inference:

X percent of observed Fs are Gs;
therefore, (approximately) X percent of all Fs are Gs.

This argument schema has a categorical counterpart:

All of observed Fs are Gs;
therefore, all Fs are Gs.

or – since the induced rule need not be a material implication –

All objects in the sample satisfy $P(x)$;
therefore, all objects in the population satisfy $P(x)$.

These formulations of inductive generalization, however, obscure a crucial issue: normally, the predicate P to be used in the general rule is not explicitly given in the observations. Rather, the key step in induction is to distill, out of all the available information about the sample, the property that is common to all objects in the sample and that will generalize reliably to the population. I will refer to this step as *hypothesis generation*.

Hypothesis generation is an often ignored step in philosophy of science. For instance, in *Conjectures and Refutations* Popper describes at length how to test a conjecture, but remains silent about how to come up with a conjecture in the first place. To refer to ill-understood phenomena such as creativity in this context is to define the problem away. Moreover, if we want to automate scientific discovery or learning (object of study in the subfield of artificial intelligence called machine learning), we have to approach hypothesis generation in a principled way. Hypothesis generation is not a wholly irrational process, and the question thus becomes: what are the rationality postulates governing inductive hypothesis generation?

In fact, this question was already considered by the American philosopher Charles Sanders Peirce, who wrote in 1903:

> Long before I first classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis – which is just what abduction is – was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts or some of them. The form of inference, therefore, is this:
>
> The surprising fact, C, is observed;
>
> But if A were true, C would be a matter of course,
>
> Hence, there is reason to suspect that A is true.
>
> Thus, A cannot be abductively inferred, or if you prefer the expression, cannot be abductively conjectured until its entire content is already present in the premiss, "If A were true, C would be a matter of course." (Peirce 1958: 5.188–9)

Here, Peirce calls the process of explanatory hypothesis generation *abduction* (while he uses the less tentative phrase "adopting an explanatory hypothesis" above, elsewhere (5.171) he defines abduction as "the process of forming an explanatory hypothesis," i.e. "abduction merely suggests that something *may be*").

Nowadays people use the term 'abduction' in various senses (even Peirce had initially a different, syllogistic view of abduction), so a brief digression on these issues may be in order – the interested reader is referred to (Flach and Kakas 2000b) for a more extensive discussion. In philosophy, it is customary to view abduction as 'reasoning to the best explanation' (Lipton 1991). This, however, combines hypothesis generation with hypothesis selection, only the former being a purely logical process amenable to logical analysis. In artificial intelligence, abduction is usually perceived as reasoning from effects to causes, or from observations to explanations: here, an abductive hypothesis is not a general rule or theory, as in induction, but rather a specific explanation or cause relating to the observed individual. Thus, abductive hypotheses explain but do not generalize. Induction, on the other hand, aims at generalizing beyond the observed individuals. While in inductive argument schemas such as the above the induced hypotheses entails the observations, this is not an explanation in the same sense as a cause explains an effect.

In general, we cannot distinguish between abductive explanations and inductive generalisations by methods based on entailment alone, including the method I am about to describe. However, the view of induction as generalization does suggest an alternative formalization which is closer to both confirmation theory (in the qualitative sense of Hempel) and Kraus et al.'s (1990) account of plausible reasoning. In the remainder of this section I will discuss rationality postulates for explanatory reasoning, including abduction and explanatory induction, and then present alternative postulates for confirmatory induction in the next section.

Returning to Peirce, the logical form of abductive hypothesis generation he suggested can be simplified to 'from C, and $A \models C$, abduce A' or, introducing the symbol $\mid\!\sim$ for abductive inference, 'if $A \models C$, then $C \mid\!\sim A$'. We can now use Kraus et al.'s (1990) consequence relation methodology to formulate rationality postulates for hypothesis generation. We start with some general principles:

Verification: if $\alpha \mid\!\sim \beta$ and $\alpha \wedge \beta \models \gamma$, then $\alpha \wedge \gamma \mid\!\sim \beta$;
Falsification: if $\alpha \mid\!\sim \beta$ and $\alpha \wedge \beta \models \gamma$, then $\alpha \wedge \neg\gamma \not\mid\!\sim \beta$.

Verification and Falsification state that if β is a possible hypothesis given observations α, and γ is a prediction on the basis of β (and α), then β is not ruled out by observing that γ is true, but falsified by observing that γ is false. (While the names of these rules have been inspired by the debate between the logical positivists and Popper, it should be stressed that – under my interpretation of $\alpha \mid\!\sim \beta$ as 'β is a *possible* hypothesis given evidence α' – Verification is a fairly weak rule to which one can hardly object.)

Falsification is different from the rules we have seen until now, because it draws negative conclusions about the consequence relation $\mid\!\sim$. This means that some of Kraus et al.'s (1990) rules need to be adapted when formulated in our framework. For instance, the following set of 'explanatory' rules is obtained by rewriting the rules given in Section 3 for deduction, substituting $\beta \mid\!\sim \alpha$ for $\alpha \vdash \beta$ (we use the variant with Right And and Right Implication):

Reflexivity: $\alpha \mid\!\sim \alpha$ for all α;
Right Strengthening: if $\beta \mid\!\sim \gamma$ and $\alpha \models \gamma$, then $\beta \mid\!\sim \alpha$;
Left Weakening: if $\beta \mid\!\sim \alpha$ and $\beta \models \gamma$, then $\gamma \mid\!\sim \alpha$;
Left And: if $\beta \mid\!\sim \alpha$ and $\gamma \mid\!\sim \alpha$, then $\beta \wedge \gamma \mid\!\sim \alpha$;
Left Implication: if $\gamma \mid\!\sim \alpha \wedge \beta$, then $\beta \rightarrow \gamma \mid\!\sim \alpha$.

The last three rules make immediate sense for explanatory hypothesis generation. In particular, Left Weakening states that the set of explanations decreases monotonically when the observations increase; it is a convergence property for induction (it can be combined with Verification into a single rule). Left And states that if α is a possible hypothesis explaining β and γ observed separately, it also explains β and γ observed together; this enables incremental induction. Left Implication deals with background knowledge: if β is a necessary part of the explanation of γ, then it can also be added as a condition to the observation.

On the other hand, the first two rules contradict Falsification and need to be weakened by adding an admissibility requirement on α (for instance, that α explains some-

thing – e.g. itself). Without going into details, we mention that the following set of rules has been demonstrated to characterize consistent explanatory reasoning:

Explanatory Reflexivity: if $\beta \mid\!\sim \beta$ and $\neg\alpha \not\mid\!\sim \beta$, then $\alpha \mid\!\sim \alpha$;
Admissible Right Strengthening: if $\beta \mid\!\sim \gamma$, $\alpha \mid\!\sim \alpha$ and $\alpha \models \gamma$, then $\beta \mid\!\sim \alpha$;
Predictive Left Weakening: if $\beta \mid\!\sim \alpha$ and $\alpha\wedge\beta \models \gamma$, then $\gamma \mid\!\sim \alpha$;
Left And: if $\beta \mid\!\sim \alpha$ and $\gamma \mid\!\sim \alpha$, then $\beta\wedge\gamma \mid\!\sim \alpha$;
Left Implication: if γ **k** $\alpha\wedge\beta$, then $\beta\rightarrow\gamma \mid\!\sim \alpha$;
Left Consistency: if $\alpha \mid\!\sim \beta$ then $\neg\alpha \not\mid\!\sim \beta$.

While some of these postulates may be debatable (for instance, one may argue that explanatory reasoning is inherently irreflexive), they do provide a starting point for studying various forms of explanatory reasoning. Instead of a single 'logic of induction', I have proposed a modular system of meta-level rationality postulates that can be adapted to model various forms of reasoning. In addition, one can study semantic characterizations of these postulates. The interested reader is referred to (Flach 2000a, 2000b).

5 Confirmatory Induction

The preceding set of postulates concentrated on induction and abduction as explanatory reasoning. There is an alternative view of induction as inferring hypotheses that are *confirmed* by the observations. This view was pioneered by Carl G. Hempel, who proposed both a set of rationality postulates (or, as he called them, adequacy conditions) and a material definition of confirmation. The following is a list of Hempel's adequacy conditions (Hempel, 1945: 103–6, 110), reformulated in our meta-language:

Entailment: if $\alpha \models \beta$, then $\alpha \mid\!\sim \beta$;
Right Weakening: if $\alpha \mid\!\sim \beta$ and $\beta \models \gamma$, then $\alpha \mid\!\sim \gamma$;
Right And: if $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$, then $\alpha \mid\!\sim \beta\wedge\gamma$;
Consistency: if $\alpha \mid\!\sim \beta$ and $\alpha \not\models \neg\alpha$, then $\alpha \not\models \neg\beta$;
Left Logical Equivalence: if $\alpha \mid\!\sim \beta$ and $\models\alpha\leftrightarrow\gamma$, then $\gamma \mid\!\sim \beta$;

For instance, the first condition states that entailment 'might be referred to as the special case of conclusive confirmation' (Hempel 1945: 107). Each of these postulates is reasonable, except perhaps Right And which seems unjustified if the evidence is too weak to rule out incompatible hypotheses – in other words, it expresses a completeness assumption regarding the observations.

The main reason for Hempel to formulate his adequacy conditions was to verify his material definition of confirmation against them – consequently, there is no guarantee that they are complete in any sense. The following set of rationality postulates for confirmatory induction can be shown to be complete with respect to a suitably devised semantics:

Confirmatory Reflexivity: if $\beta \mid\!\sim \beta$ and $\beta \not\mid\!\sim \neg\alpha$, then α **k** α;
Predictive Right Weakening: if $\alpha \mid\!\sim \beta$ and $\alpha\wedge\beta \models \gamma$, then $\alpha \mid\!\sim \gamma$;

Right And: if $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$, then $\alpha \mid\!\sim \beta \wedge \gamma$;
Right Consistency: if $\alpha \mid\!\sim \beta$ then $\alpha \not\mid\!\sim \neg\beta$;
Left Logical Equivalence: if $\alpha \mid\!\sim \beta$ and $\models \alpha \leftrightarrow \gamma$, then $\gamma \mid\!\sim \beta$;
Strong Verification: if $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$, then $\alpha \wedge \gamma \mid\!\sim \beta$;
Left Or: if $\alpha \mid\!\sim \gamma$ and $\beta \mid\!\sim \gamma$, then $\alpha \vee \beta \mid\!\sim \gamma$.

As before, I disallow contradictory observations (unlike Hempel) – a weaker form of Entailment follows from Predictive Right Weakening, and a weak form of Reflexivity has been added as a separate rule (notice that Reflexivity was implied by Hempel's rules as an instance of Entailment). Two new rules have been added. Whereas Verification states that predictions γ can be added to confirming observations α for hypothesis β, Strong Verification states that this can also be done whenever γ is confirmed by α. As with Right And, the underlying assumption is that the observations are complete enough to have all confirmations 'point in the same direction.' Left Or can be seen as a variant of Left Weakening, discussed in the context of explanatory reasoning. While Left Weakening is clearly invalid in the confirmatory case (if we weaken the observations, there will presumably come a point where they cease to confirm the hypothesis), Left Or states that separate observations confirming a hypothesis can be weakened by taking their disjunction.

The semantics against which these postulates are provably complete is a variant of Kraus et al.'s (1990) preferential semantics for plausible reasoning. In fact, the postulates for confirmatory induction are closely related to postulates considered in Section 3: for instance, Strong Verification is identical with Cautious Monotonicity. This is perhaps surprising at first sight, but can be explained by noting that plausible and confirmatory reasoning make similar assumptions in order to go beyond deduction: while in plausible reasoning one commonly assumes that anything which is not known to be an exception conforms to the rule, in induction one assumes that unknown objects behave similarly to known objects.

We end this section on a philosophical note. Hempel's name is associated with a number of paradoxes, one of which is the *confirmation paradox*. This paradox arises when one considers to add a variant of Right Strengthening to the postulates for confirmatory induction. To borrow Hempel's example:

> Is it not true, for example, that those experimental findings which confirm Galileo's law, or Kepler's laws, are considered also as confirmation Newton's law of gravitation?' (Hempel 1945: 104)

The problem is that the combination of Right Weakening and Right Strengthening immediately leads to a collapse of the system, since arbitrary observations now confirm arbitrary hypotheses. However, Hempel confuses confirmation with explanation here. Explanatory hypotheses can be arbitrarily strengthened (as long as they remain consistent with the observations), but not necessarily weakened; confirmed hypotheses can be arbitrarily weakened, but only strengthened under certain conditions. It might be possible to formalize a form of hypothesis generation where hypotheses both explain and are confirmed by the observations (this is an open problem), but then there would be strong conditions on both strengthening and weakening of the

hypothesis. Distinguishing between explanatory and confirmatory induction solves the confirmation paradox.

6 Concluding Remarks

This short chapter has been written as a re-appraisal of logic as the science of knowledge. The goal of logic is to provide a catalogue of reasoning forms. Deduction is but one of the possible forms of reasoning, easiest to formalize but with limited importance for intelligence. It is possible to characterize non-deductive or defeasible reasoning forms mathematically, by concentrating on their purely logical part, *viz.* hypothesis generation. I have suggested that such characterization is best performed on the meta-level, stating postulates that circumscribe rational behavior of reasoning agents. Possible rationality postulates for plausible, explanatory, and confirmatory reasoning have been discussed at some length.

A final word on the issue of hypothesis *selection*, which is the equally crucial but complementary step in intelligent reasoning. In my view, the process of evaluating possible hypotheses to determine which one(s) will be actually adopted is an extra-logical one. By this I mean that it does not give rise to a proof theory in any interesting sense. Furthermore, any hypothesis evaluation procedure will be construed from measures of probability, interestingness, or information content. Logic deals with possible conclusions, not actual ones. This is even true for deduction, which only characterizes tautologies, not interesting mathematical theorems. My conjecture is that successful evaluation procedures (e.g. based on Bayesian or subjective probabilities) will be applicable across a range of different reasoning forms. Thus, while hypothesis generation distinguishes reasoning forms, hypothesis evaluation unifies them.

References

van Benthem, J. (2000) Reasoning in reverse. In Flach and Kakas (2000a): ix–xi.

Flach, P. A. (1998) Comparing consequence relations. In A. G. Cohn, L. Schubert and S. C. Shapiro (eds.), *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann: 180–9.

Flach, P. A. (2000a) On the logic of hypothesis generation. In Flach and Kakas (2000a): 89–106.

Flach, P. A. (2000b) Logical characterisations of inductive learning. In D. M. Gabbay and P. Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management*, vol. 4: *Abduction and Learning*. Dordrecht: Kluwer Academic.

Flach, P. A. and Kakas, A. C. (eds.) (2000a) *Abduction and Induction: Essays on their Relation and Integration*. Dordrecht: Kluwer Academic.

Flach, P. A. and Kakas, A. C. (2000b) Abductive and inductive reasoning: background and issues. In Flach and Kakas (2000a): 1–27.

Gabbay, D. M. (1994) Classical vs non-classical logics (the universality of classical logic). In D. M. Gabbay, C. J. Hogger and J. A. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 2: *Deduction Methodologies*. Oxford: Clarendon Press, 359–495.

Hempel, C. G. (1945) Studies in the logic of confirmation. *Mind*, 54(213 & 214), 1–26 and 97–121.

Kraus, S., Lehmann, D. and Magidor, M. (1990) Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44, 167–207.

Makinson, D. (1994) General patterns in nonmonotonic reasoning. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 3. Oxford: Clarendon Press, 35–110.

Lipton, P. (1991) *Reasoning to the Best Explanation*. London: Routledge and Kegan Paul.

Peirce, C. S. (1958) C. Harstshorne, P. Weiss and A. Burks (eds.), *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.

43

Actions and Normative Positions: A Modal-Logical Approach

ROBERT DEMOLOMBE AND ANDREW J. I. JONES

1 An Approach to the Logic of Action

Influenced by the earlier work of, in particular, Alan Ross Anderson (1967), Stig Kanger (1957; Kanger and Kanger 1966), and Georg Henrik von Wright (1963), Ingmar Pörn produced in 1970 a work entitled *The Logic of Power*.

The aim of the book was to develop some modal-logical tools and to apply them to the characterization of such concepts as *influence*, *control*, *right*, and *norm* – concepts which figure centrally in our understanding of social systems. Not surprisingly, a logic of action was one of the core components of Pörn's formal-logical framework.

Action sentences of the kind

(1) John opens the door

were assigned the logical form

(2) $D_i A$

to be read as '*i* brings it about that A,' where D_i is a relativized modal operator and A describes the state of affairs brought about. Pörn (1970: 4–5) recognized that the logical form he adopted for (1) was a simplification. Although (1) entails

(3) John brings it about that the door is open.

(3) certainly does not entail (1). If, for example, it is the case that

(4) John keeps the door open

then (3) is true whilst (1) may well be false. As Pörn pointed out, the difference in sense between (1) and (4) may be explained by reference to pairs of successive occasions. The truth of (1) requires that, on the earlier of two occasions, the door in question is not open, and then John does what he does and – as a result – the door is open on the later occasion. Whereas the truth of (4) requires the door to be open on the earlier occasion

and – as a result of John's action – *still* open on the later occasion. The 'brings it about that . . .' representation of action sentences is a simplification in (at least) the sense that (2) does not discriminate between (1) and (4). Marking an important point of contrast with the approach of von Wright (1963), Pörn noted that ". . . the notion of a pair of successive occasions is not fundamental to our logic of action" (1970: 4). We might say that Pörn's logic of action sentences is an abstraction, which ignores the change-of-state-over-time aspect of actions, and focuses instead on just two factors: who the agent is, and what state of affairs it is that results from the agent's action. For certain purposes – and in particular for the applications of the logic of action that interested Pörn – an abstraction of this kind is entirely appropriate. We may also note, in passing, that Pörn's approach ignored too the question of the *means by which* an agent secured, through his action, a particular result. (But in his later work, Pörn (1977: chapter 3) gave an analysis of sentences of the kind 'i brings it about that A by bringing it about that B' which drew on automata theory.)

The logic Pörn assigned to sentences of the form D_iA was that of a (relativized) normal modality of type KT in the Chellas (1980) classification. (We ignore here Pörn's treatment of quantification and modality, and restrict attention to the propositional modal logic). In barest outline, a semantical characterization of the D_i-logic may be given as follows: a standard model M is a triple $\langle W,R^{D}_{i},V\rangle$, where W is a set of possible worlds, R^{D}_{i} is a binary relation on W (defined for each agent *i*), and V assigns to each atomic sentence a subset of W (the set of worlds at which that atomic sentence is true). R^{D}_{i} is required to be reflexive: that is, for each world $u \in W$, and for each agent *i*, $\langle u,u\rangle \in R^{D}_{i}$. Truth conditions for non-modal sentences are specified in the usual way for classical propositional logic, and for modal sentences as follows:

(C.D) $M,u \models D_iA$ iff $M,v \models A$ for all $v \in W$ such that $\langle u,v\rangle \in R^{D}_{i}$
(C.C) $M,u \models C_iA$ iff $M,v \models A$ for at least one $v \in W$ such that $\langle u,v\rangle \in R^{D}_{i}$

As usual, a sentence is said to be valid iff it is true at all worlds in all models, and where A is valid we write $\models A$.

Pörn read sentences of the form C_iA as "it is possible for all that *i* does that A." Given the structure of the truth condition (C.C), it is apparent that the intuitive understanding of the accessibility relation R^{D}_{i} is as follows: $\langle u,v\rangle \in R^{D}_{i}$ iff v is possible relative to u with respect to all that *i* does at u. It is readily shown that sentences of the following forms are valid:

DDC. $D_iA \leftrightarrow \neg C_i \neg A$
DM. $D_i(A \wedge B) \rightarrow (D_iA \wedge D_iB)$
DC. $(D_iA \wedge D_iB) \rightarrow D_i(A \wedge B)$
DK. $(D_iA \wedge D_i(A \rightarrow B)) \rightarrow D_iB$
DT. $D_iA \rightarrow A$

Furthermore, the following rule holds:

DRK. If $\models (A_1 \wedge A_2 \wedge \ldots \wedge A_n) \rightarrow A$ then
$\models (D_iA_1 \wedge D_iA_2 \wedge \ldots \wedge D_iA_n) \rightarrow D_iA$ for $n \geq 0$.

DT. expresses what is sometimes referred to as the 'success' condition, and captures the obvious truth that if an agent brings it about that A, then A is indeed the case. The validity of DT. turns essentially on the reflexivity of the accessibility relation.

For the cases n = 0 and n = 1, we have the following instances, respectively, of DRK.:

DRN. If $\models A$ then $\models D_iA$
DRM. If $\models (A_1 \rightarrow A)$ then $\models (D_iA_1 \rightarrow D_iA)$

As logical properties of the action operator, both of these two rules are intuitively problematic. The first says that each agent brings about all logical truths – but, surely, that which is logically true is unavoidably the case, and thus falls outside the scope of anyone's agency? The second says that any agent brings about all of the logical consequences of that which he brings about. So, for instance, if *i* brings it about that *j* brings it about that A, then – in virtue of DRM. and DT. – *i* brings it about that A. But there are certainly interpretations of 'bringing it about' for which we would not want a property of this kind to hold, as when we say that although *i* brought it about that *j* brought it about that A, *i* did not *himself* bring it about that A. A second problematic instance of DRM. arises if we consider expressions of the kind '*i* brings it about that *j* knows that A.' Since *j*'s knowing that A logically implies the truth of A, it will now follow from DRM. that *i* brings it about that A if he brings it about that *j* knows that A.

It is fair to say that problems of the kind raised by DRN. and DRM. led Pörn (and Kanger) to move away from using a normal modality (in the sense of Chellas (1980)) for the characterization of 'brings it about that . . .' (all normal modalities are closed under logical consequence in the sense expressed by the rule DRK.).

Pörn (1977) abandoned the idea that the logic of expressions of the kind '*i* brings it about that A' could be articulated in terms of D_iA alone. Following Kanger (1972), he adopted the hypothesis that sentences of the form D_iA should be read "it is necessary for something which *i* does that A," and that "*i* brings it about that A" entails D_iA. The question then, of course, is to decide what *else*, in addition to 'necessity for something which *i* does' is involved in '*i* brings it about that . . .'. The answer Pörn and Kanger provided can best be introduced by the following remark:

> The ascription of causality to an agent normally suggests either that but for his action it would not be the case that A or that but for his action it might not be the case that A. The notions of counteraction conditionality are not present in the concept of that which is necessary for something that an agent does. As evidence of this one may cite the fact . . . that if it is logically necessary and hence unavoidable that A, then A is also necessary for something that an agent does. (Pörn 1977: 5)

To capture the notion of counteraction conditionality, Pörn introduced modal expressions of the form D'_iA, read as 'but for *i*'s action it would not be the case that A.' In the semantics, a new accessibility relation $R_i^{D'}$ (relativized to each agent *i*) was incorporated; where $\langle u,v\rangle \in R_i^{D'}$, v is said to represent a situation in which *i* does not do any of the things that he does in u.[1] D'-expressions were assigned the following truth condition:

(C.D′) $M,u \models D'_i A$ iff $M,v \models \neg A$ for all $v \in W$ such that $\langle u,v \rangle \in R_i^{D'}$.

The new relation, $R_i^{D'}$, was required to be irreflexive and serial. (We note in passing, without entering into details, that Pörn also adopted conditions linking the two accessibility relations R_i^{D} and $R_i^{D'}$, and that in Pörn (1977) R_i^{D} was required to be both reflexive *and* transitive).

Expressions of the form $C'_i A$ were read 'but for *i*'s action it might not be the case that A' and assigned the following truth conditions:

(C.C′) $M,u \models C'_i A$ iff $M,v \models \neg A$ for at least one $v \in W$ such that $\langle u,v \rangle \in R_i^{D'}$.

It is now readily shown that sentences of the following forms are valid:

D′D′C′. $D'_i A \leftrightarrow \neg C'_i \neg A$
D′D. $D'_i A \rightarrow C'_i A$

Furthermore, D′ is a normal modality, and thus the counterparts to the schemas DM., DC., and DK., and to the rule DRK., also hold for the D′ modality.

