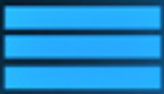


CAMBRIDGE STUDIES IN PHILOSOPHY

JOSHUA GERT

BRUTE RATIONALITY

NORMATIVITY AND HUMAN ACTION

CAMBRIDGE www.cambridge.org/9780521833189

This page intentionally left blank

Brute Rationality
Normativity and Human Action

This book presents a new account of normative practical reasons and the way in which they contribute to the rationality of action. Rather than simply 'counting in favor of' actions, normative reasons play two logically distinct roles: requiring action and justifying action. The distinction between these two roles explains why some reasons do not seem relevant to the rational status of an action unless the agent cares about them, while other reasons retain all their force regardless of the agent's attitude. It also explains why the class of rationally permissible action is wide enough to contain not only all morally required action, but also much selfish and immoral action. The book will appeal to a range of readers interested in practical reason in particular, and moral theory more generally.

Joshua Gert is Assistant Professor at the Department of Philosophy, Florida State University. He has published in a number of philosophical journals including *American Philosophical Quarterly*, *Ethics*, and *Noûs*.

CAMBRIDGE STUDIES IN PHILOSOPHY

General editors E. J. LOWE and WALTER SINNOTT-ARMSTRONG

Advisory editors
JONATHAN DANCY *University of Reading*
JOHN HALDANE *University of St Andrews*
GILBERT HARMAN *Princeton University*
FRANK JACKSON *Australian National University*
WILLIAM G. LYCAN *University of North Carolina, Chapel Hill*
SYDNEY SHOEMAKER *Cornell University*
JUDITH J. THOMSON *Massachusetts Institute of Technology*

RECENT TITLES
JOSHUA HOFFMAN & GARY S. ROSENKRANTZ *Substance among other categories*
PAUL HELM *Belief policies*
NOAH LEMOS *Intrinsic value*
LYNNE RUDDER BAKER *Explaining attitudes*
HENRY S. RICHARDSON *Practical reasoning about final ends*
ROBERT A. WILSON *Cartesian psychology and physical minds*
BARRY MAUND *Colours*
MICHAEL DEVITT *Coming to our senses*
SYDNEY SHOEMAKER *The first-person perspective and other essays*
MICHAEL STOCKER *Valuing emotions*
ARDA DENKEL *Object and property*
E. J. LOWE *Subjects of experience*
NORTON NELKIN *Consciousness and the origins of thought*
PIERRE JACOB *What minds can do*
ANDRE GALLOIS *The world without, the mind within*
D. M. ARMSTRONG *A world of states of affairs*
DAVID COCKBURN *Other times*
MARK LANCE & JOHN O'LEARY-HAWTHORNE *The grammar of meaning*
ANNETTE BARNES *Seeing through self-deception*
DAVID LEWIS *Papers in metaphysics and epistemology*
MICHAEL BRATMAN *Faces of intention*
DAVID LEWIS *Papers in ethics and social philosophy*
MARK ROWLANDS *The body in mind: understanding cognitive processes*
LOGI GUNNARSSON *Making moral sense: beyond Habermas and Gauthier*
BENNETT W. HELM *Emotional reason: deliberation, motivation, and the nature of value*
RICHARD JOYCE *The myth of morality*
ISHTIYAQUE HAJI *Deontic morality and control*
ANDREW NEWMAN *The correspondence theory of truth*
JANE HEAL *Mind, reason, and imagination*
PETER RAILTON *Facts, values and norms*
CHRISTOPHER S. HILL *Thought and world*
WAYNE DAVIS *Meaning, expression and thought*
ANDREW MELNYK *A physicalist manifesto*
JONATHAN L. KVANVIG *The value of knowledge and the pursuit of understanding*
WILLIAM ROBINSON *Understanding phenomenal consciousness*
MICHAEL SMITH *Ethics and the a priori*
D. M. ARMSTRONG *Truth and truthmakers*

Brute Rationality

Normativity and Human Action

Joshua Gert

Florida State University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521833189

First published in print format 2004

ISBN-13 978-0-511-21121-8 eBook (EBL)
ISBN-10 0-511-21298-4 eBook (EBL)