So the action logic now contains two normal modalities and their respective duals, in terms of which a new analysis of sentences of the type '*i* brings it about that A,' now represented by $E_i A$, can be formulated. Pörn opted for the following definition:

$E_i A =_{df} D_i A \wedge C'_i A$

So *i* brings it about that A iff A is necessary for something that *i* does *and* but for i's action it might not be the case that A. The two conjuncts represent, respectively, a positive and a negative condition on agent causation. (Here there is a clear point of similarity with the STIT-analysis of agency later put forward by Nuel Belnap and his associates (e.g. 1990)). A comparative overview is way beyond the scope of the present paper, but valuable accounts of these and related approaches to the logic of action are to be found in Elgesem (1997) and Hilpinen (1997).

The E-modality is defined as a conjunction of two normal modalities, but it is not itself normal. For instance, the counterpart to DRN.:

ERN. If $\models A$ then $\models E_i A$

does not hold. On the contrary, the following rule is valid:

ER¬N. If $\models A$ then $\models \neg E_i A$

and this captures in an obvious way the claim that logical truths fall outside the scope of anyone's agency. Furthermore, neither the counterpart to DRM. nor the counterpart to DM. is valid for the E-modality. Since the E-modality is classical in the sense of being closed under logical equivalence (see Chellas 1980), the validity of the E-counterpart to DM. – call it EM. – would carry the disastrous consequence that there are no true

sentences of the form E_iA. The explanation is this: suppose E_iA; then, since A is logically equivalent to $(A \wedge T)$, where T is any tautology, it follows that $E_i(A \wedge T)$. But then if EM. were to be valid it would follow that E_iT, a result which is of course inconsistent with the valid rule ER¬N.

The E-counterparts to DC., DK. and DT. are each valid.

An alternative definition of the 'brings it about' operator was offered by Kanger (1972: 108):

$$E_i^*A = {}_{df} D_iA \wedge D_i'A$$

according to which an agent *i* brings it about that A iff A is necessary for something that *i* does and but for *i*'s action it *would* not be the case that A. Intuitively, this version of the negative condition on agency appears to demand too much; for it may be that *i* brings it about that A, but that in *some* of the situations which could have arisen if he had not acted in the way he did, A is still the case – perhaps as a result of some other agent's action. Considerations of this sort favor Pörn's weaker formulation of the negative condition. There is also a technical difficulty with Kanger's definition, as has been pointed out by Jones (reported in Pörn 1977: 5). Suppose that *i* brings it about that A and that he brings it about that if A then B. That is, on Kanger's definition:

(5) $D_iA \wedge D_i'A \wedge D_i(A \rightarrow B) \wedge D_i'(A \rightarrow B)$

The second and fourth conjuncts require that, in all of the counteraction conditional alternatives to the given world, both ¬A and ¬(A→B) are true. But since the conditional here is the truth-functional conditional, a contradiction is implied. (In virtue of the seriality of $R_i^{D'}$ there will be at least one counteraction conditional alternative to each world.) Thus there can be no true act descriptions of the form $E_i^*A \wedge E_i^*(A \rightarrow B)$.[2]

It has often been observed that the Pörn–Kanger approach fails to provide an adequate analysis of the concept of action, since the accessibility relations used in the semantics are themselves articulated in terms of what is necessary for what an agent *does* and in terms of what might or would happen if the agent did not act as he does (see Hilpinen 1997: 5). Similar accusations of circularity have been leveled against the possible-worlds semantics of alethic, deontic, and epistemic modalities. If the aim of these semantical treatments of modality had been to *reduce* the concepts concerned to other concepts, then of course the criticism would be justified. But in the case of Pörn – and of many of those who have worked in applied modal logic over the last four decades – the criticism is misplaced. Pörn himself doubted whether a reduction of 'brings it about' to other notions was even possible:

> the principal construction employed, *viz.* "*i* brings it about that A", pertains to agent causality. It is not certain that this construction can be analysed in terms of anything simpler or more fundamental than itself. But it can be elaborated by means of concepts that make it possible to set out the principles of our reasoning with it. (Pörn 1977: 5)

Just the same point may be made in regard, for instance, to Hintikka's (1962) work in epistemic and doxastic logic, and in regard to much of what has been done in deontic

logic. The task has been to provide a formal framework within which our reasoning with the concepts concerned can be systematically investigated, not to effect a reduction of these concepts. Furthermore, a point which applies particularly to Pörn and Kanger, the aim has been to use action modalities and deontic modalities as basic building blocks in the construction of formal characterizations of norm-governed systems. An example of work of that kind will be described in Section 2.

However, other criticisms of Pörn's approach have addressed its adequacy as a basis for analyzing our reasoning about actions. For instance, Dag Elgesem has made some interesting observations about the negative condition in Pörn's definition of 'bringing it about,' suggesting that it collapses two distinct ideas into one:

> The first is that of avoidability in the sense that what is brought about is not logically true . . . The second idea, quite distinct, is that a necessary condition for agency is that the agent's activity is *instrumental* in the production of the result. (Elgesem 1997: 10)

Elgesem develops a new logic of action in which an attempt is made to characterize this distinction. He also notes that his criticism of Pörn's negative condition applies equally well to the version of the negative condition which appears in Belnap's STIT-theory (Elgesem 1997: 18).

2 Normative Act Positions

We now pose the following question: in regard to a particular state of affairs, and a particular agent, what is the class of possible relations between that state of affairs and the successful actions of the agent? We answer the question by generating the class of possible *act-positions* for a given agent *i* vis-à-vis a state of affairs A.

The state of affairs A either obtains or does not obtain; that is either A or $\neg$A holds. Now prefix each of A and $\neg$A with, first, the operator E_i and, second, its internal negation $E_i\neg$. Four formulas result:

E_iA, $E_i\neg A$, $E_i\neg A$, $E_i\neg\neg A$

Of these, the second and third are syntactically identical, and the first and fourth are logically equivalent, given that – as was observed in the previous section – the action operator is closed under logical equivalence. Now form the external negations of these two remaining act expressions (E_iA, $E_i\neg A$), and arrange the four resulting expressions in the form of two truth-functional tautologies:

(i) $E_iA \vee \neg E_iA$
(ii) $E_i\neg A \vee \neg E_i\neg A$

There are of course four distinct ways of choosing just one disjunct from each of (i) and (ii):

(E0) $E_iA \wedge E_i\neg A$
(E1) $E_iA \wedge \neg E_i\neg A$

(E2) $\neg E_i A \wedge E_i \neg A$
(E3) $\neg E_i A \wedge \neg E_i \neg A$

We now recall that the success condition (the counterpart to DT.) is valid for the E-operator:

ET. $E_i A \rightarrow A$

Thus (E0) is a logical contradiction, and does not represent a possible act-position. Furthermore, $E_i A$ logically implies $\neg E_i \neg A$ and $E_i \neg A$ logically implies $\neg E_i A$. So the class of possible act-positions (for one agent and one state of affairs) may be re-written as:

(E1) $E_i A$
(E2) $E_i \neg A$
(E3) $\neg E_i A \wedge \neg E_i \neg A$

The members of the set {(E1),(E2),(E3)} are mutually exclusive, and their disjunction is a tautology. That is to say, for any given agent *i*, and for any state of affairs A, precisely one of (E1), (E2), (E3) holds: either *i* brings it about that A, or *i* brings it about that $\neg$A, or *i* is passive (he does not bring it about that A and he does not bring it about that $\neg$A). We have now answered our first question by giving an exhaustive characterization of the class of one-agent act-positions vis-à-vis a given state of affairs.

Let us now introduce the normative/deontic modality O, and read expressions of the form OA as 'it is obligatory that A.' We define expressions of the form PA, 'it is permitted that A' as follows:

(Def.P) $PA =_{df} \neg O \neg A$

We may now use the set {(E1),(E2),(E3)} of one-agent act-positions as a basis on which to construct, or generate, the class of one-agent normative act-positions. First, prefix each of (E1)–(E3) with the operator O, and then prefix each of them with O$\neg$. From these six expressions generate six more, by negating each one of them. Display the resulting twelve expressions as a set of six tautologous disjunctions:

(iii) $OE_i A \vee \neg OE_i A$
(iv) $OE_i \neg A \vee \neg OE_i \neg A$
(v) $O \neg E_i A \vee \neg O \neg E_i A$
(vi) $O \neg E_i \neg A \vee \neg O \neg E_i \neg A$
(vii) $O(\neg E_i A \vee \neg E_i \neg A) \vee \neg O(\neg E_i A \vee \neg E_i \neg A)$
(viii) $O \neg (\neg E_i A \wedge \neg E_i \neg A) \vee \neg O \neg (\neg E_i A \wedge \neg E_i \neg A)$

There are 64 ways of choosing just one disjunct from each of (iii)–(viii). That is, from (iii)–(viii) we may generate 64 distinct conjunctions, each of which contains 6 conjuncts. Suppose now that the logic of the O-modality is that of Standard Deontic Logic (SDL), which is a normal modal system of type KD. This means that SDL is based on classical propositional logic, and contains (Def.P), the axiom schema:

OD. $OA \rightarrow PA$

and the rule

ORK. $\dfrac{(A_1 \wedge A_2 \wedge \ldots \wedge A_n) \rightarrow A}{(OA_1 \wedge OA_2 \wedge \ldots OA_n) \rightarrow OA}$ $n \geq 0$

Given these logical properties, and those already assigned to the E-modality, it may be shown that 57 of the 64 conjunctions are logically inconsistent. In virtue of relations of logical implication between their conjuncts, each of the seven remaining conjunctions may be simplified by removing redundant conjuncts. The result is the following set of one-agent normative act-positions:

(N1) $PE_iA \wedge PE_i\neg A \wedge P(\neg E_iA \wedge \neg E_i\neg A)$
(N2) $PE_iA \wedge O\neg E_i\neg A \wedge P(\neg E_iA \wedge \neg E_i\neg A)$
(N3) $PE_iA \wedge PE_i\neg A \wedge O(E_iA \vee E_i\neg A)$
(N4) $O\neg E_iA \wedge PE_i\neg A \wedge P(\neg E_iA \wedge \neg E_i\neg A)$
(N5) OE_iA
(N6) $O(\neg E_iA \wedge \neg E_i\neg A)$
(N7) $OE_i\neg A$

The members of this set of positions are mutually exclusive, and their disjunction is a tautology. Thus, for any given agent *i*, and for any state of affairs A, precisely one of these normative act-positions holds. The seven positions correspond to Lars Lindahl's (1977: 92) basic types of one-agent legal positions. Lindahl's book develops in some detail the pioneering work of Kanger, who combined action and deontic modalities in an attempt to systematise further W. N. Hohfeld's (1923) theory of rights-relations. (The account of how to generate normative positions, given above, differs from that of Lindahl. It is taken from Jones and Sergot (1993), to which the reader is also referred for some comparisons of this approach with those of Kanger and Lindahl. Note, in particular, that the basic structure of the generation procedure itself does not turn on any particular choice of logics for the E- and O-modalities, although of course the content and size of the generated class of possibilities does depend on that choice.)

It is clear that once an exhaustive characterization of a class of positions has been specified, one can use it as a definitive guide in attempting to determine the appropriate logical form to be assigned to a particular norm. In Jones and Sergot (1993), the main example provided to illustrate this procedure was taken from a set of norms regulating access (by various categories of agents) to sensitive, confidential information. (The scenario was a psychiatric hospital, and the norms assigned/denied rights to patients, doctors, nurses, administrative staff, etc., with respect to accessing patients' medical files.) The example norm said that a patient did not have the right to access his/her own file. One interpretation of this norm would take it to be expressing a denial that a patient is empowered to insist on access to his/her file. A different interpretation views the norm as (in part) denying permission to a patient to access his/her own file.[5] Now consider this second mode of interpretation in relation to the set of seven one-agent normative act-positions, supposing *i* to be an agent in the category of patient, and letting A be the

sentence '*i* has access to *i*'s own file.' Which of (N1)–(N7) captures the appropriate logical form? Clearly (N1), (N2), (N3), and (N5) can all be ruled out immediately, since each requires that it is permitted that *i* brings it about that *i* has access to *i*'s own file. Each of the remaining three cases implies that E_iA is not permitted. Given the fact that *i* is in the category of psychiatric patient, it is perhaps unlikely that the authorities who formulated the norm intended to place *i* under an obligation to bring it about that *i* does not access *i*'s file: in which case (N7) is eliminated from the set of plausible candidates. (N6) makes it *obligatory* that *i*'s act-position (in regard to the state of affairs concerned) is one of passivity, which seems bizarre in the circumstances. Thus, from such considerations as these, (N4) emerges as the appropriate choice of logical form, making it obligatory that it is not the case that *i* accesses *i*'s own file, permitting *i* to bring it about that *i* does not access *i*'s own file, but also permitting *i* to remain passive.

The point behind the discussion of this example is this: the set of seven positions maps out, exhaustively, at a particular level of analytical detail (*one* agent, *one* state of affairs, *one* pair of interdefinable deontic operators) the class of available interpretations. Consideration of the meaning of the particular norm, and of the probable intentions of the norm-giver, then point to the most appropriate choice.

However, for an example of this kind, it would be unsatisfactory to end the search for the correct logical form at this stage. One of the fundamental insights from Hohfeld was that rights are *relational*, and cannot be completely specified in terms of an individual's permissions considered in isolation. Another example might help illustrate the point: in the eyes of the Norwegian state, a child of 12 years is permitted to place bets on sporting events at a state-owned betting shop. But the state does not thereby grant the child the *right* to place such bets, since it does not forbid some other agent (the child's parents, say) from preventing his betting activities. Returning to the access-control example, the relational aspect emerges when we address the issue of *who* it is that is likely to be assigned the *responsibility* for ensuring that *i* does not have access to *i*'s own file. Presumably *not i* himself, which is why it seemed implausible to suppose that the norm-giving authority intended (N7). Thus we see the need to bring into consideration the role of other agents: what will their normative position be vis-à-vis the state of affairs '*i* (the patient) has access to *i*'s own file'?

The generation procedure can readily be extended to facilitate a systematic investigation of this question. First rewrite (N1)–(N7), replacing each occurrence of E_i by one of E_j, but keeping the *same* interpretation as before of the scope-formula A ('*i* has access to *i*'s own file'). (We may consider, for example, that *j* is an agent in the category of doctor in the institution concerned.) There are of course 49 conjunctions obtainable by selecting one member of the set

$$\{(N1_i), (N2_i), \ldots (N7_i)\}$$

and conjoining it with one member of the set:

$$\{(N1_j), (N2_j), \ldots (N7_j)\}$$

Of these 49, 35 are internally consistent (see Lindahl 1977: 128). We may call these 35 conjunctions the set of two-agent normative act-positions. Just six of the 35 con-

junctions contain $(N4_j)$ – the interpretation suggested above for the one-agent level of analysis. In these six cases, $(N4_j)$ is conjoined with, respectively, $(N1_j)$, $(N2_j)$, $(N3_j)$, $(N4_j)$, $(N6_j)$, and $(N7_j)$. (The $(N5_j)$ case is ruled out because the conjunction $OE_jA \wedge PE_i\neg A$ is inconsistent.) What then is the most likely intended interpretation at this two-agent level? Well, each of $(N1_j)$, $(N2_j)$, and $(N3_j)$ contains PE_jA, which is clearly incompatible with the intended interpretation. $(N4_j)$ allows doctor j to remain passive with respect to *i*'s having access to *i*'s own file, whilst $(N6_j)$ makes passivity obligatory; so considerations of probable assignment of responsibility eliminate these two cases. In which case the appropriate choice appears to be $(N7_j)$, giving, finally, the following two-agent normative act-position:

$$O\neg E_iA \wedge PE_i\neg A \wedge P(\neg E_iA \wedge \neg E_i\neg A) \vee OE_j\neg A$$

The first conjunct here can be eliminated as redundant, since it is logically implied by the fourth.

And then it would be possible to complicate matters further, by introducing more categories of agents. Or perhaps (for the analysis of some other types of norms) one might be interested in starting the generation of normative positions not from one-agent act-positions, but from two-or-more-agent act-control/influence positions, expressed in terms of sequences of two or more action operators relativized to different agents. Or perhaps one might add further operators, to express not only successful action, but attempted action.

These are just *some* of the dimensions along which the complexity of the analysis might be increased, and with it the number of conjunctions to be considered. Clearly, the task of *manually* formulating the class of consistent conjunctions will soon become unmanageable: there is a need for automation of the generation procedure. Considerable progress has been made in this direction in recent work by Marek Sergot (1999). The prospect is emerging of a rather sophisticated automated support tool, which can assist in the process of drafting clear specifications of norms pertaining to the rights of agents.

Despite the expressive power of a language combining deontic and action modalities, with respect to the characterization of rights-relations, there are also some rather significant shortcomings, as has been indicated by David Makinson (1986). For instance, Kanger's framework appears to be incapable of capturing the Hohfeldian notion of *power*, and of properly representing the *directionality* which is often characteristic of rights-relations, as when one agent (the bearer) has an obligation vis-à-vis another agent (the counterparty).

As regards the first of these shortcomings, it should be noted that there is good reason to believe that an agent's being assigned certain *legal* or *institutional powers* is not to be confused with his being *permitted* to perform certain acts. Nor should it be identified with his having the physical ability to act – see Makinson (1986) and Jones and Sergot (1996) for examples and discussion. The latter paper combines the E-operator with a modal conditional connective, in an attempt to capture the idea that, within a given institution, the actions of a designated agent may *count as* a means of establishing particular kinds of normative positions, as when a priest is empowered to

create a state of marriage, or a Head of Department is empowered to assign teaching duties.

As regards the question of how to represent the *directionality* of rights-relations, the reader is referred to the works of Henning Herrestad (1996) and Christen Krogh (1997).

Hopefully these remarks suffice to indicate the potential role of the modal logic of action, in combination with deontic and other modalities, in the characterization of norm-governed systems of agents.

Notes

1 See Segerberg (1985) for a discussion of some difficulties involved in Pörn's interpretation of this accessibility relation.
2 The underlying problem here concerns the use of the truth-functional conditional to represent counterfactual situations – a task for which it is well-known to be ill-suited.
3 We return below to the distinction between *empowered* and *permitted*.

References

Anderson, A. R. (1956) *The Formal Analysis of Normative Systems*. New Haven; reprinted In N. Rescher (ed.), *The Logic of Decision and Action*. Pittsburgh, 1967, pp. 147–213.
Belnap, N. and Perloff, M. (1990) Seeing to it that: a canonical form for agentives. In H. Kyburg, et al. (eds.), *Knowledge Representation and Defeasible Reasoning* (pp. 167–90). Dordrecht: Kluwer Academic.
Chellas, B. F. (1980) *Modal Logic*. Cambridge: Cambridge University Press.
Elgesem, D. (1997) The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2, 1–46.
Herrestad, H. (1996) *Formal Theories of Rights*. Oslo: Juristforbundets Forlag.
Hilpinen, R. (1997) On action and agency. In E. Ejerhed and S. Lindström (eds.), *Logic, Action and Cognition – Essays in Philosophical Logic* (pp. 3–27). Dordrecht: Kluwer Academic.
Hintikka, J. (1962) *Knowledge and Belief*. Ithaca, NY: Cornell Univesity Press.
Hohfeld, W. N. (1923) *Fundamental Legal Conceptions as Applied in Judicial Reasoning and Other Legal Essays*, ed. W. W. Cook. New Haven.
Jones, A. J. I. and Sergot, M. J. (1993) On the characterisation of law and computer systems: the normative systems perspective. Ch. 12. In J.-J. Ch. Meyer and R. J. Wieringa (eds.), *Deontic Logic in Computer Science*. New York: John Wiley.
Jones, A. J. I. and Sergot, M. J. (1996) A formal characterisation of institutionalised power. *Journal of the IGPL*, 4(3), 429–45.
Kanger, S. (1957) *New Foundations for Ethical Theory*. Stockholm; reprinted. In R. Hilpinen (ed.), *Deontic Logic – Introductory and Systematic Readings*. (pp. 36–58). Dordrecht: Reidel.
Kanger, S. (1972) Law and logic. *Theoria*, 38, 105–32.
Kanger, S. and Kanger, H. (1966) Rights and parliamentarism. *Theoria*, 32, 85–115.
Krogh, C. (1997) *Normative Structures in Natural and Artificial Systems*. Complex 5/97, TANO-Aschehoug.
Lindahl, L. (1997) *Position and Change*. Dordrecht: Reidel.
Makinson, D. (1986) On the formal representation of rights relations. *Journal of Philosophical Logic*, 15, 403–25.
Pörn, I. (1970) *The Logic of Power*. Oxford: Blackwell.

Pörn, I. (1977) *Action Theory and Social Science*. Dordrecht: Reidel.
Segerberg, K. (1985) On the question of semantics in the logic of action: some remarks on Pörn's logic of action. In G. Holmström and A. J. I. Jones (eds.), *Action, Logic and Social Theory, Acta Philosophica Fennica*, 38, 282–98.
Sergot, M. J. (1999) Normative positions. In P. McNamara and H. Prakken (eds.), *Norms, Logics and Information Systems*. Amsterdam: IOS Press.
von Wright, G. H. (1963) *Norm and Action – A Logical Enquiry*. London: Routledge & Kegan Paul.

This page intentionally left blank

Part XIV

MECHANIZATION OF LOGICAL INFERENCE AND PROOF DISCOVERY

This page intentionally left blank

44

The Automation of Sound Reasoning and Successful Proof Finding

LARRY WOS AND BRANDEN FITELSON

1 The Cutting Edge

The consideration of careful reasoning can be traced to Aristotle and earlier authors. The possibility of rigorous rules for drawing conclusions can certainly be traced to the Middle Ages when types of syllogism were studied. Shortly after the introduction of computers, the audacious scientist naturally envisioned the *automation* of sound reasoning – reasoning in which conclusions that are drawn follow logically and inevitably from the given hypotheses. Did the idea spring from the intent to emulate Sherlock Holmes and Mr. Spock (of *Star Trek*) in fiction and Hilbert and Tarski and other great minds in nonfiction? Each of them applied logical reasoning to answer questions, solve problems, and find proofs. But can such logical reasoning be fully automated? Can a single computer program be designed to offer sufficient power in the cited contexts?

Indeed, while the use of computers was quickly accepted for numerical calculations and data processing, intense skepticism persisted – even in the early 1960s – regarding the ability of computers to apply effective reasoning. The following simple (but perhaps deceptive) example provides a taste of the type of argument that might have been used to support this skepticism.

If one is given a puzzle concerning who holds which jobs, is told that the job of nurse is held by a male, and is asked about the possible jobs for Roberta, one quickly concludes that she is not the nurse. How could a computer program rapidly draw this correct conclusion? After all, the computer would not know that Roberta is (implicitly) female, and, of greater usefulness, it would not know that being a female implies that one is not a male. In fact, even a person often does not realize that the latter fact is used in drawing the correct conclusion for this puzzle. Since the answering of deep questions and the solving of hard problems require far more lengthy paths of reasoning, where do things stand today regarding the automation of drawing conclusions that are sound and relevant, and what is the contemporary view concerning this effort?

In answer to the latter question, still debated with vigor and fascination is the value of automation both in the context of inference rule application for drawing conclusions and in the context of useful proof finding, whether the area be mathematics, logic, circuit design, program verification, or puzzle solving. This essay may settle the issue

for many. Indeed, proofs that for decades have eluded some of the greatest logicians and mathematicians have recently been obtained with a single program, William McCune's OTTER (McCune 1994). (Various other reasoning programs exist; some offer far less power, while others are special-purpose programs, for example, designed mainly for program verification. A special-purpose program is in the majority of cases not nearly as useful as a general-purpose program is in the context of attacking a wide variety of deep questions, such as offered by logic and mathematics.) In Section 3, we shall list some of the theorems that had remained elusive for many, many years, theorems that were recently proved by an automated reasoning program and, moreover, proved in but a few CPU-hours. For the eager reader, we note that the material that is offered in Section 2 is not required for an appreciation of the significance of the successes.

The material presented here is at the cutting edge, featuring proofs not found in the literature of the masters that include Hilbert and Ackermann, Tarski and Bernays, Rose and Rosser, Łukasiewicz, and Meredith. The proofs concern results that fall mainly into three classes: those proved in a nonaxiomatic manner, where an axiomatic proof is preferred; those announced without proof; and those whose proof eluded all attempts. To be current, we focus mainly on successes from mid-1998 to the present. Our presentation – emphasizing examples rather than formalism – makes the content of this essay equally accessible to student and researcher alike, and we assume no background. Nevertheless, what we discuss offers depth, scope, and challenge. Among the treasure, one finds that – through automation – various theorems have been proved whose proof waited for many, many years. One also finds open questions to consider, questions that might be attacked using the program OTTER offered in the first of two intriguing new books on automated reasoning (Wos and Pieper 1999, 2000).

Immediately one might ask how a computer program was able to extend the work of great scholars in such an impressive manner. Indeed, not much more than 50 years ago, what did the eminent logician Łukasiewicz fail to see when he asserted that a formalized proof cannot be 'discovered mechanically' but can only be 'checked mechanically' (Łukasiewicz 1948)? (His remark would still have held essentially even in the late 1970s.) Surely the execution speed of today's machine cannot be the answer: logic and mathematics are far too deep to admit such a simple solution in the context of proof finding. Nor can the answer rest with overcoming the obstacle of the implementation of sound reasoning (inference rules); this obstacle was not severe. Can it be (as some prophesied in the 1960s) that a means has been found to effectively emulate the problem-solving skills of the gifted? That explanation also misses the mark, misses it widely, for no such means has yet been devised.

Instead, (for us) the key to the discovery of so many long-sought proofs rests with the reliance on diverse strategies, some to restrict a program's reasoning and some to direct it. (Some authors, including M. Fitting, use the word *heuristic* in a manner similar to our use of the word *strategy*; other authors sometimes use *strategy* in the context of a specific problem rather than to problems in general.) The occasional ease of discovery is startling even to us who have used OTTER for years. We shall illustrate a powerful strategy that restricts reasoning and an effective strategy that directs it. The nature of both strategies, as well as that of numerous others that are offered by OTTER, permits their embedding in many unrelated reasoning programs.

We shall shed much light on the texture of the strategies that are employed, and thus show why the automation of proof finding is often so successful. To address the concerns of many, and to complete much of the picture, among the pressing questions, this essay answers the following. What, if any, are the important differences between a program's reasoning and that of a person? What are the advantages and disadvantages of reliance on a program that applies logical reasoning? What (if any) means are employed to enable a program to reason effectively, in contrast to merely accruing new conclusions until by accident the goal is reached? How does such a program 'know' when an assignment has been completed, in particular, that a proof has been found? To what practical uses can such a program be put? Which significant open questions have been answered by such a program, and how much guidance was provided by the researcher?

2 Automated Reasoning, Principles and Elements

The breakthrough leading to the more successful mechanization (automation) of inference rule application and proof finding can perhaps be traced to the formulation of and adherence to a few principles. The first of these principles asserts that more general statements are preferred over less general. The second (which overlaps the first) concerns the avoidance of what might be termed 'person-oriented reasoning.' To illustrate the two principles, a single example taken from everyday language suffices; it also provides a taste of the language typically used by the more powerful reasoning programs and a glimpse of the typical test (discovery of a proof by contradiction) used to determine assignment completion.

Consider the following two statements, innocently uttered by someone in casual conversation. "Plato likes everybody" and "Nobody likes Plato." A casual interpretation of the given two utterances perhaps leads to mere acceptance and to no conclusion. However, closer inspection shows that the two statements contradict each other. Indeed, formally, the first can be written, for all x, LIKES(Plato,x). Where '–' denotes logical **not**, the second can be formally written, for all y, –LIKES(y,Plato). The contradiction is quickly made transparent by substituting Plato for both x and y in the respective two statements. In other words, overlooked in the casual interpretation is the fact that the first statement includes the case that Plato likes himself, where the second says that he does not. Indeed, where everyday language typically would have permitted the two statements to be accepted simultaneously without blanching, logically the two form a contradictory pair.

If the explicit use of 'for all' (*universal quantification*) is removed with the corresponding variables treated as implicitly meaning 'for all,' then one has two examples of a *clause* and a small taste of the basic linguistic unit used to present information to an automated reasoning program. (For OTTER, a variable is denoted by an expression beginning with a letter between lower-case u and z inclusive.) The example also provides a glimpse of the typical test for assignment completion that an automated reasoning program relies upon.

Regarding both the principle of generality preference and that of person-oriented reasoning avoidance, the detection of contradiction (inconsistency) by the program

does not require the application of the cited substitution to explicitly produce, respectively, the two variable-free statements LIKES(Plato,Plato) and –LIKES(Plato,Plato). Indeed, the program prefers the two (original) statements as uttered, in their given generality, *without* making the substitution that would emulate what a person would most likely do. To further clarify the picture focusing on both the preference for generality and the avoidance of person-oriented reasoning, two additional examples are in order, examples offering more depth.

When a researcher, such as an algebraist, is producing a proof, the conclusions that are presented are often influenced by their intended use. Therefore, although the conclusion in the middle of the presentation that the square of x is the identity e may be justified, instead one might find the conclusion $(yz)\,(yz) = e$. Because of the basic mechanisms relied upon (which will be illustrated), the type of reasoning program under discussion would prefer the conclusion EQUAL(prod(x,x),e) and would avoid EQUAL(prod(prod(y,z), prod(y,z)),e). (Each equality is a clause, more examples of the language used in automated reasoning.) The first equality offers more generality; the second emulates the kind of reasoning more typical of the researcher not relying on a reasoning program.

Although most likely far from obvious, the preference for generality contributes markedly to effectiveness. (For a pertinent example of the type of general reasoning, in the context of equality, applied by a reasoning program but not ordinarily by a person, see the discussion of *paramodulation* in the section entitled "Inference Rules" below.) Also far from obvious, no effective automated technique is known for wisely choosing which of the myriad of less general conclusions to draw, indeed, how to effectively emulate that aspect of person-oriented reasoning. In other words, automated reasoning programs do not offer the type of reasoning called *instantiation*, which can be used to yield the second equality from the first by replacing (instantiating) x by yz. Although instantiation serves logicians and mathematicians well, unless an effective strategy is discovered to control its use, instantiation is unneeded and even unwanted in the context of mechanizing inference rule application and proof finding. Indeed, its use (in effect) conflicts with a reasoning program's preference for generality that in turn contributes to effectiveness.

For the promised second example, an aspect of logic suffices. If one browses with some care in the literature focusing on *implication* (denoted here by i), one finds within proofs the deduction of formulas such as $i(i(x,y),i(x,y))$, where the deduction of the formula $i(z,z)$ would have been justified and sound. Generality was not the choice; rather, the choice was based on what was deemed more convenient for subsequent steps. Similar to the preceding discussion of the two equalities, the automated reasoning program would have deduced the latter formula, because of its generality, and would have avoided the former even though its use emulates the mind of an expert. Because of this practice, with reliance on the program, occasionally more general proofs are found and more general theorems are proved (which we shall cite).

The basic elements of automated reasoning

The paradigm (for the automation of logical reasoning) featured in this essay rests on six elements: a language for presenting the question or problem under study; types of

reasoning (inference rules) for drawing conclusions some of which are adjoined to the supplied information; strategies for controlling the reasoning; a means for simplifying and canonicalizing information; a means for purging types of redundant information; and a means for determining assignment completion (most often, proof finding). Regarding other paradigms, some differ by addition, some by subtraction. Specifically (in the spirit of addition), some offer induction, where such is not the case for the paradigm in focus here. As for subtraction, some paradigms do not retain new conclusions, which (to us) accounts in part for their lack of power compared with that which (for example) OTTER offers by accruing sometimes a vast number of new conclusions. Equally serious, but of a different nature, many paradigms do not emphasize the use of types of strategy, indispensable for attacking deep questions and hard problems in our view. Regarding another crucial omission, some paradigms do not offer a built-in treatment of equality.

Language

For presenting a question or problem for study by an automated reasoning program, the *clause language* (a dialect of first-order predicate calculus) serves nicely. Its lack of richness is an asset, not a liability. Indeed, rich languages offer more obstacles for formulating effective strategies for reasoning within them. However, the nature of the clause language does present at least an annoyance for one who wishes to enlist the aid of a reasoning program.

In the clause language, only two logical connectives are explicitly permitted, **not** (denoted by '–') and **or** (denoted by '|'). Between each pair of clauses logical **and** is present implicitly. In place of logical **if-then** (logical **implies**), **not** and **or** suffice; one simply replaces **if** P **then** Q with **not** P **or** Q. This replacement rule dictates what must be done for the logical operator **equivalent.**

Regarding variables, every variable within a clause is implicitly treated as meaning 'for all,' universally quantified. Existentially quantified variables are replaced with appropriate functions, Skolem functions and constants. Explicit quantification is not permitted. The scope of a variable is limited to the clause in which it occurs. Therefore, if a variable, say x, appears in two different clauses, it is treated as merely a coincidence, as if the two names are distinct. A few examples illustrate how it works.

For the assertion that Nan **and** Larry like cats, one writes two clauses, (1) LIKES(Nan,cats) and (2) LIKES(Larry,cats). The clause language implicitly assumes (logical) **and** between every pair of clauses. If one prefers to be more formal and be more precise, one writes –IS(x,cat)|LIKES(Nan,x) and its counterpart.

Since programs such as OTTER offer a built-in treatment of equality, one can write for the equality of x and minus(minus(x)) the clause EQUAL(x,minus(minus(x))). For the statement that for all x there exists a y with y greater than x, one writes GREATER(x,f(x)), where the function f is a Skolem function introduced for the existentially quantified variable y. The clause exhibits the dependence of y, in the form of $f(x)$, on x.

Inference rules

At the heart of all of the inference rules that are used by the type of program featured here (of which OTTER is but one example) is a procedure called *unification*, a

procedure that looks for substitutions that modify variables as little as need be to find a common expression. For example, from the clause IS(Snowflake,cat) and the clause –IS(x,cat)|LIKES(Nan,x), a program can deduce LIKES(Nan,Snowflake) by replacing x by the constant Snowflake and applying *modus ponens*, the rule that asserts the deducibility of Q from the pair P and P **implies** Q. (Recall that logical **if-then, implies,** can be replaced by using logical **not** and logical **or**.) Similarly, for the two clauses cited earlier focusing on Plato, (as noted) a substitution into each was possible that yielded a contradiction.

In contrast, if one considers the clause Q(x,x) and the clause –Q(y,f(y)), no contradiction can be found because no appropriate substitution exists. The general rule when applying unification (in the context of the preceding example) asserts that one is not allowed to substitute a term containing as a subterm a variable for that variable. As the following illustrates, most-general substitutions are always what the program seeks to find, which is not always the case in the literature of logic and mathematics (as discussed somewhat differently earlier). In the spirit of syllogism, from the clause –P(x)|Q(x,a,u) and the clause –Q(b,y,v)|R(y,v), the program can deduce the clause –P(b)|R(a,u). Although the reasoning would be sound, the program would not, for example, deduce –P(b)|R(a,a).

To determine whether two expressions are unifiable, one seeks a table of substitutions of terms for variables. An effective approach is to, first, rename all the variables so that no variable name appears in common in the two expressions and, second, proceed left to right, continually updating the table. Unification can fail for a variety of reasons, such as when one finds a term containing a variable opposite that same variable.

The arsenal of inference rules that is offered is not restricted to those that consider hypotheses taken two at a time. Indeed, one of those rules (*hyperresolution*) serves perfectly for the study of many areas of logic, as the following shows. First, consider a mundane example concerning relations among people. From the clause –PARENT(x,y)|–FEMALE(x)|MOTHER(x,y) and the clause PARENT(G,K) and the clause FEMALE(G), an application of hyperresolution yields the clause MOTHER(G,K). This inference rule, which by definition is required to yield clauses free of logical **not**, considers the three clauses simultaneously.

Not far removed from this mundane example is the following incarnation of the inference rule *condensed detachment*, frequently used in logic. Indeed, consider the clause –P(i(x,y))|–P(x)|P(y), which is quite reminiscent of *modus ponens*, asserting that the presence of x **implies** y and x justifies the conclusion of y. In the given three-literal clause, the expression unified with the first literal is called the *major premise*, and that unified with the second literal the *minor premise*. If P(i(i(x,y),i(i(y,z),i(x,z)))) is the major premise and P(i(i(u,v),i(v,u))) is the minor premise, the use of condensed detachment yields P(i(i(v,u),z),i(i(u,v),z)) as the conclusion. For the conclusion, no substitution is required for the variables in the minor premise.

Far more complicated (and clearly not easily seen) is the case (taken from equivalential calculus, with the function i replaced by the function e) in which both the major and minor premises are P(e(e(e(e(x,e(y,z)),e(y,x)),e(z,u)),u)) and condensed detachment is applied. The conclusion that is yielded is P(e(x,x)), requiring a nontrivial substitution for the variables in both the major and the minor premise. Such complicated unifica-

tions are in no way difficult for an automated reasoning program, but they can be tiresome (or worse) for an unaided researcher.

Of the various inference rules, one of the more complicated is *paramodulation*, which enables an automated reasoning program to treat equality as if it is 'understood.' Paramodulation – which is the best example of a computer-oriented inference rule, and one that a person probably should not apply by hand – generalizes the usual notion of equality substitution. The following example illustrates the cited generalization and demonstrates that paramodulation is indeed computer oriented. Paramodulating *from* the equation $x + (-x) = 0$ *into* the equation $y + (-y + z) = z$ yields in a single step the conclusion $y + 0 = -(-y)$.

Strategy

Because the space of deducible conclusions is so huge (many, many millions), without the use of strategy to restrict and strategy to direct the reasoning, an attempt to find significant proofs would be doomed. In contrast to reasoning programs, researchers succeed because of much knowledge, intuition, and experience. But often proofs that are desired escape even the masters. In Section 3, we give examples of such proofs – proofs that were missing for decades, but that were found through automation.

Two strategies provide a fair taste of what is needed and of what has made the difference. The first strategy, the *set of support strategy*, was formulated to restrict a program's reasoning. For this strategy, the term 'special hypothesis' was introduced, referring to that part of the problem presentation that is outside of the set of axioms and conclusion to be proved. In one of the two strongly recommended uses, the strategy allows a program to draw a conclusion only if it can be recursively traceable to either the special hypothesis or the denial of the conclusion to be proved. For example, if one is asked to prove that rings in which the cube of x equals x (for all x) are commutative, the special hypothesis consists of the property that $xxx = x$. The denial of the conclusion, in the preceding, consists of the assumption that such rings are *not* commutative, that there exist two elements a and b with ab not equal to ba.

In general, when one asks a reasoning program to attempt to find a proof, one supplies a set of statements that include those that correspond to assuming that the conclusion of the theorem under study is false. As indicated earlier, the test that is used for assignment completion, especially for the determination that a proof has been completed, is the detection of a contradiction. For the set of support strategy in its purest form, (put another way) the program is restricted from applying the chosen inference rules to sets of hypotheses all of which are members of the axioms. By imposing such a restriction, the program is prevented from exploring the underlying theory and, instead, is forced to key recursively on the special hypothesis and the denial of the conclusion. Often, our preference is to instruct the program to recursively key on the special hypothesis alone, using the denial of the conclusion solely to detect that the assignment has been completed. The following simple syntactic example illustrates the use of the set of support strategy.

Let the axioms consist of three clauses: P|Q; –Q|R; –R|S. Let the special hypothesis consist of the single clause –P, and let the conclusion to be proved consist of the single clause S. The denial of the conclusion is, therefore, –S. The search for a proof can begin by focusing mainly on the axioms until the clause P|S is deduced, and then hyperreso-

lution can be used to consider that clause with the special hypothesis and the denial of the conclusion to show that a contradiction has been found. However, if one imagines the case in which the set of axioms is far, far richer, one can easily conjecture that the program might get lost (among a huge set of deduced-and-retained conclusions) and never find a proof. Instead, with the set of support strategy, keying on the special hypothesis, in succession, Q is deduced, then R, then S, which with –S provides the sought-after contradiction.

In contrast to the preceding strategy (which restricts the reasoning of a program), the *resonance strategy* directs the reasoning. With this strategy, the researcher supplies formulas or equations (resonators) that are deemed attractive in the sense that any formula or equation that is similar to a resonator is given preference for driving the program's reasoning, where 'similar' means that there is an exact match if all variables are treated as indistinguishable. To illustrate the use of the resonance strategy, let us consider the following clauses that axiomatize two-valued sentential (or propositional) calculus, where the function *i* denotes **implication**, the function *n* denotes **negation**, and the predicate *P* denotes 'provable.'

```
%  Lukasiewicz 1 2 3.
P (i(i(x,y),i(i(y,z),i(x,z)))).
P(i(i(n(x),x),x)).
P(i(x,i(n(x),y))).
```

If the researcher conjectures that any formula that is similar (in the sense just given) to one of the axioms merits immediate attention, then the three axioms are placed in an appropriate list. Because of being similar to the first of the three axioms – with the use of the resonance strategy – the formula P(i(i(x,x),i(i(y,u),i(x,v)))) will be given prompt consideration for initiating applications of, say, condensed detachment when and if it is deduced and retained.

With either of the two given strategies, the researcher can provide substantial aid to a reasoning program by a judicious choice of, respectively, which clauses to recursively key upon and which to consider heavily and immediately.

Canonicalization and redundancy

Another aspect of automated reasoning that contributes to effectiveness is its ability, when given the appropriate equalities, to automatically canonicalize and simplify information. For example, in the presence of the equality EQUAL(sum(0,x),x), if the program is instructed to do so, new conclusions are automatically rewritten, with subterms of the form sum(0,t) for terms *t* replaced by *t*. The procedure is called *demodulation*. With its use as described, a class of redundant information is purged. In particular, quite similar items are not kept in the many forms that might otherwise be kept. Such is the case for laws that include associativity, where, if so instructed, the program will not retain the many associated forms of a given expression.

Another mechanism, called *subsumption*, is relied upon to purge different identical copies of the same conclusion and, perhaps more important, to purge proper instances of retained conclusions. For example, if a program has retained EQUAL(prod(x,x),e) (for the identity *e*, as in the study of group theory), it will immediately purge through the

use of subsumption items such as EQUAL(prod(prod(y,z),prod(y,z)),e). By taking such an action and others of a more complicated nature, a reasoning program focuses again on generality and avoids emulation (in the sense that a person might retain less general information in the presence of more general).

An intriguing proof

We close this section with an impressively short proof that nicely illustrates: (1) that which an automated reasoning program does well and (2) that which might have eluded fine minds for a long time. The proof is also of value to logic because of focusing on two three-axiom systems for two-valued sentential (or propositional) calculus. The first system (consisting of what Łukasiewicz denotes as theses 19, 37, and 60) was found through automation; the second (consisting of theses 19, 37, and 59) is due to Łukasiewicz himself. Because the two axiom systems share in common two members (theses 19 and 37), what is required (to prove that the set of formulas consisting of 19, 37, and 60 is an axiom system) is a deduction of thesis 59 (whose negation is found as the input clause numbered 6) from theses 19, 37, and 60 (respectively, the input clauses numbered 7 through 9).

When the eminent logician Dana Scott was notified of the following four-step proof, his reaction, by e-mail, was that it might indeed be "a very neat proof that would not be obvious to a human investigator." Scott explained that it is not particularly easy to do unification in one's head – and is he ever right!

A neat proof focusing on the Wos axiom system for two-valued sentential calculus

```
5   [ ]  -P(i(x,y))|-P(x)|P(y).
6   [ ]  -P(i(i(n(p),r),i(i(q,r),i(i(p,q),r))))|$ANS(negation_thesis_59).
7   [ ]  P(i(i(i(x,y),z),i(y,z)))   # label(thesis_19).
8   [ ]  P(i(i(i(x,y),z),i(n(x),z)))   # label(thesis_37).
9   [ ]  P(i(i(u,i(n(x),z)),i(u,i(i(y,z),i(i(x,y),z)))))   # label(thesis_60).
-------------------------------
16  [hyper,5,9,8]   P(i(i(i(x,y),z),i(i(u,z),i(i(x,u),z)))).
23  [hyper,5,16,7]  P(i(i(x,i(y,z)),i(i(i(u,y),x),i(y,z)))).
30  [hyper,5,23,7]  P(i(i(i(x,y),i(i(z,y),u)),i(y,u))).
34  [hyper,5,30,9]  P(i(i(n(x),y),i(i(z,y),i(i(x,z),y))))   # label(thesis_59).
```

Clause (34) contradicts clause (6), and the proof is complete.

3 Significant Successes

We begin the discussion of successes obtained via automation with a success that is of especial satisfaction and significance. The reasons for assigning such importance to it will become clear almost immediately. Where the function i denotes *implication*, the function n denotes **negation**, and the predicate P denotes 'provable,' the success to be discussed first concerns a 23-letter single axiom (the following, found in 1936 by Łukasiewicz) for two-valued sentential (or propositional) calculus.

P(i(i(i(x,y),i(i(i(n(z),n(u)),v),z)),i(w,i(i(z,x)i(u,x))))).

In his 1936 paper (footnote 10), Łukasiewicz suggests how difficult finding proofs of single axioms is. He laments: "Such research is ... so laborious that it cannot be said when, if ever, it will be completed."

What made finding a proof that the given single axiom suffices for two-valued sentential calculus unusually satisfying was the fact that no proof was given by Łukasiewicz, and, from what we can ascertain, no proof was ever published – until the automated reasoning program OTTER was brought into play in mid-1999. In fact, as far as we know, not even a hint was provided in the literature concerning a method for finding such a proof, nor was a hint provided concerning the target for such a proof – although one might surmise that Łukasiewicz had in mind his three-axiom system rather than, say, the axiom system of Hilbert or some other system.

The finding of the desired proof through mechanization can justly be viewed as a reward for adhering to one of the key principles regarding experimentation. In particular, the formulation of a promising methodology demands its testing on difficult problems that are not required to be related to its wellspring. The genesis of the methodology (whose key aspects will be given shortly) was the study of an even shorter single axiom for two-valued sentential calculus, the following.

P(i(i(i(i(i(x,y),i(n(z),n(u))),z),v),i(i(v,x),i(u,x)))).

That axiom was provided by Meredith (1953) 16 years after Łukasiewicz presented his 23-letter axiom and (most likely) in response to an implied Łukasiewicz challenge about finding a single axiom with strictly fewer than 23 letters.

Although Meredith supplied what amounts to a 41-step proof (relying, in effect, on condensed detachment), our goal was to find a means for a reasoning program to produce a proof *without* guidance from the researcher. We had sought such an approach for at least five years, and, in mid-1999, we formulated one that indeed produced a proof, a proof substantially different from Meredith's. When we applied the new methodology to the study of the Łukasiewicz 23-letter axiom, in but four runs, in one afternoon, OTTER produced the first proof we had ever seen, one of length 200 (applications of condensed detachment).

Regarding the methodology and its key aspects, first, it is iterative. It relies on the use of the set of support strategy, adjoining to the appropriate list from run n (to be used in run $n + 1$) results obtained in run n. Although various known axiom systems were admitted as targets to determine that a desired proof had been completed, for our attack on the 23-letter formula, the main target was the Łukasiewicz three-axiom system for this area of logic. The resonance strategy also plays a key role. As for resonators – keeping in mind that we had no clue about the nature of the sought-after proof – we chose to use 68 theses proved by Łukasiewicz, theorems that hold in two-valued sentential calculus. Finally, based on numerous earlier successes, we chose to take an action that one might indeed find counterintuitive, for the action on the surface made the task harder to complete. Specifically, we chose to instruct OTTER to avoid the use of double negation, avoid retaining any new conclusion that contained a term of the form $n(n(t))$ for any term t.

As for properties of the 200-step proof that was found, only eight of its steps are among the 68 theses used as resonators, and only 22 of the 200 steps match one of the 68 resonators (treating all variables as indistinguishable). We include this data in part to address the understandable concern that the researcher may have played an unintentional but key role, in other words, provided much guidance. Such was not the case; we were merely testing the methodology, with, of course, the hope that great fortune would occur; we knew nothing relevant to a possible proof. Double-negation terms are indeed absent. Thus we offer the following open question. Where P and Q may each be collections of formulas, if **T** is a theorem asserting the deducibility of Q from P such that Q is free of double negation, what conditions guarantee that there exists a proof relying solely on condensed detachment all of whose deduced steps are free of double negation?

Among the other successes we obtained through mechanization – some of which were missing for decades – are the following. In infinite-valued sentential calculus, where logical **or** of x and y can be represented with $i(i(x,y),y)$, one can prove that **or** is associative. This area of logic can be axiomatized with the following five formulas (represented as clauses), where the fifth is dependent on the first four.

```
P(i(x,i(y,x))).
P(i(i(x,y),i(i(y,z),i(x,z)))).
P(i(i(i(x,y),(y),i(i(y,x),x))).
P(i(i(n(x),n(y)),i(y,x))).
%  Following is MV5, which is a dependent axiom.
P(i(i(i(x,y),i(y,x)),i(y,x))).
```

As far as we know, until mid-1999, no condensed detachment proof (of associativity) had been reported in the literature. Not only did automation find such a proof – where, before, the only published proofs were not purely axiomatic because they relied partially on reasoning in the metatheory – a more general theorem was proved by OTTER. The generality of the basic mechanisms relied upon by automated reasoning, for example, unification, may have been the primary key to finding the more general result.