ISBN-13 978-0-521-83318-9 hardback
ISBN-10 0-521-83318-3 hardback

To my parents, my sister Heather,
and my wife Victoria

Contents

Preface and acknowledgements

I would guess that the first time I read any real philosophy was when I was about ten years old. Sitting and reading aloud on the living room couch with my father, I took the part of Hylas in Berkeley's *Three Dialogues*. It is a happy memory for me, despite the fact that I turned out, as those familiar with that dialogue will know, not to have very many lines, and always to be wrong. I also have a very distinct visual memory, from roughly the same period, of the moment my father presented the open question argument to me. He didn't explain the problems with the argument, and if he had, I doubt I would have understood what he was saying. I was just sophisticated enough that the argument seemed to me to show exactly what Moore thought it showed. I didn't like having to believe in non-natural properties. I didn't even have any clear idea what they were. But I had to do it. Twenty-seven years later, I think I might have gotten out of the problem.

Those two memories may be the oldest ones I have of doing any philosophy with my father, but they are by no means the only ones. Later memories are less distinct, probably because philosophical discussion became as common as eating dinner. But as far as I can recall, all of these memories of talking philosophy with my father – of arguing and criticizing, and, generally, of being shown that I didn't know what I was talking about – are uniformly happy. My love for philosophy is, I am sure, continuous with my great love for my father. There is no doubt that it is my father who has had the most profound philosophical impact on me. Indeed, I am pleased to think of myself, in many parts of this book, as refining, building upon, and modifying his views, just as other philosophers have refined, built upon, and modified the views of their advisors. Given that my father's influence began early, I cannot adequately express how lucky I feel that so many of his starting points have turned out to be so fruitful. For it is hard to deny that the students of Kantians tend to become Kantians, and the students of Humeans tend to become Humeans. When I consider

the strength of this law of philosophical inheritance, and realize how easily I might have fallen under the spell of a mainstream view (or, worse, a currently fashionable one), I am reminded of the huge role fortune plays in all the achievements for which we would like to take exclusive credit.

Other than my father, I would like to thank a number of people with whom I have had profitable conversations or correspondence on the topics I address in the following chapters. I should single out Daniel Callcut, Charles Chastain, and John Deigh, both for the sheer volume of conversation, and also for entering into the discussion with sufficient sympathy to understand the whole picture. Thanks also to Peter Achinstein, Ken Akiba, Robert Audi, John Broome, María Victoria Costa, Jonathan Dancy, Heather Gert, Peter Hylton, Anthony Laden, Paul McNamara, Al Mele, Andrew Melnyk, Karen Neander, Brian Neuslein, Joseph Raz, Thomas Scanlon, Jerome Schneewind, Paul Weirich, and Susan Wolf.

I also owe a great debt to Oscar Jorge Mainoldi, who accidentally taught me Spanish, and who superintended the writing of virtually the whole of this book during five successive summers at what must be the world's most fertile environment for the production of philosophy: the bar/café "Portofino," at the corner of 13th and 42nd, in La Plata, Argentina. For being among the truthmakers behind this fact, thanks also to Daniela "Pichu" Memna, Mercedes Mirabella, Martín Zamudio, Rubén Peralta, and Sebastián Alvarez.

Chapter 1 takes the form it does largely because I was invited to give an overview of my account of reasons and rationality at the Universidad Nacional de La Plata in La Plata, Argentina, in the summer of 2002. I am very grateful to Pedro Karczmarczyk and Martín Daguerre for organizing that talk, and I am grateful to all the members of that audience for their patience with my Spanish. The material for chapter 2 was previously published as "Practical Rationality, Morality, and Purely Justificatory Reasons" in *American Philosophical Quarterly* 37 (3), 227–43. I am grateful to the Executive Editor of *APQ*, Nicholas Rescher, for permission to use that material here. Most of the material for chapters 3 and 7 was previously published in the *Southern Journal of Philosophy* as "Skepticism about Practical Reasons Internalism" 39 (1), 59–77 and "Two Concepts of Rationality" 41 (3), 367–98, and some material for chapter 3 was also taken from "Korsgaard's Private-Reasons Argument" *Philosophy and Phenomenological Research* 64 (2), 303–24. I thank the Editor of the *Southern Journal*, Nancy Simco, and the editors of *PPR* for permission to use that material here. Chapter 4 appears as "A Functional Role Analysis of Reasons" in