Next meriting mention are various distributive laws that hold in infinite-valued sentential calculus. Some forms of distributivity have been proved using a combination of axiomatic and metatheoretic reasoning. But some valid forms have eluded proof of any kind. For example, if we define x **or** y as $i(i(x,y),y)$ and x **and** y as $n(i(i(n(x),n(y)),n(y)))$, then **or** and **and** distribute over each other in infinite-valued logic. This can easily be established semantically, but proving these distributivity laws from the complete set of axioms given earlier for infinite-valued logic (together with the rule of condensed detachment) is something that eluded even Rose and Rosser (1959: 12), who wrote the definitive treatise on infinite-valued logic.

Again, mechanization of proof finding met the test, producing the missing proofs based solely on condensed detachment. For the curious who wonder about the appeal of axiomatic proofs and, even more, of proofs relying on a single inference rule, note that they are often more enlightening and often easier to understand.

Of a different nature are questions focusing on possible axiom dependence. Indeed, a book by Epstein (1994) poses several such questions. Automation has quickly settled

some of Epstein's open questions. In one case (Epstein 1994: 85, problem 12), an appropriate dependence proof was found, when OTTER showed how one of the axioms, the axiom $i(x,i(y,x))$ in Epstein's axiomatization of two-valued sentential calculus, could be proven from the others. Then, appropriate models establishing the independence of the remaining set of axioms were found using William McCune's program MACE, which searches for finite models of sets of clauses.

We close this section by turning to questions concerning proof elegance. More than occasionally a theorem has been proved, but the proof is far, far from elegant. It may be much longer than need be, according to experience and intuition. It may rely on formulas that are extremely complex, and, again, educated opinion suggests such is not required. Among the other inelegant features that may be present is that of requiring the use of some unwanted terms. A program such as OTTER has proved useful in all cited areas, often finding a proof offering far more elegance than can be found in the literature. We are content to cite but one example in this essay.

Meredith (1959) proved (in approximately 38 condensed detachment steps, making extensive use of double negation) axiom MV5 of Łukasiewicz's axioms for infinite-valued sentential calculus from the remaining four axioms. Thus he established a dependence within Łukasiewicz's axiomatization. Because of our emphasis on the avoidance of double negation and the conjecture that its avoidance might enable a reasoning program to find shorter proofs, we embarked on an automated search for a shorter, double-negation-free proof of the dependence of axiom MV5. Success was ours: OTTER produced a 32-step proof, and one in which double negation is absent, thus addressing two elements of increased elegance.

For the student or researcher who might enjoy a question focusing on finding a possibly shorter proof, we suggest the Meredith single axiom. Can one find a proof relying on 40 or fewer applications of condensed detachment showing that the Meredith axiom suffices for all of two-valued sentential calculus? To make the open question more precise, the sought-after proof must complete with the deduction of one of the known axiom systems for that area of logic. If one wishes an open question focusing on the need for double negation, the following might be of interest. In infinite-valued sentential calculus, can one find a proof free of double negation that establishes the deducibility from either the four independent or five dependent axioms for that area of logic of the distributive law P(i(i(n(x),n(i(i(n(y),n(z)),n(z)))),n(i(i(n(i(n(x),y)),n(i(n(x),(n(x),z))),n(i(n(x),z)))))? This question can be rephrased, asking that one prove without the use of double negation that x **or** (y **and** z) **implies** (x **or** y) **and** (x **or** z), where **and** is defined as earlier but **or** is defined as i(n(x),y).

4 Myths, Mechanization, and Mystique

The myths that surround the mechanization of inference rule application and proof finding are many. *Utter pessimism:* effective mechanization is not possible, especially in the context of answering deep, open questions. *Self-worship:* if effective mechanization is possible, emulation of the minds of masters is required. *Uselessness:* one cannot learn from proofs produced from a computer program. The *0/1 myth:* either the program completes the given assignment, or absolutely nothing is produced of value. *Fear:*

reasoning programs will eventually obviate the role of logicians, mathematicians, and the like.

Other than the last given myth (which we shall dispatch shortly), this chapter provides some evidence and some clues that unmask each of the given myths and others unnamed. Indeed, regarding the Utter pessimism myth, the list of open questions (some of which remained open for many decades) that have been answered through heavy reliance on automation is lengthy and continues to grow. As for the Self-worship myth, the more successful and powerful reasoning programs clearly do not emulate person-oriented reasoning. For but two examples, paramodulation (applied so effectively by a computer for equality-oriented reasoning) is the type of inference rule that understandably *is not* used by unaided researchers, and instantiation, which *is* heavily used, is not offered by the type of reasoning program in focus because it appears not to admit effective control.

The Uselessness myth is quickly dispatched. Indeed, although we are far from expert in the areas of logic that we have attacked with OTTER, we do continually learn from the proofs it supplies and, sometimes, from its failures. Even more can, and sometimes is, learned by a master examining the efforts of a reasoning program. Regarding the 0/1 myth, the output file that can be produced may offer a new key lemma and, even better, may contain a proof that the skilled researcher was unable to find unaided. Even if a proof of the desired type is not found, one can study the set of retained conclusions resulting from an unsuccessful attempt and discover precisely what is needed to reach the objective in the next automated attack.

As for the last myth, Fear, it is utter nonsense. The mind of the logician, mathematician, or other scientist will never be replaced, only supplemented! The explanation for the significant contributions to logic and mathematics resulting from the joint effort of program and researcher rests to a great extent with the fact that the general approach (as discussed in this chapter) taken by the more effective reasoning programs differs sharply from that taken by the successful researcher. The two approaches complement each other, and that is the key.

A mystique regarding the automation of reasoning still exists. For but one example, the literature strongly suggests that the proof of numerous deep theorems requires the use of double negation, which in fact is not the case, as shown with the use of OTTER. Is it certain that such success (with the dispensing of double negation) rests with the sharp increase in useful information when compared with total information that is considered? For a second example, who would have thought possible that automated reasoning would yield the answer to a deep question that had remained open since the mid-1930s and that had defied fine minds (that included Tarski)? Specifically, McCune's program EQP – in approximately 10 CPU-days – found a proof showing that every Robbins algebra is a Boolean algebra (McCune 1997).

Part of the mystique, as espoused throughout this chapter, rests with the intense and explicit use of various types of strategy. We strongly conjecture that the successes reported here, as well as numerous others not touched upon, would have been out of reach without the program's reliance on strategy. The formulation of some of the strategies resulted directly from an attempt to answer, through automation, an open question. Because we intend to continue to augment reasoning programs by formulating new strategies and new methodologies, we ask assistance in accruing

new open questions to study. An effective way to convey questions to us is by e-mail: wos@mcs.anl.gov.

Regarding source books and a program that might prove useful and intriguing, two books provide much of what is needed. The first book (Wos and Pieper 1999) serves well as a text, assumes no background, discusses various applications of automated reasoning, offers numerous open questions for consideration, and includes a CD-ROM on which one finds the program OTTER as well as various other useful files. In addition to logic and mathematics, the discussed practical applications include circuit design and validation; the important use of automated reasoning for program verification is not discussed. The second (two-volume) book (Wos and Pieper 2000) consists of reprints of published papers that enable one to follow the development of the field from the early 1960s to the late 1990s. The two books connect in a rather unusual manner: The first contains a long chapter whose subsections each correspond to one of the reprinted papers, giving an overview and appropriate problems. For further information on automated reasoning at Argonne National Laboratory, see http://www.mcs.anl.gov/AR/, which gives all of the needed pointers for new results, for various neat proofs, and for puzzles.

Acknowledgments

This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, US Department of Energy, under Contract W-31-109-Eng-38.

References

Epstein, R. (1994) *The Semantic Foundations of Logic: Propositional Logics*, 2nd edn. New York: Oxford University Press.

Łukasiewicz, J. (1948) The shortest axiom of the implicational propositional calculus. *Proceedings of the Royal Irish Academy*, 52A, 3, 25–33.

Łukasiewicz, J. (1970) Logistic and philosophy. In L. Borkowski (ed.), *Jan Łukasiewicz: Selected Works*. Amsterdam: North-Holland (original work of Łukasiewicz published in 1930).

McCune, W. (1994) *OTTER 3.0 Reference Manual and Guide*. Technical report ANL-94/6. Argonne, IL: Argonne National Laboratory.

McCune, W. (1997) Solution of the Robbins problem. *Journal of Automated Reasoning*, 19, 263–76.

Meredith, C. A. (1953) Single axioms for the systems (C,N), (C,O), and (A,N) of the two-valued propositional calculus. *Journal of Computing Systems*, 1, 155–64.

Meredith, C. (1959) The dependence of an axiom of Łukasiewicz. *Transactions of the American Mathematical Society*, 87, 54.

Rose, A. and Rosser, J. B. (1959) Fragments of many-valued statement calculi. *Transactions of the American Mathematical Society*, 87, 1–53.

Wos, L. and Pieper, G. W. (1999) *A Fascinating Country in the World of Computing: Your Guide to Automated Reasoning*. Singapore: World Scientific.

Wos, L. and Pieper, G. W. (2000) *Collected Works of Larry Wos*. Singapore: World Scientific.

Further Reading

We recommend the following, listed in order of importance:

Wos, L. (1996) *The Automation of Reasoning: An Experimenter's Notebook with OTTER Tutorial*. New York: Academic Press.

Wos, L. (1993) The kernel strategy and its use for the study of combinatory logic. *Journal of Automated Reasoning*, 10, 287–343.

Wos, L. (1995) Searching for circles of pure proofs. *Journal of Automated Reasoning*, 15, 279–315.

Boyer, R. S. and Moore, J. S. (1998) *A Computational Logic Handbook*, 2nd edn. New York: Academic Press.

McCune, M. and Padmanabhan, R. (1996) *Lecture Notes in Computer Science*, vol. 1095: *Automated Deduction in Equational Logic and Cubic Curves*. New York: Springer.

45

A Computational Logic for Applicative Common LISP

MATT KAUFMANN AND J. STROTHER MOORE

1 Introduction

Perhaps one of the most ambitious goals for mathematical logic was put forth by one of its earliest advocates.

> If we had some exact language . . . or at least a kind of truly philosophic writing, in which the ideas were reduced to a kind of alphabet of human thought, then all that follows rationally from what is given could be found by a kind of calculus, just as arithmetical or geometrical problems are solved. (Leibniz, 1646–1716)

Mathematical logic casts too harsh a light to be appropriate for the 'rationalization' of many human endeavors. Can one axiomatize good and evil, or even the aerodynamics of the African sparrow, so that all that follows by mathematical proof is truly believable?

But Leibniz' dream was aided immeasurably by the invention of the digital computer because the computer not only provided a platform on which to build a reasoning engine but provided a source of problems to tackle with it.

> Instead of debugging a program, one should prove that it meets its specifications, and this proof should be checked by a computer program. (John McCarthy, "A Basis for a Mathematical Theory of Computation," 1961)

Computing systems, such as microprocessors, switches, file servers, compilers, encryption devices, control programs, financial software, etc., are naturally described in the precise language of mathematical logic. If the logical 'model' of the system accurately describes what is built, then the logical properties of the model accurately predict the behavior of the artifact.

But is proving theorems about computing systems practical? Is it cost effective? Here are two more quotations that shed some light on those questions.

> An elusive circuitry error is causing a chip used in millions of computers to generate inaccurate results. (*NY Times*, "Circuit Flaw Causes Pentium Chip to Miscalculate, Intel Admits," November 11, 1994)

> Intel Corp. last week took a $475 million write-off to cover costs associated with the divide bug in the Pentium microprocessor's floating-point unit. (*EE Times*, January 23, 1995)

It is possible to prove a lot of theorems for $475 million.

'ACL2' stands for 'A Computational Logic for Applicative Common Lisp.' It is the name of a programming language, a first-order mathematical logic based on recursive functions, and a mechanical theorem prover for that logic. ACL2 is designed for use in reasoning about computing systems, both those implemented in hardware and those implemented in software.

The human user of ACL2 can formalize or model a computing system by defining functions that simulate the operation of the system. Since ACL2 is a programming language, such an operational model is just a computer program that can be run on concrete data to produce concrete results. With this program the user might test the behavior of the system on some finite number of example inputs. Since ACL2 is also a mathematical logic, the user might prove theorems about the model, possibly establishing properties that hold for an infinite number of inputs. Finally, using ACL2's interactive theorem proving program, the user might check these proofs mechanically, thereby eliminating the all-too-frequent errors that crop up in 'hand proofs.'

This is not just a mathematical fantasy. For example, ACL2 was used to prove the correctness of the circuitry implementing the elementary floating point operations on the AMD Athlon™ processor[1] with ACL2. Most major chip manufacturers have personnel devoted to proving theorems or otherwise formally checking properties of their designs.

In this article we describe ACL2 briefly, present a simple modeling problem and its solution in ACL2, and describe some of ACL2's recent applications.

We assume the reader has had a little experience with computing and programming. Also helpful would be an introductory course in first order predicate calculus.

2 The ACL2 System

Here we briefly discuss ACL2 as a programming language, a logic, and a mechanical theorem prover or proof checker. The ACL2 system is available under the GNU General Public License and without fee from its home page, `http://www.cs.utexas.edu/users/moore/acl2`. Installation instructions and documentation are included. We discuss how to learn to use ACL2 in Section 5.

ACL2 is just one of several mechanical theorem proving programs used for hardware and software verification. Among the others are HOL (Gordon and Melham 1993), Otter (McCune 1994), and PVS (Owre et al. 1992). See the *Related Web Sites* link under the *Books and Papers* link of the ACL2 home page for lists of dozens of other theorem provers. Theorem provers are still research vehicles, even though some, like the ones mentioned above, are being used by researchers in industry. Each is designed to explore a different part of the theorem proving problem. ACL2 is first order with considerable automation, with heuristics tailored to recursive definitions and induction. Otter supports first-order predicate calculus, with full support for quantification. HOL and PVS both support higher-order logics. Of major concern to the developers of HOL

was how to build a theorem prover that was both user-extensible and sound. The PVS and ACL2 developers were primarily concerned with building tools that people without research backgrounds in automated reasoning could use off the shelf to prove theorems about computing systems. The Otter team focused on finding automatic proof techniques so that Otter, rather than its human users, gets full credit for its proofs. We should emphasize, however, that all of these tools address all of these issues to varying degrees. For example, ACL2 addresses user extensibility, and Otter requires the user to interact by setting parameters.

The programming language

As a programming language, ACL2 is a variant of Common Lisp (Steele 1990). 'Lisp,' which stands for 'list processing,' is commonly used for artificial intelligence applications because it facilitates symbol manipulation. Lisp was invented by John McCarthy in the late 1950s as part of his visionary project towards a mechanized theory of computation (McCarthy 1960, 1962, 1963).

ACL2 is a *functional* or *applicative* version of Lisp, meaning that ACL2 programs are mathematical functions of their arguments. They do not have side-effects and are not sensitive to implicit 'global variables' or implicit 'state.'

ACL2 terms are written in prefix notation. A term is a variable, a constant, or the application of a function symbol, f, of k arguments to k terms, $a_1, \ldots, a_k$, written $(f\ a_1 \ldots a_k)$. Here is how one might write $a^2 + ab$ in ACL2: `(+ (expt a 2) (* a b))`. The ACL2 runtime system provides facilities for calculating the values of terms under assignments of values to their free variables. For example, if `a` has the value `3` and `b` has the value `5`, then the term above is calculated to have the value `24`.

In addition to the numbers (integers, rationals, and complex numbers with rational coefficients) ACL2 supports several other data types. These include strings (such as `"Hello World"`), symbols (such as `LOAD` and `X`), and ordered pairs. Primitive functions are provided for manipulating each type of data. For example, the function `cons` takes two arguments and returns an ordered pair containing them. The functions `car` and `cdr` take one argument, which is normally an ordered pair, and return the first and second components, respectively. The function `consp` takes one argument and returns the constant `T` ('true') if the argument is an ordered pair and `NIL` ('false') otherwise.

Ordered pairs are written in parenthesized 'dot notation.' For example, the pair traditionally written as $\langle 3, \mathtt{NIL} \rangle$ is written in ACL2 as `(3 . NIL)`.

Ordered pairs can be used to encode a wide variety of abstractions. One such abstraction is linear lists, which are so common that the notation for printing ordered pairs in ACL2 (and Lisp) is oriented towards it. The constant `NIL` may be written simply as `()`, and thus plays double duty; it is used both as the false truth-value and as the empty list. The ordered pair `(3 . NIL)` may be written simply as `(3)`. The ordered pair `(2 . (3 . NIL))` may be written `(2 3)`, the ordered pair `(1 . (2 . (3 . NIL)))` may be written `(1 2 3)`, etc.

It is convenient to be able to write list constants inside terms. What is a term that evaluates to (i.e. whose meaning is) the list `(1 2 3)`? One such term is `(cons 1 (cons 2 (cons 3 NIL)))`. But another one is `'(1 2 3)`. The 'quote mark' can be used to write a term that evaluates to a given constant.

Lists are frequently used to represent still other abstractions. For example, the list `(1 2 3)` may be thought of as the *stack* obtained by pushing `1` onto the stack `(2 3)`. The following *function definitions* make these conventions easier to remember. (Actually, the symbols `push` and `pop` are defined in ACL2 and may not be redefined; to make these definitions we must actually operate in a symbol *package* other than the default one, but we do not discuss that here.)

```
(defun push    (item stack)    cons item stack))
(defun top     (stack)         (car stack))
(defun pop     (stack)         (cdr stack))
```

The first `defun` above, for example, defines `push` to be a function symbol of two arguments, `item` and `stack`, whose value is obtained by evaluating the term `(cons item stack)`. Thus, `(push 1, '(2 3))` is the stack `(1 2 3)`. The `top` of that stack is `1` and the `pop` is `(2 3)`.

Here is another common use of lists. Consider the list of two ordered pairs `((A . 7) (B . 4))`. Call this constant α. The `car` of α is `(A . 7)`. The `cdr` of α is `((B . 4))`. The `car` of the `cdr` of α is `(B . 4)`. Lists such as α are thought of as *tables* that map keys (`A` and `B`, in this case) to values (`7` and `4`, respectively). Such lists are called *association lists* or *alists* or *assignments*. The `car` of the `car` of a nonempty alist is the first key assigned in the alist; the `cdr` of the `car` is the value assigned to that key. The `cdr` of a nonempty alist is another alist that assigns the rest of the symbols.

Here is a function that looks up the value of a symbol in an alist. This function is recursive.

```
(defun lookup (sym alist)
(if (consp alist)
(if (equal sym (car (car alist)))
    (cdr (car alist))
    (lookup sym (cdr alist)))
0))
```

The definition above may be paraphrased as: If `alist` is a `cons` pair, then if `sym` is the first key assigned, return its value; otherwise `lookup sym` in the rest of `alist`. If `alist` is not a `cons` pair, return `0`.

After this definition, `(lookup 'B '`α`)` evaluates to `4`. But `(lookup 'C '`α`)` evaluates to `0`.

ACL2 supports a variety of syntactic extensions. Another way to define lookup is shown below.

```
(defun lookup (sym alist)
  (cond
    ((endp alist) 0)
    ((equal sym (car (car alist)))
     (cdr (car alist)))
    (t (lookup sym (cdr alist)))))
```

The cond special form is just a nest of ifs. `(Endp alist)` is equivalent to `(not (consp alist))`.

For a more thorough introduction to ACL2 as a programming language see Kaufman et al. (2000b). The link labeled *Hyper-Card* on the ACL2 home page contains a quick introduction to Lisp and a reference card to the programming language. The *User's Manual* link contains several megabytes of hypertext documentation.

The Logic

ACL2 is formalized as a first-order mathematical logic. Any standard formulation of first-order logic will serve our purposes. See also Kaufmann and Moore (to appear). Axioms describe the primitive functions. For example, here are several of the axioms. (Actually, `(consp NIL) = NIL` is not an axiom but an easily proved theorem.)

Axioms

```
x = NIL → (if x y z) = z.
x ≠ NIL → (if x y z) = y.
(consp NIL) = NIL
(consp (cons x y)) = t
(car (cons x y)) = x
(cdr (cons x y)) = y
```

Using the natural numbers and ordered pairs, a representation of the ordinals up to ε_0 is introduced. For example, the ordinal $\omega^2 + \omega \times 4 + 3$ is represented in ACL2 by the list `(2 1 1 1 1 . 3)`. An axiom defines a relation (a function returning `T` or `NIL`), named `e0-ord-<`, corresponding to the well-founded ordering relation on these ordinals. Another axiom introduces the predicate, `e0-ordinalp`, which recognizes the ACL2 ordinals.

The principle of mathematical induction, in ACL2, is then stated as a rule of inference that allows induction up to ε_0. To prove a conjecture by induction one must identify some ordinal-valued measure function. The induction principle permits one to assume inductive instances of the conjecture being proved, provided the instance has a smaller measure according to the chosen measure function.

Finally, a principle of definition is provided, by which the user can extend the axioms by the addition of equations defining new function symbols. To admit a new recursive definition, the principle requires the identification of an ordinal measure function and a proof that the arguments to every recursive call decrease according to this measure. Only terminating recursive definitions can be so admitted under the definitional principle. ('Partial functions' can be axiomatized; see Maniolos and Moore (2000).)

The successful admission of a definition adds a new axiom. For example, the definition of `push` above adds

Axiom

```
(push item stack) = (cons item stack).
```

The two variables are (implicitly) universally quantified.

The measure used to justify the recursive function `lookup` is `ac12-count`. The `ac12-count` of a natural number is that number. The `ac12-count` of an ordered pair is one more than sum of the counts of the `car` and the `cdr`. `Ac12-count` always returns a natural number.

As is customary in formal treatments of mathematical logic, from such basics a variety of other rules of inference are derived to make proofs more practical. A more thorough treatment of the logic is presented in Kaufmann et al. (2000b: Chapter 6). The solutions to the exercises for that chapter (see the *Books and Papers* on the home page and follow the obvious links) contain formal proofs of many elementary theorems and a sketch of how more elaborate rules of inference are justified. See also Kaufmann and Moore (1997).

The theorem prover

The ACL2 theorem prover is a symbolic manipulation engine driven from a collection of rules in a database. The user determines the available rules, but in an indirect way. The rules are derived from theorems posed as challenges by the user and proved by the system. Thus, the logical soundness of the theorem prover cannot be imperiled by the user. But the strategy employed by the theorem prover can be largely determined by the experienced user who understands how rules are derived from theorems and what the effects of those rules are. The user may also direct the system to read in all the rules in previously certified 'books,' thereby enabling the sharing of results in the ACL2 community. In addition, the user may supply hints to affect the system's decisions and may specify low-level proof steps via an interactive loop.

The system has many heuristics for determining its behavior. For example, heuristics determine when it expands recursive function definitions, when it inducts, and what induction hypotheses it assumes.

The system also contains many decision procedures and other high-level derived rules of inference. For example, it can use a BDD procedure (Bryant 1992) to recognize propositional tautologies, it has built in knowledge about linear arithmetic inequalities (Boyer and Moore 1997), and it can use calculation to compute the values of functions on constants.

The system prints a description of its evolving 'proof' as it proceeds. It does not produce a formal proof, but when it says 'Q.E.D.' we, the authors of ACL2, believe that the computation it did is sufficient to guarantee the existence of a formal proof in the logic described. The system often fails, either by abandoning the proof attempt or running until the user aborts the attempt. In either case, it is up to the human user to 'fix' the situation, by reformulating the conjecture to prove or the hints provided, or by further developing a database of rules.

The ACL2 theorem prover is an improved version of the Boyer–Moore theorem prover, Nqthm (Boyer and Moore 1979, 1997; Boyer et al. 1995), adapted to applicative Common Lisp. For more details of how it works, see Kaufmann et al. (2000b). We illustrate it in the section entitled "Sample Output" below.

How good is ACL2's theorem prover? That is, how automatic is it? At one level that depends on how good a database of rules it has and whether the conjecture at hand

falls in the class of formulas handled by that set of rules. But perhaps the intent of the question is deeper. How far away from its rules can it operate successfully? How deep are its proofs? The answer depends on whether you view the question from the perspective of the logician or the more traditional mathematician, who have very different ideas of what the word 'proof' means. Logicians think of proofs as sequences or trees of formulas, expressed in a precisely defined syntax and related to one another by a precisely defined set of inference rules. Most mathematicians think of proofs as informal but convincing arguments. The logician might very well consider ACL2 an automatic theorem prover because it is not always obvious how to construct formal proofs of some of the theorems it proves automatically. But the mathematician would probably think of ACL2 as a proof checker, at best. The mathematician would find virtually everything ACL2 proves automatically to be 'self-evident' or 'obvious' from the theorems and definitions ACL2 had previously been led to accept. To the mathematician, ACL2 is not so much *finding* a proof as it is *checking* one presented to it by the human. This will become more obvious in Section 3.

3 A Modeling Problem

In this section we will deal with a simple variant of a classic example in the verification literature, first done 'by hand' in McCarthy and Painter (1967) and by machine with the Boyer–Moore prover in Boyer and Moore (1979). We will model an assembly language for a push-down stack machine, formalize a simple arithmetic expression language, implement a compiler that translates from arithmetic expressions to assembly code, and prove the compiler is correct. In addition to illustrating the formalization of some central ideas in computing – state machines, language semantics, compilation – this example is appropriate because it deals with a few of the same issues that arise in model theory, for example the assignment of meaning to the sentences of a formal language.

The assembly language

An *instruction* is a nonempty list. The *opcode* is the first element of the list. Some instructions have an *operand*, which is the second element.

```
(defun opcode (inst) (car inst))
(defun operand (inst) (car (cdr inst)))
```

The opcodes on our machine and their informal semantics are: `(LOAD` *var*`)` pushes the value of *var* onto the stack, `(PUSH` *c*`)` pushes the constant *c* onto the stack, `(DUP)` duplicates the top of the stack, `(ADD)` pops two items off the stack and pushes their sum, and `(MUL)` pops two items off the stack and pushes their product. We formalize this with the function `step`, which takes an instruction to execute, an alist giving the variable values, and a stack; the function returns the new value of the stack.

```
(defun step (inst alist stk)
  (let ((op (opcode inst)))
    (cond
   ((equal op 'LOAD)
    (push (lookup (operand inst) alist) stk))
   ((equal op 'PUSH)
    (push (operand inst) stk))
   ((equal op 'DUP)
    (push (top stk) stk))
   ((equal op 'ADD)
    (push (+ (top (pop stk)) (top stk))
          (pop (pop stk))))
   ((equal op 'MUL)
    (push (* (top (pop stk)) (top stk))
          (pop (pop stk))))
(t stk))))
```

A *program* is a sequence of instructions. They are executed sequentially with a given alist and some initial stack. The final stack is returned.

```
(defun m (program alist stk)
  (cond ((endp program) stk)
        (t (m (cdr program)
              alist
              (step (car program) alist stk)))))
```

The function `m` formalizes the semantics of this simple programming language. For example,

```
(m '((LOAD A) (DUP) (ADD))
   '((A . 7) (B . 4))
   '(1 2 3))
```

'simulates' the execution of a program that pushes the value of `A`, duplicates it, and adds the two values together. It does so in an environment in which the value of `A` is 7 and the value of `B` is `4`. The initial stack is `(1 2 3)`, a stack with 1 on top. The result of this execution is the stack `(14 1 2 3)`.

An expression language

An *expression* (and its *value* under an assignment) is a variable symbol (whose value is specified by the assignment), a numeric constant (which is its own value), or a list of one of the following forms (where the $expr_i$ are expressions): `(INC` $expr_1$`)` (whose value is one more than that of $expr_1$), `(SQ` $expr_1$`)`, (whose value is the square of that of $expr_1$), `(`$expr_1$ + $expr_2$`)` (whose value is the sum of those of the two

subexpressions), or $(expr_1 * expr_2)$ (whose value is the product of those of the two subexpressions).

We can formalize this as follows.

```
(defun eval (x alist)
  (cond
  ((atom x)
   (cond ((symbolp x) (lookup x alist))
         (t x)))
  ((equal (fn x) 'INC)
   (+ 1 (eval (argl x) alist)))
  ((equal (fn x) 'SQ)
   (* (eval (argl x) alist)
      (eval (argl x) alist)))
  ((equal (fn x) '+)
   (+ (eval (argl x) alist)
      (eval (arg2 x) alist)))
  ((equal (fn x) '*)
   (* (eval (arg1 x) alist)
      (eval (arg2 x) alist)))
  (t 0)))
```

where

```
(defun fn (expr)
  (if (equal (len expr) 2) (car expr) (car (cdr expr))))
(defun arg1 (expr)
  (if (equal (len expr) 2) (car (cdr expr)) (car expr)))
(defun arg2 (expr)
  (car (cdr (cdr expr)))).
```

Eval formalizes the semantics of this expression language. We can test it. For example, here is a transcript showing that the `eval` of a certain expression is equal to `400`.

```
COMP !> (eval '(SQ (INC (A + (3 * B)))) '((A. 7) (B. 4)))
400
COMP !>
```

A compiler

A compiler is a translator from one language to another. We will compile arithmetic expressions, as above, into our assembly language. The goal is to produce a program that, when executed under a given assignment, will push the value of the expression on the stack. The method is straightforward. To compile a product, say, we concatenate

the compiled code for the two subexpressions and then generate an `(MUL)` instruction to pop the two intermediate values off the stack and push their product. To compile `(SQ` *expr*$_1$`)`, we will compile the subexpression and then generate a `(DUP)` followed by an `(MUL)`. The others are similar. Here is the compiler.

```
(defun compile (x)
  (cond
   ((atom x)
    (cond
     ((symbolp x) (list (list 'LOAD x)))
     (t (list (list 'PUSH x)))))
   ((equal (fn x) 'INC)
    (append (compile (arg1 x))
            '((PUSH 1) (ADD))))
   ((equal (fn x) 'SQ)
    (append (compile (arg1 x))
            '((DUP) (MUL))))
   ((equal (fn x) '+)
    (append (compile (arg1 x))
            (compile (arg2 x))
            '((ADD))))
   ((equal (fn x) '*)
    (append (compile (arg1 x))
            (compile (arg2 x))
            '((MUL))))
   (t (list (list 'PUSH 0)))))
```

`Append` concatenates its arguments. We illustrate the compiler below.

Specification

The output of compile on the expression `(SQ (INC (A + (3 * B))))` is the program shown below.

```
COMP ! > (compile '(SQ (INC (A + (3 * B)))))
((LOAD A)
 (PUSH 3)
 (LOAD B)
 (MUL)
 (ADD)
 (PUSH 1)
 (ADD)
 (DUP)
 (MUL))
COMP ! >
```

This program is 'correct' in the sense that executing it leaves the value of the given expression on top of the stack.

The *specification* of compile is that it produces correct programs for every expression. A formalization of this claim is `(equal top (m (compile x) a s)) (eval x a))`. We will name this conjecture `main`.

Mechanical proof

All of the definitions involved in the formalization above are automatically admitted by the mechanical theorem prover. `Ac12-count` is the only measure needed and the system 'guesses' that.

If we then submit `main` as a challenge conjecture, the ACL2 theorem prover runs for 11 seconds (on a 731 MHz Pentium III) and gives up. Inspection of the proof attempt using 'The Method,' described in Kaufmann et al. (2000b) and in the on-line ACL2 manual, produces the following insights. First, the proof will clearly involve induction on the form of the expression `x`. Second, `main` is not strong enough to prove by induction. We must prove the conjecture that says 'execution of the compiled code *pushes* the value of the expression onto the pre-existing stack (leaving the other items there intact).' Our `main` does not insure that other intermediate values are not removed and hence cannot be used to explain how the compiler works. Note that it is a common mathematical trick to generalize a conjecture before doing proof by induction, and although ACL2 provides a little support for making such generalizations automatically, it is generally up to the user to do so.

The stronger conjecture is `(equal (m (compile x) a s) (push (eval x a) s))`, which, if proved, clearly implies `main`. We name this conjecture `lemma`. The attempt to prove `lemma` fails in about 6 seconds. Inspection of the failed proof reveals that ACL2 chose an inadequate induction scheme. Consider the case for compiling a sum expression. The theorem prover inductively assumes `lemma` for both subexpressions. But it is obvious to the human user that the second induction hypothesis (that for the second argument of the sum) must use the instance in which the stack `s` is the pre-existing with one more thing pushed onto it: the value of the first subexpression.

Induction schemes are described to ACL2 by defining recursive functions that instantiate their arguments appropriately. Here is the necessary definition, which is admitted automatically.

```
(defun hintfn (x a s)
  (cond
   ((atom x) (list x a s))
   ((equal (fn x) 'INC)
    (hintfn (arg1 x) a s))
   ((equal (fn x) 'SQ)
    (hintfn (arg1 x) a s))
   ((equal (fn x) '+)
    (cons (hintfn (arg1 x) a s)
          (hintfn (arg2 x) a (push (eval (arg1 x) a) s))))
   ((equal (fn x) '*)
```

```
   (cons (hintfn (arg1 x) a s)
         (hintfn (arg2 x) a (push (eval (arg1 x) a) s))))
  (t (list x a s))))
```

The value of this function is irrelevant. What matters is the case analysis it does and the way it instantiates its arguments in recursion.

If we then tell ACL2 to prove `lemma`, advising it to induct the way `hintfn` recurs, the proof attempt again fails. Inspection reveals that the system must be able to simplify `(m (append x y) a s)`. This is obvious in retrospect: the compiler concatenates two recursively obtained code sequences and we must know how the machine deals with concatenated programs. The obvious relationship is given in the theorem below.

```
(defthm composition
  (equal (m (append x y) a s)
         (m y a (m x a s))))
```

This is how the user actually submits a challenge to the theorem prover. The formula above *alleges* that the execution of the concatenation of program `x` followed by program `y` is equal to the execution of program `y` starting with the stack produced by the execution of program `x`. If ACL2 can prove this, it will build it in as a rewrite rule (by default) and name the theorem `composition`. In fact, the system successfully proves `composition`, by induction on `x` and simplification. We show the output under "Sample output" below.

The system can now prove `lemma`, and can then use it to prove `main`. The two commands, in full, are shown below.

```
(defthm lemma)
  (equal (m (compile x) a s)
         (push (eval x a) s))
  :hints (("Goal" :induct (hintfn x a s))))
(defthm main
  (equal (top (m (compile x) a s))
         (eval x a)))
```

The total amount of time to replay the entire successful proof sequence (including the admission of all of the definitions) is approximately 2 seconds. All the necessary user input has been exhibited here. An experienced ACL2 user might well have recognized the importance of `composition` and `lemma` from the outset and would thus have stated them as part of the initial proof plan. We mention the discovery process because it is important in more complicated proofs where all the necessary lemmas are rarely recognized in advance.

Sample output

Here is the output of the theorem prover on the `composition` theorem.

```
COMP ! > (defthm composition
                 (equal (m (append x y) a s)
                        (m y a (m x a s))))
```

Name the formula above *1.
Perhaps we can prove *1 by induction. Three induction schemes are suggested by this conjecture. These merge into two derived induction schemes. However, one of these is flawed and so we are left with one viable candidate.

We will induct according to a scheme suggested by (M X A S), but modified to accommodate (APPEND X Y). If we let (:P A S X Y) denote *1 above then the induction scheme we'll use is

```
(AND (IMPLIES (AND (NOT (ENDP X))
                   (:P A (STEP (CAR X) A S) (CDR X) Y))
              (:P A S X Y))
     (IMPLIES (ENDP X) (:P A S X Y))).
```

This induction is justified by the same argument used to admit M, namely, the measure (ACL2-COUNT X) is decreasing according to the relation EO-ORD-< (which is known to be well-founded on the domain recognized by EO-ORDINALP). Note, however, that the unmeasured variable S is being instantiated. When applied to the goal at hand the above induction scheme produces the following two nontautological subgoals.
Subgoal *1/2

```
(IMPLIES (AND (NOT (ENDP X))
              (EQUAL (M (APPEND (CDR X) Y)
                        A (STEP (CAR X) A S))
                     (M Y A (M (CDR X) A (STEP (CAR X) A S)))))
         (EQUAL (M (APPEND X Y) A S)
                (M Y A (M X A S)))).
```

By the simple :definition ENDP we reduce the conjecture to
Subgoal *1/2'

```
(IMPLIES (AND (CONSP X)
              (EQUAL (M (APPEND (CDR X) Y)
                        A (STEP (CAR X) A S))
                     (M Y A (M (CDR X) A (STEP (CAR X) A S)))))
         (EQUAL (M (APPEND X Y) A S)
                (M Y A (M X A S )))).
```

But simplification reduces this to T, using the :definitions BINARY-APPEND, M, OPCODE, OPERAND, POP, PUSH, STEP and TOP, the :executable-counterpart of EQUAL, primitive type reasoning and the :rewrite rules CAR-CONS, COR-CONS, COMMUTATIVITY-OF-* and COMMUTATIVITY-OF-+.
Subgoal *1/1

```
(IMPLIES (ENDP X)
         (EQUAL (M (APPEND X Y) A S)
                (M Y A (M X A S )))).
```

By the simple :definition ENDP we reduce the conjecture to
Subgoal *1/1'

```
(IMPLIES (NOT (CONSP X))
         (EQUAL (M (APPEND X Y) A S)
                (M Y A (M X A S )))).
```

But simplification reduces this to T, using the :definitions BINARY-APPEND and M and primitive type reasoning.
That completes the proof of *1.
Q. E. D.
Summary

```
Form:   ( DEFTHM COMPOSITION . . .)
Rules: ((:DEFINITION BINARY-APPEND)
        (:DEFINITION ENDP)
        (:DEFINITION M)
        . . . material deleted . . .
        (:REWRITE CDR-CONS)
        (:REWRITE COMMUTATIVITY-OF-*)
        (:REWRITE COMMUTATIVITY-OF-+))
Warnings: None
Time: 0.09 seconds (prove: 0.06, print: 0.02, other: 0.01)
COMPOSITION
```

4 Case Studies

The compiler example illustrates two different models, a theorem relating them, and the role of the user in structuring ACL2's proofs by the discovery of appropriate lemmas. At `http://www.cs.utexas.edu/users/moore/publications/flying-demo/scipt.html` you will find this example and many others, including the correctness of an insertion sort function, the correctness of a binary adder and of a multiplier, the formal semantics of a simple netlist description language – a language like that used to describe circuits – and the correctness of a function that generates a description of an adder, and some theorems abut Java byte code programs. The web pages show the definitions, many example computations, and most of the proofs, including all of the proofs for the compiler example discussed here.

These models are suggestive of how ACL2 is used. But they are trivial by the standards of industrial machine designs and realistic programming languages. The stack machine above is fully specified in about two dozen lines of code; the proof required two lemmas. Industrial applications of ACL2 have involved hundreds of pages of code to formalize a single model and thousands of lemmas to relate two such models.

We now briefly describe a few such applications. For more details, see Kaufmann et al. (2000a) and Kaufmann and Moore (2000), collections of case studies written by ACL2 users.

ACL2 has been used to model several industrial microprocessors. The models are similar to that for `m`: a 'state' is formalized as an n-tuple of various components like stacks, registers, etc., and a state-transition function, `step`, is defined. The Motorola CAP digital signal processor (DSP) (Brock et al. 1996; Brock and Hunt 1997; Gilfeather et al. 1994) was modeled at two levels: the pipeline level, where several instructions are simultaneously being decoded and carried out; and the user level, where instructions are executed sequentially. Both models were bit- and cycle-accurate in the sense that they specified all the state components completely on every step. The two models were shown equivalent under certain conditions on the program being executed. Another commercial microprocessor modeled with ACL2 is the Rockwell JEM1 (Greve and Wilding 1998; Wilding et al. to appear) – the world's first silicon Java Virtual Machine. ACL2 has been used to verity commercial DSP (Brock and Moore 1999) microcode. It has been used to prove the IEEE compliance of the `FDIV` microcode for the AMD-K5™ processor[2] (Moore et al. 1998) and of the circuit descriptions implementing each of the elementary floating-point operations on the AMD Athlon (Russinoff 1998; Russinoff and Flatau 2000). It has been used to verify a pipelined machine providing interrupts and exceptions in the face of speculative out-of-order execution (Sawada and Hunt 1998) and a security model for the boot code of the IBM 4758 (Smith and Austel 1998).

Not all of ACL2's applications are at the hardware level. ACl2 is being used to prove properties of Java byte code (Moore 1999; Moore and Porter 2000a, 2000b), including multi-threaded programs.

ACL2 has been used to provide a trusted (verified) proof-checker for the Otter theorem proving system (McCune and Shumsky 2000). Otter is perhaps the preeminent resolution-style theorem prover and has been under development at Argonne National Labs for decades. When Otter claims success, it can give its proof to a much simpler theorem prover for checking, and one such checker was verified to be sound for finite models by ACL2. In a similar kind of work, ACL2 was used to verify a checker for an off-line compiler for safety-critical train-borne real-time control software (Bertoli and Traverso 2000).

ACL2 has been used to prove the correctness of a model checker (Manolios 2000), the alternating-bit protocol (Manolios et al. 1999), a BDD package (Sumners 2000), and many other algorithms.

An extension of the system by Ruben Gamboa (1999) adds the real numbers via nonstandard analysis and many interesting theorems in real analysis have been proved including trigonometric identities, Euler's identity, the fundamental theorem of calculus (Kaufmann 2000) and theorems about continuity and differentiability (Gamboa 2000). See also Gamboa and Kaufmann (1999).

The ACL2 home page contains links to many other papers reporting ACL2 applications.

Notes

1 AMD, the AMD logo, and combinations thereof, and AMD Athlon are trademarks of Advanced Micro Devices, Inc.
2 AMD, the AMD logo, and combinations thereof, and AMD-K5 are trademarks of Advanced Micro Devices, Inc.

References

Bertoli, P. and Traverso, P. (2000) Design verification of a safety-critical embedded verifier. In Kaufmann et al. (2000) pp. 233–46.

Boyer, R. S. and Moore, J. S. (1979) *A Computational Logic*. New York: Academic Press.

Boyer, R. S. and Moore, J. S. (1988) Integrating decision procedures into heuristic theorem provers: A case study of linear arithmetic. In *Machine Intelligence 11* (pp. 83–124). Oxford: Oxford University Press.

Boyer, R. S. and Moore, J. S. (1997) *A Computational Logic Handbook*. 2nd edn. New York: Academic Press.

Boyer, R. S., Kaufmann, M. and Moore, J. S. (1995) The Boyer–Moore theorem prover and its interactive enhancement. *Computers and Mathematics with Applications*, 5(2), 27–62.

Brock, Bishop and Warren Hunt, A. Jr. (1997) Formally specifying and mechanically verifying programs for the Motorola complex arithmetic processor DSP. In *1997 IEEE International Conference on Computer Design* (pp. 31–6). IEEE Computer Society, October.

Brock, B. and Moore, J. S. (1999) A mechanically checked proof of a comparator sort algorithm. Submitted for publication.

Brock, B., Kaufmann, M. and Moore, J. S. (1996) ACL2 theorems about commercial microprocessors. In M. Srivas and A. Camilleri (eds.), *Formal Methods in Computer-Aided Design (FMCAD'96)* (pp. 275–93). New York: Springer-Verlag, LNCS 1166, November.

Bryant, R. E. (1992) Symbolic Boolean manipulation with ordered binary decision diagrams. *ACM Computing Surveys*.

Gamboa, R. (1999) Mechanically verifying real-valued algorithms in ACL2. PhD thesis, University of Texas at Austin.

Gamboa, R. (2000) Continuity and differentiability. In Kaufmann et al. (2000) pp. 301–17.

Gamboa, R. and Kaufmann, M. (forthcoming) Non-standard analysis in ACL2. *Journal of Automated Reasoning*.

Gilfeather, S., Gehman, J. and Harrison, C. (1994) Architecture of a complex arithmetic processor for communication signal processing. In *International Symposium on Optics, Imaging, and Instrumentation, 2296, Advanced Signal Processing: Algorithms, Architectures, and Implementations V* (pp. 624–5). SPIE.

Gordon, M. and Melham, T. (1993) *Introduction to HOL: A Theorem Proving Environment for Higher Order Logic*. Cambridge University Press.

Greve, D. A. and Wilding, M. M. (1998) Stack-based Java a back-to-future step. *Electronic Engineering Times*, Jan. 12, p. 92.

Kaufmann, M. (2000) Modular proof: The fundamental theorem of calculus. In Kaufmann et al. (2000), pp. 75–92.

Kaufmann, Matt and Moore, J. (1997) A precise description of the ACL2 logic. In http://www.cs.utexas.edu/users/moore/publications/km97.a.ps.Z. Department of Computer Sciences, University of Texas at Austin.

Kaufmann, M. and Moore, J. S. (2000) *Proceedings of ACL2 Workshop 2000*. Department of Computer Sciences, Technical Report TR-00-29. http://www.cs.utexas.edu/ftp/pub/techreports/tr00-29.dir.

Kaufmann, M. and Moore J. S. (2001) Structured theory development for a mechanized logic. *Journal of Automated Reasoning*, 26, 161–203.

Kaufmann, M., Manolios, P. and Moore, J. S. (eds.) (2000a) *Computer-Aided Reasoning: ACL2 Case Studies*. Dordrecht: Kluwer Academic Press.

Kaufmann, M., Manolios, P. and Moore, J. S. (2000b) *Computer-Aided Reasoning: An Approach*. Dordrecht: Kluwer Academic Press.

Manolios, P. (2000) Mu-calculus model-checking. In Kaufmann et al. (2000) pp. 93–112.

Manolios, P. and Moore, J. Partial functions in ACL2. In Kaufmann and Moore (2000). http://www.cs.utexas.edu/ftp/pub/techreports/tr00-29.dir.

Manolios, P., Namjoshi, K. and Sumners, R. (1999) Linking theorem proving and model-checking with well-founded bisimulation. In *Computed Aided Verification, CAV '99* (pp. 369–79). New York: Springer-Verlag, LNCS 1633.

McCarthy, J. (1960) Recursive functions of symbolic expressions and their computation by machine (part I). *CACM*, 3(4), 184–95.

McCarthy, J. (1962) Towards a mathematical science of computation. In *Proceedings of IFIP Congress* (pp. 21–8). Amsterdam: North-Holland.

McCarthy, J. (1963) A basis for a mathematical theory of computation. In *Computer Programming and Formal Systems*. Amsterdam: North-Holland.

McCarthy, John and Painter, James (1967) Correctness of a compiler for arithmetic expressions. In *Proceedings of Symposia in Applied Mathematics*, vol. 19. American Mathematical Society.

McCune, W. (1994) *Otter 3.0 Reference Manual and Guide*, Technical Report ANL-94/6, Argonne National Laboratory, Argonne, IL. See also URL http://www.mcs.-anl.gov/AR/otter/.

McCune, W. and Shumsky, O. (2000) Ivy: A preprocessor and proof checker for first-order logic. In Kaufmann et al. (2000), pp. 265–82.

Moore, J. S. (1999) Proving theorems about Java-like byte code. In E.-R. Olderog and B. Steffen (eds.), *Correct System Design – Recent Insights and Advances* (pp. 139–62). LNCS 1710.

Moore, J. S. and Porter, G. (2000a) An executable formal JVM thread model. Submitted for publication.

Moore, J. S. and Porter, G. (2000b) Mechanized reasoning about Java threads via a JVM thread model. Submitted for publication.

Moore, J. S., Lynch, T. and Kaufmann, M. (1998) A mechanically checked proof of the correctness of the kernel of the AMD5K86 floating point division algorithm. *IEEE Transactions on Computers*, 47(9), 913–26.

Owre, S., Rushby, J. and Shankar, N. (1992) PVS: A prototype verification system. In D. Kapur (ed.), *11th International Conference on Automated Deduction (CADE)*, (pp. 748–52). Lecture Notes in Artificial Intelligence, vol. 607. New York: Springer-Verlag.

Russinoff, D. (1998) A mechanically checked proof of IEEE compliance of a register-transfer-level specification of the AMD-K7 floating-point multiplication, division, and square root instructions. *London Mathematical Society Journal of Computation and Mathematics*, 1, 148–200.

Russinoff, D. M. and Flatau, A. (2000) RTL verification: A floating-point multiplier. In Kaufmann et al. (2000), pp. 201–32.

Sawada, J. and Hunt, W. (1998) Processor verification with precise exceptions and speculative execution. In *Computed Aided Verification, CAV '98* (pp. 135–46). New York: Springer-Verlag. LNCS 1427.

Smith, S. W. and Austel, V. (1998) Trusting trusted hardware: Towards a formal model for programmable secure coprocessors. In *The Third USENIX Workshop on Electronic Commerce*, September.

Steele, G. L. Jr. (1990) *Common Lisp: The Language*, 2nd edn. Burlington, MA: Digital Press.

Sumners, R. (2000) Correctness proof of a BDD manager in the context of satisfiability checking. In Kaufmann and Moore (2000). http://www.cs.utexas.edu/ftp/pub/techreports/tr00–29.dir.

Wilding, Matthew, Greve, David and Hardin, David (to appear) Efficient simulation of formal processor models. *Formal Methods in System Design*. Draft TR available as http://pobox.com/users/hokie/docs/efm.ps.

Further Reading

If you are interested in reading more about ACL2, the definitive book is Kaufmann et al. (2000b), which explains the programming language, the logic, the theorem prover, and how to use them. The book contains exercises, and the solutions to the exercises are available on the Web through the ACL2 home page http:/www.cs.utexas.edu/users/moore/acl2. A wealth of additional reading material is available from the home page.

In addition, ACL2 is available at no fee under the GNU General Public License. You may install it and then define functions, execute them, and learn to prove theorems with the ACL2 theorem prover. Installation instructions and several megabytes of hypertext documentation are available on the ACL2 home page. Also of value are the *two short tours* link on the home page and the previously mentioned flying demo, http://www.cs.utexas.edu/users/moore/publications/flying-demo/script.html.

46

Sampling Labeled Deductive Systems

D. M. GABBAY

1 Labeled Deductive Systems in Context

In the past 30 years logic has undergone a serious evolutionary development. The meteoric rise of the applied areas of computer science and artificial intelligence put pressure on traditional logic to evolve. There was the urgent need to develop new logics in order to provide better models of human behavior and actions. Such models are used to help design products which aid/replace the human in his daily activity. As a result, a rich variety of new logics have been developed and there was the need for a new unifying methodology for the chaotic landscape of the new logics.

Such a methodology is *Labeled Deductive Systems* (LDS).

The purpose of this chapter is to introduce Labeled Deductive Systems and show that many logical systems, new and old, monotonic and non-monotonic all fall within this new framework. This chapter is based on Gabbay (1996).

We begin with the traditional view of what is a logical system.

Traditionally, to present a logic **L**, we need to first present the set of well-formed formulas of that logic. This is the *language* of the logic. We specify the sets of atomic formulas, connectives, quantifiers, and the set of well-formed formulas. Secondly, we mathematically define the notion of consequence, that is, for sets of formulas Δ and formulas Q, we define the consequence relation $\Delta \vdash_{\mathbf{L}} Q$, which is read '$Q$ follows from Δ in the logic **L**.'

The consequence relation is required to satisfy the following intuitive properties: (Δ, Δ' abbreviates $\Delta \cup \Delta'$).

Reflexivity

$\Delta \vdash Q$ if $Q \in \Delta$

Monotonicity

$$\frac{\Delta \vdash Q}{\Delta, \Delta' \vdash Q}$$

Transitivity

$$\frac{\Delta \vdash A; \Delta, A \vdash Q}{\Delta \vdash Q}$$

If you think of Δ as a database and Q as a query, then reflexivity means that the answer 'yes' is given for any Q which is already listed in the database Δ. Monotonicity reflects the accumulation of data, and transitivity is nothing but lemma generation, namely, if $\Delta \vdash A$, then A can be used as a lemma to derive B from Δ.

These three properties have appeared to constitute the minimal and most natural for a logical system, given that the main applications of logic were in mathematics and philosophy.

The above notions were essentially put forward by Tarski (1956) in 1936 and is referenced as Tarski consequence. Scott (1974), inspired by constructions in Gabbay (1991), generalized the notion to allow Q to be a set of formulas Γ. The basic relation is then of the form $\Delta \vdash \Gamma$, satisfying:[1]

Reflexivity

$\Delta \vdash \Gamma$ if $\Delta \cap \Gamma \neq \varnothing$

Monotonicity

$$\frac{\Delta \vdash \Gamma}{\Delta, \Delta' \vdash \Gamma}$$

Cut

$$\frac{\Delta, A \vdash \Gamma; \Delta' \vdash A, \Gamma'}{\Delta, \Delta' \vdash \Gamma, \Gamma'}$$

Scott further showed that for any Tarski consequence relation $\vdash$ there exist two Scott consequence relations (a maximal one and a minimal one) that agree with it, namely, that $\Delta \vdash A$ (Tarski) iff $\Delta \vdash \{A\}$ (Scott) (see Gabbay 1981).

The above notions are monotonic. However, the increasing use of logic in computer science and artificial intelligence has given rise to logical systems which are not monotonic, that is to systems in which the axiom of monotonicity is not satisfied. There are many such systems, satisfying a variety of conditions and presented in a variety of ways. Furthermore, some are characterized in a proof theoretical and some in a model theoretical manner. All these different presentations give rise to some notion of consequence $\Delta \vdash Q$, but they only seem to all agree on reflexivity.[2] The essential difference between these logics (commonly called *non-monotonic logics*) and the more traditional logics (now referred to as *monotonic logics*) is the fact that $\Delta \vdash A$ holds in the monotonic case because of some $\Delta_A \subseteq \Delta$, while in the non-monotonic case the entire set Δ is

somehow used to derive A. Thus if Δ is increased to Δ', there is no change in the monotonic case, while there may be a change in the non-monotonic case.

The above describes the situation current in the early 1980s. We have had a multitude of systems generally accepted as 'logics' without a unifying underlying theory and many had semantics without proof theory or vice versa, though almost all of them were based on some sound intuitions of one form or another. Clearly there was the need for a general unifying framework. An early attempt at classifying non-monotonic systems was Gabbay (1985). It was put forward that basic axioms for a Tarski type consequence relation should be *reflexivity, transitivity*, and *restricted monotonicity*, namely:

Restricted monotonicity (cumulativity)

$$\frac{\Delta \vdash A; \Delta \vdash B}{\Delta, A \vdash B}$$

A variety of systems seem to satisfy this axiom. See a survey in Makinson (1994) and Gabbay (1996).

Although some sort of classification was obtained and semantical results were proved, the approach does not seem to be strong enough. Many systems do not satisfy restricted monotonicity. Other systems such as relevance logic, do not even satisfy reflexivity. Others have a richness of their own which is lost in a simple presentation as an axiomatic consequence relation. Obviously a different approach is needed, one which would be more sensitive to the variety of features of the systems in the field. Fortunately, developments in a neighboring area, that of automated deduction, seem to be of help. New automated deduction methods were developed for nonclassical logics, and resolution was generalized and modified to be applicable to these logics. In general, because of the value of these logics in theoretical computer science and artificial intelligence, a greater awareness of the computational aspects of logical systems was developing and more attention was being devoted to proof-theoretical presentations. It became apparent to us that a key feature in the proof-theoretic study of these logics is that a slight natural variation in an automated or proof-theoretic system of one logic (say $\mathbf{L}_1$), can yield another logic (say $\mathbf{L}_2$).

Although $\mathbf{L}_1$ and $\mathbf{L}_2$ may be conceptually far apart (in their philosophical motivation, and mathematical definitions) when it comes to automated techniques and proof theoretical presentation, they turn out to be brother and sister. This kind of relationship is not isolated and seems to be widespread. Furthermore, non-monotonic systems seem to be obtainable from monotonic ones through variations on some of their monotonic proof-theoretical formulation, thus giving us a handle on classifying non-monotonic systems.

This phenomena has prompted Gabbay (1992) to put forward the view that a logical system $\mathbf{L}$ is not just the traditional consequence relation $\vdash$ (monotonic or non-monotonic) but a pair $(\vdash, \mathbf{S}_\vdash)$ where $\vdash$ is a mathematically defined consequence relation (i.e. the set of pairs (Δ, Q) such that $\Delta \vdash Q$) satisfying whatever minimal conditions on a consequence relation one happens to agree on, and $\mathbf{S}_\vdash$ is an algorithmic system for

generating all those pairs. Thus according to this definition classical logic $\vdash$ perceived as a set of tautologies together with a Gentzen system $\mathbf{S}_{\vdash}$ is not the same as classical logic together with the two-valued truth table decision procedure $\mathbf{T}_{\vdash}$ for it. In our conceptual framework, $(\vdash, \mathbf{S}_{\vdash})$ is *not the same logic* as $(\vdash, \mathbf{T}_{\vdash})$.

To illustrate and motivate our way of thinking, observe that it is very easy to move from $\mathbf{T}_{\vdash}$ for classical logic to a truth table system $\mathbf{T}^n_{\vdash}$ for Łukasiewicz n-valued logic. It is not so easy to move to an algorithmic system for intuitionistic logic. In comparison, for a Gentzen system presentation, exactly the opposite is true. Intuitionistic and classical logics are neighbors, while Łukasiewicz logics seem completely different. In fact, some of the examples of this chapter show proof theoretic similarities between Łukasiewicz's infinite valued logic and Girard's Linear Logic, which in turn is proof theoretically similar to intuitionistic logic.

There are many more such examples among temporal logics, modal logics, defeasible logics and others. Obviously, there is a need for a more unifying framework. The question is then whether we can adopt a concept of a logic where the passage from one system to another is natural, and along predefined accceptable modes of variation? Can we put forward a framework where the computational aspects of a logic also play a role? Is it possible to find a common home for a variety of seemingly different techniques introduced for different purposes in seemingly different intellectual logical traditions?

To find an answer, let us ask ourselves what makes one logic different from another? How is a new logic presented and described and compared to another? The answer is obvious. These considerations are usually dealt with on the meta-level. Most logics are based on *modus ponens* and the quantifier rules are formally the same anyway and the differences between them are meta-level considerations on the proof theory or semantics. If we can find a mode of presentation of logical systems where meta-level features can reside side by side with object level features then we can hope for a general framework. We must be careful here. In the logical community the notions of object-level vs. meta-level are not so clear. Most people think of *naming* and *proof predicates* in this connection. This is not what we mean by meta-level here. We need a more refined understanding of the concept. There is a similar need in computer science. In Gabbay (1996) we devote a chapter to these considerations. See also Gabbay (1992).

We found that the best framework to put forward is that of a *Labeled Deductive System, LDS*. Our notion of what constitutes a logic will be that of a pair $(\vdash, \mathbf{S}_{\vdash})$ where $\vdash$ is a set-theoretic (possibly non-monotonic) consequence relation on a language $\mathbf{L}$ and $\mathbf{S}_{\vdash}$ is an *LDS*, and where $\vdash$ is essentially required to satisfy no more than *Identity* (i.e. $\{A\} \vdash A$) and *Surgical Cut* (see below and Gabbay (1991; forthcoming)). This is a refinement of our concept of a logical system mentioned above and first presented in Gabbay (1992). We now not only say that a logical system is a pair $(\vdash, \mathbf{S}_{\vdash})$, but we are adding that $\mathbf{S}_{\vdash}$ itself has a special presentation, that of an *LDS*.

An *LDS* system is a triple $(\mathbf{L}, \Gamma, \mathbf{M})$, where $\mathbf{L}$ is a logical language (connectives and wffs) and Γ is an algebra (with some operations) of labels and $\mathbf{M}$ is a discipline of labeling formulas of the logic (from the algebra of labels Γ), together with deduction rules and with agreed ways of propagating the labels via the application of the deduction rules. The way the rules are used is more or less uniform to all systems. In the general

case we allow Γ, the algebra of labels, to be an *LDS* system itself! Furthermore, if our view of a logical system is that the declarative unit is a pair, a formula and a label, then we can also label the pair itself and get multiple labeling.

The perceptive reader may feel resistence to this idea at this stage. First be assured that you are not asked to give up your favourite logic or proof theory nor is there any hint of a claim that your activity is now obsolete. In mathematics a good concept can rarely be seen or studied from one point of view only and it is a sign of strength to have several views connecting different concepts. So the traditional logical views are as valid as ever and add strength to the new point of view. In fact, a closer examination of the material in my book would reveal that manifestations of our *LDS* approach already exist in the literature in various forms (see Anderson and Belnap (1975), Fitting (1983) and Gabbay (1996) and the references there), however, they were locally regarded as convenient tools and there was not the realization that there is a general framework to be studied and developed. None of us is working in a vacuum and we build on each others' work. Further, the existence of a general framework in which any particular case can be represented does not necessarily mean that the best way to treat that particular case is within the general framework. Thus if some modal logics can be formulated in *LDS*, this does not mean that in practice we should replace existing ways of treating the logics by their *LDS* formulation. The latter may not be the most efficient for those particular logics. It is sufficient to show how the *LDS* principles specialize and manifest themselves in the given known practical formulation of the logic.

The reader may further have doubts about the use of labels from the computational point of view. What do we mean by a unifying framework? Surely a Turing machine can simulate any logic, is that a unifying framework? The use of labels is powerful, as we know from computer science, are we using labels to play the role of a Turing machine? The answer to the question is twofold. First that we are not operating at the meta-level, but at the object level. Second, there are severe restrictions on the way we use *LDS*. Here is a preview:

1. The only rules of inference allowed are the traditional ones, modus ponens, and some form of deduction theorem for implication, for example.
2. Allowable modes of label propagation are fixed for all logics. They can be adjusted in agreed ways to obtain variations but in general the format is the same. For example, it has the following form for implications: $(A \to B)$ gets label t iff $\forall x \in \Gamma_1$ [If A is labeled x then B can be proved with labels $t + x$], where Γ_1 is a set of labels characterizing the implication in that particular logic. For example Γ_1 may be all atomic labels or related labels to t, or variations. The freedom that different logics have is in the choice of Γ_1 and the properties of '+'. For example we can restrict the use of modus ponens by a wise propagation of labels.
3. The quantifier rules are the same for all logics.
4. Meta-level features are implemented via the labeling mechanism, which is object language.

The reader who prefers to remain within the traditional point of view of: *assumptions* (*data*) *proving a conclusion* can view the labeled formulas as another form of data.

There are many occasions when it is most intuitive to present an item of data in the form $t : A$, where t is a label and A is a formula. The common underlying reason for the use of the label t is that t represents information which is needed to modify A or to supplement (the information in) A which is not of the same type or nature as (the information represented by) A itself. A is a logical formula representing information declaratively, and the additional information of t can certainly be added declaratively to A to form A', however, we may find it convenient to put forward the additional information through the label t as part of a pair $t : A$.

Take for example a source of information which is not reliable. A natural way of representing an item of information from that source is $t : A$, where A is a declarative presentation of the information itself and t is a number representing its reliability. Such expert systems exist (e.g. Mycin) with rules which manipulate both t and A as one unit, propagating the reliability values t_i through applications of *modus ponens*. We may also use a label naming the source of information and this would give us a qualitative idea of its reliability.

Another area where it is natural to use labels is in reasoning from data and rules. If we want to keep track, for reasons of maintaining consistency and/or integrity constraints, where and how a formula was deduced, we use a label t. In this case, the label t in $t : A$ can be the part of the data which was used to get A. Formally in this case t is a formula, the conjunction of the data used. We thus get pairs of the form $\Delta_i : A_i$, where A_i are formulas and Δ_i are the parts of the database from which A_i was derived.

A third example where it is natural to use labels is time stamping of data. Where data are constantly revised and updated, it is important to time stamp the data items. Thus the data items would look like $t_i : A_i$, where t_i are time stamps. A_i itself may be a temporal formula. Thus there are two times involved, the logical time s_i in $A_i(s_i)$ and the time stamping t_i of A_i. For reasons of clarity, we may wish to regard t_i as a label rather than incorporate it into the logic (by writing for example $A^*(t_i, s_i)$).

To summarize then, we replace the traditional notion of consequence between formulas of the form $A_1, \ldots, A_n \vdash B$ by the notion of consequence between labeled formulas

$$t_1 : A_1, t_2 : A_2, \ldots t_n : A_n \vdash s : B$$

Depending on the logical system involved, the intuitive meaning of the labels varies. In querying databases, we may be interested in labeling the assumptions so that when we get an answer to a query, we can record, via the label of the answer, from which part of the database the answer was obtained. Another area where labeling is used is temporal logic. We can time stamp assumptions as to when they are true and query, given those assumptions, whether a certain conclusion will be true at a certain time. Thus the consequence notion for labeled deduction is essentially the same as that of any logic: given assumptions does a conclusion follow.

Whereas in the traditional logical system the consequence is defined using proof rules on the formulas, in the *LDS* methodology the consequence is defined by using rules on both formulas and their labels. Formally we have formal rules for manipulating labels and this allows for more scope in decomposing the various features of the consequence relation. The meta features can be reflected in the algebra or logic of the labels and the object features can be reflected in the rules of the formulas.

The notion of a database or of a 'set of assumptions' also has to be changed. A database is a configuration of labeled formulas. The configuration depends on the labeling discipline. For example, it can be a linearly ordered set $\{a_1 : A_1, \ldots, a_n : A_n\}$, $a_1 < a_2 < \ldots < a_n$. The proof discipline for the logic will specify how the assumptions are to be used. We need to develop the notions of the Cut Rule and the Deduction Theorem in such an environment. This we do in a later section.

The next two sections will give many examples of *LDS* disciplines featuring many known monotonic and non-monotonic logics. It is of value to summarize our view listing the key points involved:

- The unit of declarative data is a labeled formula of the form $t : A$, where A is a wff of a language $\mathbf{L}$ and t is a label. The labels come from an algebra (set) of labels.
- A database is a set of labeled formulas.
- An *LDS* discipline is a system (algorithmic) for manipulating both formulas and their labels. Using this discipline the statement $\Delta \vdash \Gamma$ is well defined for the two databases Δ and Γ. Especially $\Delta \vdash t : A$ is well defined.
- $\vdash$ must satisfy the minimal conditions, namely

Identity

$$\{t : A\} \vdash t : A$$

Surgical cut

$$\frac{\Delta \vdash t : A, \Gamma[t : A] \vdash s : B}{\Gamma[\Delta] \vdash s : B}$$

where $\Gamma[t : A]$ means that $t : A$ is contained/occurs somewhere in the structure Γ and $\Gamma[\Delta]$ means that Δ replaces A in the structure.

- A logical system is a pair $(\vdash, \mathbf{S}_\vdash)$, where $\vdash$ is a consequence relation and $\mathbf{S}_\vdash$ is an *LDS* for it.

2 Examples from Monotonic Logics

To motivate our approach we study several known examples in this section.

Example 2.1 below shows a standard deduction from Relevance Logic. The purpose of the example is to illustrate our point of view. There are many such examples in Anderson and Belnap (1975). Example 2.3 below considers a derivation in modal logic. There we use labels to denote essentially possible worlds. The objective of the example is to show the formal similarities to the relevance logic case in Example 2.1. Example 2.4 can reap the benefits of the formal similarities of the first two examples and introduce, in the most natural way, a system of relevant modal logic. The objective of

Example 2.4 is to show that the labels in Example 2.1 and Example 2.3 can be read as determining the metalanguage features of the logic and can therefore be combined 'declaratively' to form the new system of 2.4. Example 2.5 considers strict implication. This example shows that for strict S4 implication one can read the labels either as relevance labels or as possible world labels. Example 2.6 shows how labels can interact with quantifiers in modal logic. We continue with examples of relevance reasoning, many-valued logics, formulas as types, realizability and conclude with a formal definition of an algebraic LDS for $\rightarrow$ and $\neg$.

EXAMPLE 2.1 (RELEVANCE AND LINEAR LOGIC) Consider a propositional language with implication '$\rightarrow$' only. The forward elimination rule is *modus ponens*. From the theorem proving view, *modus ponens* is an object language consideration. Thus a proof of $\vdash (B \rightarrow A) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow B))$ can proceed as follows:

Assume $a_1 : B \rightarrow A$ and show $(A \rightarrow B) \rightarrow (A \rightarrow B)$. Further assume $a_2 : A \rightarrow B$ and show $A \rightarrow B$. Further assume $a_3 : A$ and show B. We thus end up with the following problem:

Assumptions

1. $a_1 : B \rightarrow A$
2. $a_2 : A \rightarrow B$
3. $a_3 : A$

Derivation

4.	$a_2a_3 : B$	by *modus ponens* from lines (2) and (3).
5.	$a_1a_2a_3 : A$	from (4) and (1).
6.	$a_2a_1a_2a_3 : B$	from (5) and (2).
7.	$a_2a_1a_2 : A \rightarrow B$	from (3) and (6).
8.	$a_2a_1 : (A \rightarrow B) \rightarrow (A \rightarrow B)$	from (2) and (7).
9.	$a_2 : (B \rightarrow A) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow B))$	from (1) and (8).

The meta aspect of this proof is the annotation of the assumptions and the keeping track of what was used in the deduction. A meta-leval condition would determine the logic involved.

A formal definition of the labeling discipline for this class of logics is given in Gabbay (1996). For this example it is sufficient to note the following three conventions:

1. Each assumption is labeled by a new atomic label.
   An ordering on the labels can be imposed, namely $a_1 < a_2 < a_3$. This is to reflect the fact that the assumptions arose from our attempt to prove $(B \rightarrow A) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow B))$ and not for example from $(A \rightarrow B) \rightarrow ((B \rightarrow A) \rightarrow (A \rightarrow B))$ in which case the ordering would be $a_2 < a_1 < a_3$. The ordering can affect the proofs in certain logics.

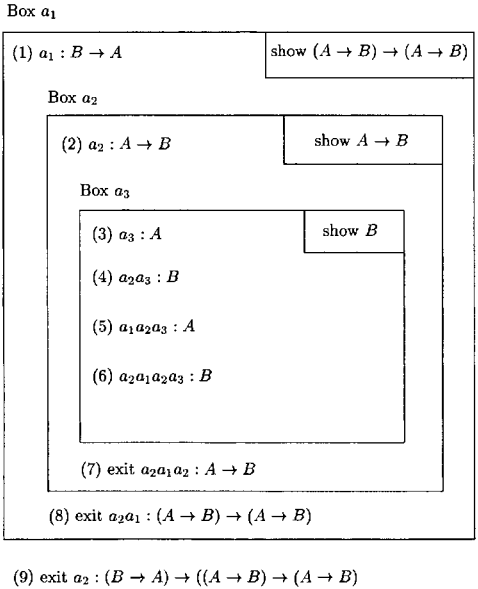


Figure 46.1

2. If in the proof, A is labeled by the multiset α and $A \to B$ is labeled by β then B can be derived with a label $\alpha \cup \beta$ where '$\cup$' denotes multiset union.
3. If B was derived using A as evidenced by the fact that the label α of A is a submultiset of the label β of B ($\alpha \subseteq \beta$) then we can derive $A \to B$ with the label $\beta - \alpha$ ('$-$' is multiset subtraction).

The derivation can be represented in a more graphical way.

To show $(B \to A) \to ((A \to B) \to (A \to B))$. See figure 46.1.

The above is the *metabox* way of representing the deduction. Note that in line 8, multiset subtraction was used and only one copy of the label a_2 was taken out. The other copy of a_2 remains and cannot be cancelled. Thus this formula is not a theorem of linear logic, because the outer box does not exit with label $\varnothing$. In relevance logic, the discipline uses sets and not multisets. Thus the label of line 8 in this case would be a_1 and that of line 9 would be $\varnothing$. The above deduction can be made even more explicit as follows:

$(B \to A) \to ((A \to B) \to (A \to B))$ follows with a label from Box a_1.

Box a_1

a_1 :	$B \to A$ assumption
a_2a_1 :	$(A \to B) \to (A \to B)$ from Box a_2

Box a_2

a_2 :	$A \to B$ assumption
$a_2a_1a_2$:	$A \to B$ from Box a_3

Box a_3

$a_3 : A$	assumption
$a_2 : A \to B$	reiteration from box a_2
$a_2a_3 : B$	by *modus ponens*
$a_1 : B \to A$	reiteration from box a_1
$a_1a_2a_3 : A$	*modus ponens* from the two preceding lines
$a_2 : A \to B$	repetition of an earlier line
$a_2a_1a_2a_3 : B$	*modus ponens* from the two preceding lines

The following meta-rule was used:

We have a system of partially ordered metaboxes $a_1 < a_2 < a_3$. Any assumption in a box a can be reiterated in any box b provided $a < b$.

REMARK 2.2 a. The above presentation of the boxes makes them look more like possible worlds. The labels are the worlds and formulas can be exported from one world to another according to some rules. The next example 2.3 describes modal logic in just this way.

b. Note that different meta-conditions on labels and metaboxes correspond to different logics.

The following table gives intuitively some correspondence between meta-conditions and logics.

Meta-condition	**Logic**
ignore the labels	intuitionistic logic
accept only the derivations which use all the assumptions	relevance logic
accept derivations which use all assumptions exactly once	linear logic

The meta-conditions can be translated into object conditions in terms of axioms and rules. If we consider a Hilbert system with modus ponens and substitution then the additional axioms involved are given below:

Linear Logic

$A \to A$
$(A \to (B \to C)) \to (B \to (A \to C))$
$(C \to A) \to ((B \to C) \to (B \to A))$
$(C \to A) \to ((A \to B) \to (C \to B))$

Relevance Logic
Add the schema below to linear logic

$(A \to (B \to C)) \to ((A \to B) \to (A \to C))$

Intuitionistic Logic
Add the schema below to relevance logic:

$A \to (B \to A)$

The reader can note that the following axiom (Peirce Rule) yields classical logic. Further note that for example, we can define 'Linear Classical Logic' by adding Peirce Rule to linear logic. A new logic is obtained.

Classical Logic
Add the schema below to intuitionistic logic:

$((A \to B) \to A) \to A.$

EXAMPLE 2.3 This example shows the meta-level–object level division in the case of modal logic. Modal logic has to do with possible worlds. We thus think of our basic database (or assumptions) as a finite set of information about possible worlds. This consists of two parts. The configuration part, the finite configuration of possible worlds for the database, and the assumptions part which tells us what formulas hold in each world. The following is an example of a database:

	Assumptions	**Configuration**
(1)	$t : \Box\Box B$	$t < s$
(2)	$s : \Diamond(B \to C)$	

The conclusion to show (or query) is:

$t : \Diamond\Diamond C.$

The derivation is as follows:

3. From (2) create a new point r with $s < r$ and get $r : B \rightarrow C$.

We thus have

Assumptions	**Configuration**
(1), (2), (3)	$t < s < r$

4. From (1), since $t < s$ we get $s : \Box B$.
5. From (4), since $s < r$ we get $r : B$.
6. From (5) and (3) we get $r : C$.
7. From (6) since $s < r$ we get $s : \Diamond C$.
8. From (7) using $t < s$ we get $t : \Diamond\Diamond C$.

Discussion:
The object rules involved are:

$\Box E$ **Rule**:

$$\frac{t < s; t : \Box A}{s : A}$$

$\Diamond I$ **Rule**:

$$\frac{t < s, s : B}{t : \Diamond B}$$

$\Diamond E$ **Rule**:

$$\frac{t : \Diamond A}{\text{create a new point } s \text{ with } t < s \text{ and deduce } s : A}$$

Note that the above rules are not complete. We do not have rules for deriving, for example, $\Box A$. Also, the rules are all for intuitionistic modal logic.

The meta level consideration may be properties of $<$,

e.g. transitivity $t < s \wedge s < r \rightarrow t < r$ or
e.g. linearity: $t < s \vee t = s \vee s < t$ etc.

EXAMPLE 2.4 The reader can already see the benefit of separating the meta-level (the handling of possible worlds i.e. labels) and the object-level (i.e. formulas) features. We can combine both the meta-level features of Examples 2.1 and 2.3 to create for example a modal relevance logic in a natural way. Each assumption has a relevance label as well as a world label. Thus the proof of the previous example becomes the following:

	Assumptions	Configuration
(1)	$(a_1, t) : \Box\Box B$	$t < s$
(2)	$(a_2, s) : \Diamond(B \to C)$	

We proceed to create a new label r using $\Diamond E$ rule. The relevance label is carried over. We have $t < s < r$.

3. $(a_2, r) : B \to C$

Using $\Box E$ rule with relevance label carried over, we have:

4. $(a_1, s) : \Box B$
5. $(a_1, r) : B$

Using *modus ponens* with relevance label updated

6. $(a_1, a_2, r) : C$

Using $\Diamond I$ rule:

7. $(a_1, a_2, s) : \Diamond C$
8. $(a_1, a_2, t) : \Diamond\Diamond C$

(8) means that we got t: $\Diamond\Diamond C$ using both assumptions a_1 and a_2.

There are two serious problems in modal and temporal theorem proving. One is that of Skolem functions for $\exists x \Diamond A(x)$ and $\Diamond \exists x A(x)$ are not logically the same. If we skolemize we get $\Diamond A(c)$. Unfortunately it is not clear where c exists, in the current world $((\exists x = c)\Diamond A(x))$ or the possible world $(\Diamond(\exists x = c)A(x))$.

If we use labeled assumptions then, $t : \exists x \Diamond A(x)$ becomes $t : \Diamond A(c)$ and it is clear that c is introduced at t. In fact we shall write it as c^t.

On the other hand, the assumption $t : \Diamond \exists x A(x)$ will be used by the $\Diamond E$ rule to introduce a new point s, $t < s$ and conclude $s : \exists x A(x)$. We can further skolemize at s and get $s : A(c)$, with c introduced at s and write it as c^s. We thus need the mechanism of remembering or labeling constants as well, to indicate where they were first introduced, and we need rules to govern them. This is illustrated in Example 2.6 below.

Labeling systems for modal and temporal logics is studied in Gabbay (1991).

EXAMPLE 2.5 The following example describes the logic of modal S4 strict implication. In this logic the labels can be read either as relevance labels or as possible worlds. S4 strict implication $A \to B$ can be understood as a temporal connective, as follows:

'$A \to B$ is true at world t iff for all future worlds s to t and for t itself we have that if A is true at s then B is true at s.' Thus $A \to B$ reads 'From now on, if A then B.'

Suppose we want to prove that $A \to B$ and $A \to (B \to C)$ imply $A \to C$. To show this we reason semantically and assume that at time t, the two assumptions are true. We

want to show that $A \to C$ is also true at t. To prove that we take any future time s, assume that A is true at s and show that C is also true at s. We thus have the following situation:

1. $t : A \to B$
2. $t : A \to (B \to C)$
3. show $t : A \to C$
   from box

3.1	Assume $s : A$ Show $s : C$ Since s is in the future of t, we get that at s, (1) and (2) are also true.
3.2	$s : A \to B$ from (1)
3.3	$s : A \to (B \to C)$ from (2) We now use *modus ponens*, because $X \to Y$ means 'from now on, if X then Y'
3.4	$s : B$ from (3.1) and (3.2)
3.5	$s : B \to C$ from (3.2) and (3.3)
3.6	$s : C$ *modus ponens* from (3.4) and (3.5)

exit $t : A \to C$

Notice that any $t : D$ can be brought into (reiterated) the box as s: D, provided it has an implicational form, $D = D_1 \to D_2$. We can thus regard the labels above as simply naming assumptions (not as possible worlds) and the logic has the reiteration rule which says that only implications can be reiterated.

Let us add a further note to sharpen our understanding. Suppose $\to$ is read as a **K4** implication (i.e. transitivity without reflexivity). Then the above proof should fail. Indeed the corresponding restriction on modus ponens is that we do perform X, $X \to Y \vdash Y$ in a box, provided $X \to Y$ is a reiteration into the box and was not itself derived in that same box. This will block line (3.6).

EXAMPLE 2.6 Another example has to do with the Barcan formula.

This is a case of quantified modal logic. We need to organize how to deal with quantifiers in LDS. The idea is that whenever we introduce a variable or a constant under a label we must label the variable/constant as well. Thus we have the rule:

$$\frac{t : \exists x A(x)}{t : A(c^t)} \qquad \frac{t : \forall x A(x)}{t : A(x^t)}$$

we also have $t : x^t$ and $t : c^t$ holding, where $t : y$ means that y *resides* at t. A rule of the form

$$\frac{t : y}{s : y}$$

is called a visa rule, allowing for a term y residing at t also to reside at s. Thus we have the $\exists$ introduction rule as

$$\frac{t:A(y);t:y}{t:\exists yA(y)}.$$

and the universal generalization rule:

$$\frac{t:A(x);t:x,x \text{ universal variable}}{t:\forall xA(x)}$$

To get the Barcan formula we need a visa rule

$$\frac{t:y;t<s}{s:y}$$

We can now prove this formula.

	Assumption	**Configuration**
(1)	$t:\forall x\Box A(x)$	$t<s$

We show

s: $\forall xA(x)$

We proceed intuitively

1. $t:\Box A(x)$ (stripping $\forall x$, remembering x is arbitrary), and $t:x$.
2. Since the configuration contains s, $t<s$ we get

   $s:A(x)$

3. Since x is arbitrary we get by visa rule and $\Box$ rule:

   $s:\forall xA(x)$; $s:x$

The rule

$$\frac{t:\Box A(x),t<s}{s:A(x)}$$

is allowed because of the visa rule.

To have the above rule for arbitrary x is equivalent to adopting the Barcan formula axiom:

$\forall x \Box A(x) \rightarrow \Box \forall x A(x)$

To show $\Box \forall x A(x) \rightarrow \forall x \Box A(x)$, we need the visa rule:

$$\frac{t:y; s<t}{s:y}$$

The above are just a few examples for the scope we get using labels. The exact details and correspondences are worked out in our monograph Gabbay (1996).

EXAMPLE 2.7 (RELEVANCE REASONING) The indices are α, β, and $\gamma = (\beta - \alpha)$. The reasoning structure is:

Assume $\alpha : A$
Show $\beta : B$
If $\beta \supseteq \alpha$ then exit with $(\beta - \alpha) : A \rightarrow B$.

To show $A \rightarrow (B \rightarrow C) \vdash B \rightarrow (A \rightarrow C)$
Assume

$a_1 : A \rightarrow (B \rightarrow C)$

we use the metabox to show $B \rightarrow (A \rightarrow C)$. See figure 46.2.

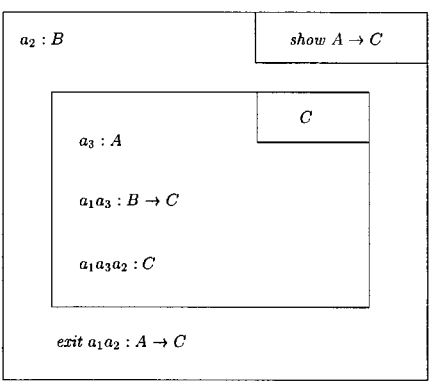


Figure 46.2

EXAMPLE 2.8 (ŁUKASIEWICZ MANY-VALUED LOGICS) Consider Łukasiewicz infinite-valued logic, where the values are all real numbers or rationals in $[0,1]$. We designate 0 as **truth** and the truth table for implication is

$$x \to y = \max(0, y - x)$$

Here the language contains atoms and implication only, assignments h give values to atoms in $[0,1]$, $h(q) \in [0,1]$ and h is extended to arbitrary formulas via the table for $\to$ above. Define the relation

$$A_1, \ldots, A_n \vdash B$$

to mean that for all h, $h(A_1) + \ldots + h(A_n) \geqq h(B)$, where $+$ is numerical addition.

This logic can be regarded as a labeled deductive system, where the labels are values $t \in [0,1]$. $t : A$ means that $h(A) = t$, for a given background assignment h. The interesting part is that to show $t : A \to B$ (i.e. that $A \to B$ has value t) we assume $x : A$ (i.e. that A has value x) and then have to show that B has value $t + x$, i.e. show $t + x : B$.

This is according to the table of $\to$.

Thus figure 46.3 shows the deduction in box form:

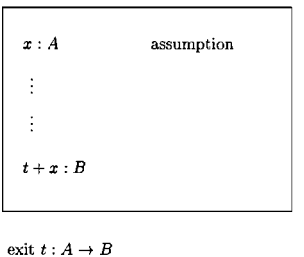


Figure 46.3

This has the *same structure* as the case of relevance logic, where $+$ was understood as concatenation.

A full study of many valued logics from the LDS point of view is given in Gabbay (1996).

EXAMPLE 2.9 (FORMULAS AS TYPES) Another instance of the natural use of labels is the Curry–Howard interpretation of formulas as types. This interpretation conforms exactly to our framework. In fact, our framework gives the incentive to extend the formulas as types interpretation in a natural way to other logics, such as linear and relevance logics and surprisingly, also many valued logics, modal logics, and intermediate logics. A formula is considered as a type and its label is a *definable* λ-term of the same

type. Given a system for defining λ-terms, the theorems of the logic are all those types which can be shown to be nonempty.

The basic propagation mechanism corresponding to *modus ponens* is:

$$\frac{t^{A} : A \qquad t^{A \to B} : A \to B}{t^{A \to B}(t^{A}) : B}$$

It is satisfied by *application*.

Thus if we read the + in $t^{A \to B} + t^{A}$ as application, we get the exact parallel to the general schema of propagation. Compare with relevance logic where + was concatenation, and with many valued logics where + was numerical addition!

To show $t : A \to B$ we assume $x : A$, with x arbitrary, that is start with a term x of type A, use the proof rules to get B. As we saw, applications of *modus ponens* generate more terms which contain x in them via application. If we accept that proofs generate functionals, then we get B with a label $y = t(x)$. Thus $t = \lambda x t(x)$. This again conforms with our general schema for $\to$.

In Gabbay and Queiroz (1992) on the Curry–Howard interpretation we exploit this idea systematically. There are two mechanisms which allow us to restrict or expand our ability to define terms of any type. We can restrict λ-abstraction (e.g. allow $\lambda x t(x)$ only if x actually occurs in t), this will give us logics weaker than intuitionistic logic, or we can increase our world of terms by requiring diagrams to be closed, for example, for any φ of classical logic such that

$$\vdash (A \to B) \to [\varphi(A) \to \varphi(B)]$$

in classical logic, we want figure 46.4 to be complete, that is for any term t there must exist a term t' (see figure 46.4).

Take for example the formula $A \to (B \to A)$ as type. We want to show a definable term of this type, we can try and use the standard proof (see figure 46.5), however, with the restriction on λ-abstraction which requires the abstracted variable to actually occur in the formula, we cannot exit the inner box. For details see Gabbay and Queiroz (1992).

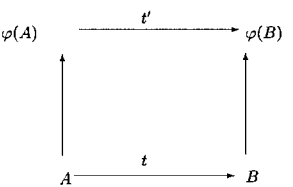


Figure 46.4

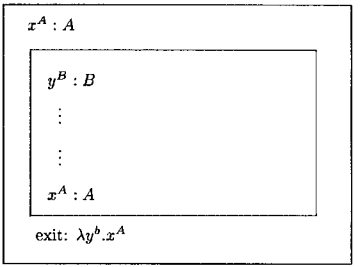


Figure 46.5

EXAMPLE 2.10 (REALIZABILITY INTERPRETATION) The well-known realizability interpretation for intuitionistic implication is another example of a functional interpretation for $\rightarrow$ which has the same universal *LDS* form. A notation for a recursive function $\{e\}$ realises an implication $A \rightarrow B$ iff for any n which realizes A, $\{e\}(n)$ realizes B. Thus

$$e : A \rightarrow B \text{ iff } \forall n[n : A \Rightarrow \{e\}(n) : B]$$

It is an open problem to find an axiomatic description of the set of all wffs which are realisable.

DEFINITION 2.11 (AN ALGEBRAIC LDS FOR IMPLICATION AND NEGATION) Let **L** be a propositional language with $\rightarrow$, $\neg$ and atoms. Let A be an algebra of labels with relations $x < y$ for priority among labels, $F(x, y)$ of compatibility among labels and functions, $f(x, y)$ for propagating labels and $\cup+$ for aggregating labels.

Given two labeled formulas $t : A$ and $s : A \rightarrow B$, $F(s, t)$ must hold in order to licence the *modus ponens*. If it does not hold, we cannot get B. If it does hold, we can get B but we must know what is the label of B. This is the job of the function $f(s, t)$. The aggregation function tells us how different proofs of the same B with different labels can reinforce one another. Thus if we have $t : B$ and $s : B$ we can aggregate and get $t \uplus s : B$. See Example 3.4 below for a very famous aggregation rule.

1. A *declarative unit* is a pair $t : A$, where A is a formula and t a term on the algebra of labels (built up from atomic labels and the functions f and $\cup+$).
2. A database is a set containing declarative units and formulae of the form $t_i < s_i$ and $F(t_i, s_i)$ for some labels $t_1, \ldots, s_i, \ldots$

3. The $\rightarrow$ elimination rule, *modus ponens*, has the form

$$\frac{t:A; s:A\rightarrow B; \mathcal{F}(s,t)}{f(s,t):B}$$

4. The $\Rightarrow$ introduction rule has the form
   - To introduce $t : A \rightarrow B$
     Assume $x : A$, for x arbitrary in the set $\{y \mid F(t, y)\}$, and show $f(t, x) : B$.
5. Negation rules have the form

$$\frac{t:B; s:\neg B}{r:C}$$

   We are not writing any specific rules because there are so many options for negation.
6. A family of *flattening rules* **Flat** of the form

$$\frac{t_1:A,\ldots,t_k:A; s_1:\neg A,\ldots,s_m:\neg A; y_i<y_j, i=1,2,\ldots, j=1,2,\ldots}{\gamma=\mathsf{Flat}(\{t_1,\ldots,t_k,s_1,\ldots,s_m\})}$$

   where γ is either 0 or 1 and is the result of applying the function **Flat** on the set containing t_i, s_j and where y_i, y_l range over $\{t_1, \ldots, t_k, s_1, \ldots, s_m\}$.[3] The meaning of γ is as follows. Since obviously we can prove both A and $\neg A$ with different labels, we need a flat decision on whether we take A, $(\gamma = 1)$ or $\neg A$, $(\gamma = 0)$.
7. Aggregation rule

$$\frac{t:A; s:A}{t \uplus s:A}$$

8. $\uplus$ is associative, commutative and f is distributive over $\cup +$.
9. A proof is a sequence of expressions which are of the form $t < s$, $F(t, s)$ or $t : A$ such that each element of the sequence is either an assumption or is obtained from previous elements in the sequence by an elimination rule or is introduced by a subcomputation via the $\rightarrow$ introduction rule. Flattening rules are to be used last.

3 Examples from Non-monotonic Logics

The examples in the previous section are from the area of monotonic reasoning. This section will give examples from non-monotonic reasoning. As we have already mentioned, we hope that the idea of *LDS* will unify these two areas.

EXAMPLE 3.1 (ORDERED LOGIC) An ordered logic database is a partially ordered set of local databases, each local database being a set of clauses. Figure 46.6 describes an ordered logic database.

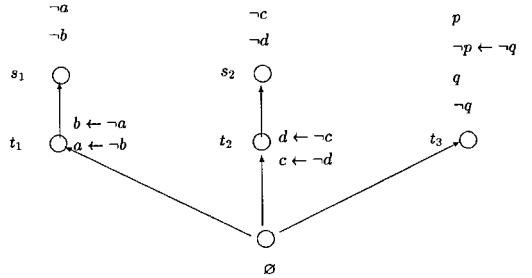


Figure 46.6

The local databases are labeled t_1, t_2, t_3, s_1, s_2 and ϕ and are partially ordered as in the figure.

To motivate such databases, consider an ordinary logic program $C_1 = \{p \leftarrow \neg q\}$. The computation of a logic program assumes that, since q is not a head of any clause, $\neg q$ is part of the data (this is the *closed world assumption*). Suppose we relinquish this principle and adopt the principle of asking an *advisor* what to do with $\neg q$. The advisor might say that $\neg q$ succeeds or might say that $\neg q$ fails. The advisor might have his own program to consult. If his program is C_2, he might run the goal q (or $\neg q$), look at what he gets and then advise. To make the situation symmetrical and general we must allow for Horn programs to have rules with both q and $\neg q$ (i.e. literals) in heads and bodies and have any number of negotiating advisors. Thus we can have $C_2 = \{\neg q\}$, $C_1 = \{q \leftarrow \neg q\}$ and C_1 depends on C_2. Ordered logic develops and studies various aspects of such an advisor system which is modeled as a partially ordered set of theories. Such a logic is useful, for example for multi-expert systems where we want to represent the knowledge of several experts in a single system. Experts may then be ordered according to an 'advisory' or a relative preference relation.

A problem to consider is what happens when we have several advisors that are in conflict. For example, C_1 depends on C_2 and C_1 depends on C_3. The two advisers, C_2 and C_3, may be in conflict. One may advise $\neg q$, the other q. How to decide? There are several options:

1. We can accept q if all advisors say 'yes' to q.
2. We can accept q if at least one advisor says 'yes' to q.
3. We can apply some non-monotonic or probabilistic mechanism to decide.

If we choose options (1) or (2) we are essentially in modal logic. To have a node t and to have $?q$ refer to advisors $t_1, \ldots, t_n$ with $t < t_i$, $i = 1, \ldots, n$ is like considering $?\Box q$ at t in modal logic with $t_1, \ldots, t_n$ possible worlds in option 1 and like considering $\Diamond q$ at t in option (2). Option (3) is more general, and here an *LDS* approach is most useful. We

see from this advisor's example an application area where the labels arise naturally and usefully. The area of ordered logic is surveyed in Vermeir and Laenens (1990).

EXAMPLE 3.2 (DEFEASIBLE LOGIC) This important approach to non-monotonic reasoning was introduced by Nute (1994). The idea is that rules can prove either an atom q or its negation $\neg q$. If two rules are in conflict, one proving q and one proving $\neg q$, the deduction that is stronger is from a rule whose antecedent is logically more specific. Thus the database:

t_1: Bird $(x) \rightarrow$ Fly (x)
t_2: Big $(x) \wedge$ Bird $(x) \rightarrow \neg$ Fly (x)
t_3: Big (a)
t_4: Bird (a)

$t_1 < t_2$
t_3
t_4

can prove:

$t_2 t_3 t_4$: $\neg$Fly(a)
$t_1 t_4$: Fly(a)

The database will entail $\neg$ Fly (a) because the second rule is more specific.

As an *LDS* system the labeling of rules in a database Δ is very simple. We label a rule by its antecedent. The ordering of the labels is done by logical strength relative to some background theory Θ (which can be a subtheory of Δ of some form). Deduction pays attention to the strength of labels.

EXAMPLE 3.3 (FALLACIES) The reader should note that our point of view and the use of labels is genuinely more general and is capable of yielding more. We describe an unexpected application of our view. There is a serious, well-motivated and well-organized community, the informal logic and argumentation community, studying the nature of human reasoning and argumentation in general and attempting to foundationally explain the role of the fallacies in human arguments. Fallacies are argument structures which appear to be correct and convincing, but are actually wrong. Many of them can be effectively used in some situations, but not in others. Any account of real life human practical reasoning must give account of the fallacies. In Hamblin (1970), a fallacy is an argument that "seems to be valid but is not so."

The handling of the fallacies in the traditional literature is divergent between two extremes.

There are those who reject the fallacies as not having any logical value (see Lambert and Ulrich 1980) and there are those who try to see some logic in them. Among the latter are John Woods and Douglas Walton. They believe that the traditional fallacies can be explained within the framework of other logics, such as inductive logics,

non-classical logics, logics of plausible reasoning, relevance logics and more. The Woods–Walton approach, see Walton (1990); Woods (1988); Woods and Walton (1989), is successful in many cases in showing and explaining how some fallacies are really not fallacies. However the Woods–Walton approach was in principle criticized by F. H. Van Emeron and R. Grootendorst (1992), who point out that this approach, although successful in many cases, creates new and serious problems. Van Emeron and Grootendorst, justly point out that every fallacy, in this approach needs, so to speak, its own logic. Van Emeron and Grootendorst say:

> For practical purposes this approach is not very realistic. In order to be able to carry out the analyses, a considerable amount of logical knowledge is required. There are also some theoretical disadvantages inherent in this approach. By relying on so many logical systems, one only gets fragmentary description of the various fallacies, and no overall picture of the domain of the fallacies as a whole. Ideally, one unified theory that is capable of dealing with all the different phenomena, is to be preferred. (van Emeron and Grootendorst 1992: 103)

We agree with both Van Emeron–Grootendorst and with Woods–Walton. There is indeed a possible candidate for a unifying logic in which suitable theories for practical reasoning and the fallacies can be formulated. It is the framework of Labeled Deductive Systems.

This example is a preliminary study at classifying and explaining some of the fallacies in *LDS*.

Here we quote Douglas Walton's words

> until we have a clearer definition of theoretical reasoning, it is not possible to refute the argument that there is one underlying kind of reasoning that has two uses – practical problem solving and theoretical problem solving. (Walton 1990: 353)

Well-known among the fallacies is the fallacy *ad hominem*, the fallacy of attacking not the argument but the person presenting it. This kind of reasoning is sometimes acceptable and sometimes not. It is generally considered nonlogical, although admittedly extensively used by the human practical reasoner. In our framework, this fallacy has a natural place.

Consider the notion of a database Δ. This is a structure of declarative units of the form $t : A$, where t is the label and A the formula. The label t annotates A. Suppose the annotation indicates the priority of the formula A and that in an external ordering $<$ gives the relative strength of the priorities. Thus a priority database can be for example

$$\{t : A, s : B, t < s\}$$

t and s can be numbers of algebraic terms and $t < s$ indicates that B has a higher priority than A. This priority can be used in derivation. For example, in the presence of $A \rightarrow \neg C$, $B \rightarrow C$ of equal priority, C will be derived.

The data items A and B are formulas of the logic $\mathbf{L}_1$, which is applied to some application area. In many areas it is quite reasonable to have the labels themselves be

formulas α, β of another language and logic $\mathbf{L}_2$, describing the origin and nature of the data items, A, B. Some reasoning in $\mathbf{L}_2$ may be available to determine the priority (if any) of α and β. A formula $\Psi(\alpha, \beta)$ and a base theory Θ (possibly dependent on Δ) of $\mathbf{L}_2$ may be used for this purpose, that is we have:

$$\alpha \leq \beta \text{ iff } \Theta \vdash_2 \Psi(\alpha, \beta).$$

The simplest condition (in case $\mathbf{L}_2$ has some form of implication) is

$$\alpha \leq \beta \text{ iff } \Theta \vdash_2 \beta \to \alpha.$$

Note that our labels are wffs α of $\mathbf{L}_2$ labelling wffs A of $\mathbf{L}_1$ and the base theory Θ determines the priorities of labels. We now explain the logical force of the fallacy by an example. Suppose we are faced with the following deduction.

$\alpha : A \to \neg C$
$\beta : B \to C$
$\gamma : A$
$\gamma : B$
$\Theta \vdash_2 \beta \to \alpha$

We must conclude C, because β has higher priority than α. To counter this argument, we may either prove $\neg C$ from additional data or we may attack the source of information, that is add Θ_0 to Θ or try and show that $\Theta \cup \Theta_0 \nvdash_2 \beta \to \alpha$?, (Note that $\mathbf{L}_2$ reasoning is also non-monotonic!). This move appears to us as attacking, not the argument, but its source. However, in the correct context (priority logic) it is a correct move. Other fallacies which are explainable in this framework are *ad verecundiam*, appeal to unsuitable authority, where the labeling is incorrect and fallacies of irrelevance. A systematic study of the fallacies in our context will (hopefully) be done elsewhere.

To make the above database more concrete consider the following scenario. A man is imprisoned for fraud for a long period of time. During that period, medical evidence emerges that the prisoner has terminal cancer. The question is whether to release him from jail. One legal argument supports an early release. The problem seems to be that the prisoner made some threats during the trial and a social and psychological report cannot exclude the possibility that the prisoner might use his remaining free days for revenge. Our database now reads

$m : B$	medical file m supporting the statement that the prisoner has cancer
$p : A$	social workers report supporting the statement that the prisoner is seeking revenge
$\alpha : A \to \neg C$	legal precedents α supporting the rule that in case of possible revenge the prisoner should not be released

$\beta : B \rightarrow C$	legal reasoning β supporting that in case of cancer the prisoner should be released
$p < m$	medical files are stronger than 'psychological' files'

From the above data we can conclude

$$\beta * m : C$$

and

$$\alpha * p : \neg C$$

Since both β and m have higher priority, C will follow by the *flattening* process.

If we want to change the conclusion (to get $\neg C$), we must either attack the medical file m, discrediting the medical evidence or boost up the credibility of the psychological report.

EXAMPLE 3.4 (DEMPSTER–SHAFER RULE) The present example presents a very well-known rule of aggregation, the Dempster–Shafer rule. Our exposition relies on Ng and Subrahmanian (1994).

The algebra A we are dealing with is the set of all subintervals of the unit interval $[0,1]$. The Dempster–Shafer addition on these intervals is defined by

$$[a,b] \oplus [c,d] = \left[\frac{a \cdot d + b \cdot c - a \cdot c}{1-k}, \frac{b \cdot d}{1-k} \right]$$

where $k = a \cdot (1 - \mathrm{d}) + c \cdot (1 - \mathrm{b})$, where '$\cdot$', '$+$', '$-$' are the usual arithmetical operations. The compatibility condition required on a, b, c, d is

$$F([a, b], [c, d]) \equiv k \neq 1.$$

The operation $\oplus$ is commutative and associative. Let $\mathbf{e} = [0,1]$.

The following also holds:

- $[a, b] \oplus \mathbf{e} = [a, b]$
- For $[a, b] \neq [1,1]$ we have $[a, b] \oplus [0,0] = [0,0]$
- For $[a, b] \neq [0,0]$ we have $[a, b] \oplus [1,1] = [1,1]$
- $[a, b] \oplus [c, d] = \phi$ iff either $[a, b] = [0,0]$ and $[c, d] = [1,1]$ or $[a, b] = [1,1]$ and $[c, d] = [0,0]$.

In this algebra, we understand the declarative unit $[a, b] : A$ as saying that the probability of the event represented by A lies in the interval $[a, b]$. We have, of course

$$\frac{[a,b] : A \rightarrow B; [c,d] : A}{[a,b] \oplus [c,d] : B},$$

provided $F([a, b], [c, d])$ holds.

It is also possible to move to a higher language and write clauses of the form

$$t : (t_1 : A_1) \rightarrow ((t_2 : A_2) \rightarrow (t_3 : A_3))$$

which is more like the way clauses are used in traditional Dempster–Shafer applications.

4 Conclusion and Further Reading

Logic is widely applied in computer science and artificial intelligence. The needs of the application areas in computing are different from those in mathematics and philosophy. In response to computer science needs, intensive research has been directed in the area of nonclassical and non-monotonic logic. New logics have been developed and studied. Certain logical features, which have not received extensive attention in the pure logic community, are repeatedly being called upon in computational applications. Two features in logic seem to be of crucial importance to the needs of computer science and stand in need of further study. These are:

1. The meta-level features of logical systems
2. The 'logic' of Skolem functions and unification

The meta-language properties of logical systems are usually hidden in the object language. Either in the proof theory or via some higher-order or many-sorted devices. The logic of Skolem functions is nonexistent. Furthermore, the traditional presentation of classical and nonclassical logics is not conducive to bringing out and developing the features needed for computer science applications. The very concept of what is a logical system seems to be in need of revision and clarification. A closer examination of classical and nonclassical logics reveals the possibility of introducing a new approach to logic; the discipline of *Labeled Deductive Systems* (*LDS*) which, I believe, will not only be ideal for computer science applications but will also serve, I hope, as a new unifying logical framework of value to logic itself. What seem to be isolated local features of some known logics turn out to be, in my view, manifestations of more general logical phenomena of interest to the future development of logic itself.

Semantics for LDS logics is presented in my book on Fibring Logics (Gabbay 1998).

LDS is part of a more general view of logic. This view is discussed elsewhere (Gabbay 1991, 1996, forthcoming), however in brief, we claim the following. The new concept of a logical system is that of a *network* of *LDS* systems which has mechanisms for *communication* (through the labels, which code meta-information) and *evolution* or change.

Evaluation is a general concept which can embrace updating, abduction, consistency maintenance, action, and planning. The above statement of position is vague but it does imply that we believe that notions like abduction and updating are logical notions of equal standing to those of provability. See Gabbay and Woods (to appear).

Notes

1 The similarity with Gentzen sequents is obvious. A sequent $\Delta \vdash \Gamma$ is a relation between Δ and Γ. Such a relation can either be defined axiomatically (as a consequence relation) or be generated via closure conditions like $A \vdash A$ (initial) and other generating rules. The generating rules correspond to Gentzen rules. In many logics we have $\Delta \vdash \Gamma$ iff $\emptyset \vdash \wedge\Delta \rightarrow \vee\Gamma$, which gives an intuitive meaning to $\vdash$.

2 Recently logical systems were put forward by Makinson–Torre (2001) which do not satisfy reflexivity.

3 **Flat** is a function defined on any set of labels and giving as value a new label. To understand this, recall another function on numbers which we may call **Sum**. It adds any set of numbers to give a new number: their sum!

References

Anderson, A. R. and Belnap, N. D. (1975) *Entailment*. Princeton, NJ: Princeton University Press.

Basin, D., D'Agostino, M., Gabbay, D. M., Matthews, S. and Vigano, L. K. (eds.) (2000) *Labelled Deduction*. Dordrecht: Kluwer.

Van Emeron, F. H. and Grootendorst, R. (1992) *Argumentation, Communication and Fallacies*. New York: Lawrence Elbaum.

Fitting, M. (1983) *Proof Methods for Modal and Intuitionistic Logic*. Dordrecht: Kluwer.

Gabbay, D. M. (1981) *Semantical Investigations in Heyting's Intuitionistic Logic*. Amsterdam: Reidel.

Gabbay, D. M. (1985) Theoretical foundations for non-monotonic reasoning, in K. Apt (ed.), *Expert Systems, Logics and Models of Concurrent Systems* (pp. 439–59). Berlin: Springer Verlag.

Gabbay, D. M. (1992) Theory of algorithmic proof. In S. Abramsky, D. M. Gabbay and T. S. E. Maibaum (eds.), *Handbook of Logic in Theoretical Computer Science*, vol. 1 (pp. 307–408). Oxford: Oxford University Press.

Gabbay. D. M. (1998) *Fibring Logics*. Oxford: Oxford University Press.

Gabbay, D. M. and Woods, J. (to appear) *Agenda Relevance, I and II*.

Gabbay, D. M. (1969) The Craig interpolation theorem for intuitionistic logic I and II. In R. O. Gandy (ed.), *Logic Colloquium 69*, (pp. 391–410). Amsterdam: North Holland.

Gabbay, D. M. (1991) Abduction in labelled deductive systems, a conceptual abstract. In R. Krose and P. Siegel (eds.) *ECSQAU 91*, Lecture notes in Computer Science 548 (pp. 3–12). Berlin: Springer Verlag.

Gabbay, D. M. (1991) *Theoretical Foundations for Non Monotonic Reasoning, Part 2: Structured Non-Monotonic Theories*. In *SCAI '91, Proceedings of the Third Scandanavian Conference on AI*. (pp. 19–40). Amsterdam: IOS Press.

Gabbay, D. M. (1991) Modal and temporal logic programming II. In T. Dodd, R. P. Owens and S. Torrance (eds.), *Logic Programming – Expanding the Horizon* (pp. 82–123). New York: Ablex.

Gabbay, D. M. (1992) How to construct a logic for your application. In *Proceedings of the 16th German AI Conference, GWAI 92*, Springer Lecture Notes on AI, vol. 671, pp. 1–30.

Gabbay, D. M. (1992) Modal and temporal logic programming III: metalevel features in the object language. In L. F. del Cerro and M. Penttonen (eds.), *Non-Classical Logic Programming* (pp. 85–124). Oxford: Oxford University Press.

Gabbay, D. M. (1994) Labelled deductive systems and situation theory. In P. Aczel, D. Israel, Y. Katagin and S. Peters (eds.), *Situation Theory and Applications*, vol. 3 (pp. 89–118). Stanford, CA: CSLI.

Gabbay, D. M. (1996) *Labelled Deductive Systems*, vol. 1. Oxford: Oxford University Press.
Gabbay, D. M. (forthcoming) *A General Theory of Structured Consequence Relations*. To appear in a volume of substructured logics, ed. P. Schröder-Heister and K. Dosen. Oxford: Oxford University Press.
Gabbay, D. and Olivetti, N. (2000) *Goal Directed Proof Theory*. Dordrecht: Kluwer.
Gabbay, D. M. and Queiroz, R. J. G. B. (1992) Extending the Curry–Howard interpretation to linear, relevance and other resource logics. *Journal of Symbolic Logic*, 57, 1319–66.
Gabbay, D. M. and Woods, J. (to appear) *Reach of Abduction*, vols. 1 and 2.
Hamblin, C. L. (1970) *Fallacies*. London: Methuen.
Kraus, S., Lehmann, D. and Magidor, M. (1990) Preferential models and cumulative logics. *Artificial Intelligence*, 44, 167–07.
Lambert, K. and Ulrich, W. (1980) *The Nature of Argument*. New York: Macmillan.
Lehmann, D. (1989) What does a conditional knowledge base entail? In *KR 89, Toronto, May 89* (pp. 1–18). New York: Morgan Kauffman.
Makinson, D. (1988) General theory of cumulative inference. In M. Reinfrank, J. de Kleer, M. L. Ginsberg and E. Sandewall (eds.), *Non-monotonic Reasoning*. Springer Verlag Lecture Notes on Artificial Intelligence No. 346.
Makinson, D. (1994) General patterns in nonmonotonic reasoning. In D. M. Gabbay, C. J. Hogger, and J. A. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 3 (pp. 35–110). Oxford: Oxford University Press.
Makinson, D. and van der Torre, L. (2001) Constraints for input/output logics. *Journal of Philosophical Logic*, 30, 155–85.
Ng, R. and Subrahmanian, V. (1994) Dempster–Shafer logic programs and stable semantics. In J. N. Crossley and J. Be Remmel (eds.), *Logical Methods* (pp. 654–704). Birkhauser.
Nute, D. (1994) Defeasible logic. In D. M. Gabbay, C. J. Hogger, and J. A. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 3 (pp. 353–98). Oxford: Oxford University Press.
Scott, D. (1974) Completeness and axiomatizability in many valued logics. *Proceedings of Tarski Symposium*, American Mathematical Society, Providence, Rhode Island, 411–36.
Tarski, A. (1956) On the concept of logical consequence, in Polish (1936). Translation in *Logic Semantics Metamathematics*. Oxford: Oxford University Press.
Vermeir, D. and Laenens, E. (1990) An overview of ordered logic in *Abstracts of the Third Logical Biennial*. Bulgaria: Varga.
Vigano, L. (1999) *Labelled Non-classical Logics*. Dordrecht: Kluwer.
Walton, D. (1990) *Practical Reasoning*. New York: Rowman and Littlefield.
Woods, J. (1988) Are fallacies theoretical entities? *Informal Logic*, 10, 67–76.
Woods, J. and Walton, D. (1989) *Fallacies: Selected Papers, 1972–1982*. Dordrecht: Kluwer.

This page intentionally left blank

Resources for Further Study

A number of valuable resources are available for further study of philosophical logic. In addition to the books and articles cited in the references at the end of each chapter included in this volume, there are four general categories of resources that can be consulted for information about the history and current research developments in philosophical logic. Additional materials can be found by soliciting advice from logicians, philosophers, and mathematicians at local colleges and universities.

Logic Handbooks

Many university and independent presses publish books on or related to mathematical and philosophical logic. There are also several special series of original monographs in logic that are worth investigating. The literature is too vast to justify a selection of the individual books that have contributed to the development of logic. We can nevertheless identify special categories of texts of special interest, beginning with handbooks and book series dedicated to logic and philosophical logic. Here are some recent relevant publications:

Boyer, Robert S. (1988) *A Computational Logic Handbook*. Boston, MA: Academic Press.

Sherwood, John C. (1960) *Discourse of Reason: A Brief Handbook of Semantics and Logic*. New York: Harper & Row.

Handbook of Fuzzy Computation, ed. Enrique H. Ruspini, Piero P. Bonissone and Witold Pedrycz. Philadelphia, PA: Institute of Physics Publications, 1998.

Handbook of Logic and Language, ed. Johan van Benthem and Alice ter Meulen. Cambridge, MA: MIT Press, 1997.

Handbook of Logic in Artificial Intelligence and Logic Programming, ed. Dov Gabbay, C. J. Hogger, and J. A. Robinson. Oxford: Clarendon Press, 1993–8.

Emmet, E. R. (1984) *Handbook of Logic*. Totowa, NJ: Rowman & Allanheld.

Cowan, Sam (1985) *Handbook of Mathematical Logic*. Englewood Cliffs, NJ: Prentice-Hall.

Handbook of Mathematical Logic, ed. Jon Barwise et al. Amsterdam: North-Holland, 1977.

Handbook of Philosophical Logic, ed. Dov Gabbay and F. Guenthner. Dordrecht: Kluwer, 1983–9.

Handbook of Tableau Methods, ed. Marcello D'Agostino, Dov Gabbay, Reiner Hähnle, and Joachim Posegga. Dordrecht: Kluwer, 1999.

Logic Designer's Handbook: Circuits and Systems, ed. E. A. Parr. Oxford: Newnes, 1993.

Non-Classical Logics and Their Applications to Fuzzy Subsets: A Handbook of the Mathematical Foundations of Fuzzy Set Theory, ed. Ulrich Höhle and Erich Peter Klement. Dordrecht: Kluwer, 1999.

Programmable Logic Handbook, ed. Geoff Bostock. Oxford: Newnes, 1993.

Programmable Logic Handbook: PLDs, CPLDs, and FPGAs, ed. Ashok K. Sharma. New York: McGraw-Hill, 1998.

Logic Books Series

Addison-Wesley Series in Logic. Reading: Addison-Wesley Publishing Co.
Algebra, Logic and Applications. Amsterdam: Gordon and Breach Science Publishers.
Applied Logic Series. Dordrecht: Kluwer Academic.
Clarendon Library of Logic and Philosophy. Oxford: Clarendon Press.
De Gruyter Series in Logic and its Applications. Berlin: Walter de Gruyter.
History of Logic. Napoli: Bibliopolis.
International Series in Logic Programming. Reading: Addison-Wesley.
Lecture Notes in Logic. New York: Springer-Verlag.
Library of Philosophy and Logic. Oxford: Basil Blackwell.
Logic and Computation in Philosophy. Oxford: Oxford University Press.
Logic Programming. Cambridge: The MIT Press.
Oxford Logic Guides. Oxford: The Clarendon Press.
Perspectives in Mathematical Logic. New York: Springer-Verlag.
Progress in Computer Science and Applied Logic. Boston: Birkhäuser.
Studies in Logic and Computation. Oxford: Clarendon Press.
Studies in Logic and the Foundations of Mathematics. Amsterdam: North-Holland Publishing.
Studies in Logic, Language, and Information. Stanford: CSLI Publications.
SUNY Series in Logic and Language. Albany: State University of New York Press.
Trends in Logic. Dordrecht: Kluwer Academic.

Journals

Numerous journals are devoted specifically to topics in symbolic logic. Some of these feature articles that are highly technical contributions to mathematical logic, while others specialize in research papers concerning historical and philosophical aspects of logic. Many philosophy journals that are not designated logic journals also frequently include essays about logic or make use of logic in presenting philosophical arguments. The following are the principal journals in the field that can be considered as primary sources for contemporary work in mathematical and philosophical logic. Details can be found by consulting publishers' websites indicated below and publications and computer databases that index journal articles in the field, including but not limited to *The Philosophers' Index*, *Web of Science*, *Humanities Index*, *Social Science Index*, and *Mathematical Reviews*. The most important journals in logic or in philosophy that often publish work in logic include: *Analysis*; *Annals of Pure and Applied Logic*; *Archive for Mathematical Logic*; *Bulletin of Symbolic Logic*; *Bulletin of the Section of Logic*, *Fundamenta Mathematicae*; *Historia Mathematica*; *History and Philosophy of Logic*; *Informal Logic*; *Israel Journal of Mathematics*; *Journal of Applied Non-Classical Logics*; *Journal of Formalized Mathematics*; *Journal of Logic and Computation*;

Journal of Logic, Language and Information; *Journal of Mathematical Logic*; *Journal of Philosophical Logic*; *Journal of Philosophy*; *Journal of Symbolic Logic*; *Logic Journal of the IGPL*; *Logical Analysis and History of Philosophy*; *Logique et Analyse*; *Mathematical Logic Quarterly*; *Mind*; *Modern Logic*; *Nordic Journal of Philosophical Logic*; *Notre Dame Journal of Formal Logic*; *Noûs*; *Philosophia Mathematica*; *Studia Logica*; *Synthese*; *Zentralblatt MATH*.

Internet

The internet is another increasingly useful source of information for following new developments in philosophical logic. Some logicians post results or discussions of philosophical topics related to symbolic logic on the web. Publishers and journals also announce new publications of interest to logicians, including journal contents, which can be used to gather information about new findings. Conferences in symbolic logic are also frequently listed, with lists of speakers and presentation titles or abstracts. There are also logic chat rooms, and special interest networks in which logicians present problems and exchange ideas about logic. The best general advice about using the internet is to be persistent in pursuing interesting leads, but also to be highly selective and to treat whatever is found there with a grain of salt and a healthy skepticism. There is no refereeing of information on the web, and people are free to post whatever ideas they like, without editorial scrutiny. Some of the most important and interesting websites for philosophical logic include:

http://www.nd.edu/~ndjfl/index.html
Notre Dame Journal of Formal Logic

http://www.aslonline.org/
Association for Symbolic Logic

http://www.earlham.edu/~peters/philinks.htm
Guide to Philosophy on the Internet
(Peter Suber, Philosophy Department, Earlham College)

http://www-personal.monash.edu.au/~dey/phil/
Philosophy in Cyberspace

http://web.phil.ufl.edu/SEP/index.html
Society for Exact Philosophy

http://www.math.uu.se/logik/logic-server/
Research groups in Logic and Theoretical Computer Science

http://www.math.uu.se/logik/logic-server/collection.html
List of Research Groups in Logic and Theoretical Computer Science Worldwide

http://www.math.fu-berlin.de/~dvmlg/
Deutsche Vereinigung für Mathematische Logik und für Grundlagen der Exakten Wissenschaften

http://www.wkap.nl/jrnlsubject.htm/5+0+0+0
Listing of Journal Homepages by Subject, including Logic

http://www.logic.at/kgs/www_logicien.html
Resources in Computer Science, Logic, Mathematics and Philosophy
(Kurt Gödel Society)

www.jneedle.demon.co.uk/vague/articles.htm
Online Articles on Concept and Logic of Vagueness

http://www.uni-bonn.de/logic/world.html
Mathematical Logic around the world
(service provided by the Mathematical Logic Group, University of Bonn, and the Institute for Logic, University of Vienna)

http://www.rbjones.com/rbjpub/logic/index.htm
Factasia

http://www2.galaxy.com/galaxy/Humanities/Philosophy/Logic.html
Galaxy

http://logic.stanford.edu:80/
Logic Group at Stanford University

http://www.sjsu.edu:80/depts/itl/
Mission: Critical
(San Jose State University's Critical Thinking Web Page)

Organizations

There are special philosophical organizations that promote the study of symbolic and philosophical logic. Many of them publish newsletters or maintain websites to which you can subscribe or to which a local library might offer access. A few of the most important philosophical societies related to logic include: Association for Automated Reasoning; Association for Logic Programming; Association for Symbolic Logic; Associazione Italiana di Logica e sue Applicazioni (University of Genoa); Austrian Ludwig Wittgenstein Society; Belgian National Center of Research in Logic; Berkeley Group in Logic and the Methodology of Science (University of California at Berkeley); Bertrand Russell Society; British Logic Colloquium; Center for Critical Thinking; Center for Fuzzy Logic, Robotics, and Intelligent Systems (Texas A&M University); Centre de Logique et de Philosophie des Sciences; COMPULOG Americas; Deutsche Vereinigung für Mathematische Logik und für Grundlagen der Exakten Wissenschaften; European Association for Computer Science Logic (University of Udine, Italy); European Association for Logic, Language, and Information (Amsterdam); Fuzzy Logic Research (University of Missouri at Columbia); Helsinki Logic Group; Indiana University Logic Group; Institut für mathematische Logik und Grundlagen der Mathematik (Freiburg); Institut für mathematische Logik und Grundlagenforschung (Münster); Institute for Logic, Language, and Computation (Amsterdam); Interest Group in Pure and Applied Logics (IGPL) (Imperial College, London); Italian Society of Logic and Philosophy of Science; Kurt Gödel Society; Laboratory for Applied Logic (Brigham

Young University); Logic Group at Stanford University; Mathematical Logic Group (University of Bonn); Mind Association; Swiss Society for Logic and Philosophy of Sciences (Zurich); Society for Exact Philosophy; Society for Symbolic Logic. There are also regular sessions and special symposia on topics in philosophical logic sponsored by the three annual meetings (Eastern, Central, and Pacific Divisions) of the American Philosophical Association.

Index