Philosophical Studies (2004). A version of chapter 5 was published as "Requiring and Justifying: Two Dimensions of Normative Strength" *Erkenntnis* 59 (1), 5–36, and appears here with kind permission of Kluwer Academic Publishers. A distant ancestor of chapter 7 was published, in Spanish, in *Revista Latinoamericana de Filosofía* 25 (2), 255–81, after having been presented to the members of the Centro de Investigaciones Filosóficas in Buenos Aires, Argentina, in 1998. I am grateful to María Julia Bertomeu for her invitation to address this group. Chapter 8 first appeared in *The Philosophical Forum* 34 (1), 53–72, and chapter 9 appeared in *Noûs* 37 (3), 417–46. I acknowledge the kind permission of Blackwell Publishing to reprint both of them here.

Finally, I should thank Florida State University for summer funding provided through their FYAP Summer Grant program. It was during the summer in which I received this funding that I was able to revise the manuscript to deal – I cannot say how successfully – with the comments of Russ Shafer-Landau and Michael Ridge, who reviewed the original manuscript for Cambridge University Press. My final thanks go to them for their sympathetic and open-minded attitude, and for many useful criticisms.

1

What would an adequate theory of rationality be like?

THE FUNDAMENTAL NORMATIVE NOTION

When we argue with other people about what to do, very often we appeal to principles. Certainly when philosophers offer moral theories, and argue that we should be moral, they appeal to principles. And even when we, or they, offer reasons in place of principles, it is reasonable to think of such arguments as shorthand for appeals to principles. For no one would advocate an action simply because there was *some* reason in its favor, if it were clear that there were compelling reasons against performing it. Thus when reasons are cited in arguments, there is some idea that all the relevant reasons, taken together, support the action. This implies that there is some principle in the background that produces overall verdicts based on all those reasons: perhaps it is the simple principle 'perform the action supported by the most reasons', or perhaps it is some more complicated principle. One cites particular reasons in order to suggest that those reasons are sufficient to determine the outcome of the application of such a principle. The very plausible idea that two actions to which the same reasons are relevant must have the same rational status also suggests that reason-based arguments are backed by a unique principle: a principle that takes those reasons as input and yields the status of the action as output.

When a principle is made explicit in an argument, it is often appropriate to ask 'Why should I follow that principle?' And when an answer is given, in terms of some other principle, it is often appropriate to ask exactly the same question. In some cases there will be no good answer to this question, and then the recommendations that flow from the principle may lose their authority. There is, however, a significant philosophical tradition according to which this sequence of principles and questions, and more basic principles and further questions, cannot go on forever. At some point, after the articulation of one of these principles, it will no longer make sense, or be appropriate, to ask 'But why should I follow

that principle?' That is, there is a philosophical tradition that asserts the existence of a *fundamental normative principle applicable to action*. Perhaps this tradition goes back as far as Aristotle, who asserted that there was one governing end according to which all human action was to be judged. Hume, when he argued that reason cannot by itself direct the will, was reacting against the majority opinion of his contemporaries, according to whom reason *could* do so. That is, the philosophers against whom Hume was arguing held that if it could be shown that reason required or prohibited an action, that was the end of the practical argument about that action: no further appeal could possibly be made that could legitimately alter such a judgment. Kant also is a prominent member of this tradition, advocating the existence of a categorical imperative that tells one how one *must* act, and against which no further consideration can have any legitimate force.

Contemporary philosophers also defend the existence of a fundamental normative principle, or set of principles. Indeed, this is the sense of 'rational' that is central to contemporary ethical theorizing. For example, Stephen Darwall writes that "It is part of the very idea of the [rationally normative system] that its norms are *finally authoritative* in settling questions of what to do." Thomas Nagel writes that it should not be possible to ask why one should do what one has reason to do, and that for this reason there cannot be a *justification* for acting rationally. And Allan Gibbard's notion of rationality "settles what to do . . . what to believe, and . . . how to feel."[1] According to all of these philosophers it is a conceptual truth that there cannot be a sufficient reason to act irrationally and that there is a reason *not* to do so. Therefore, according to these philosophers, the question 'Could I have a sufficient reason to do an irrational act?' is as misguided (or trivial) as the question 'Could there be an unmarried bachelor?'

When we are presented with any proposal regarding this fundamental normative principle, there are two tests we can apply to see whether it is adequate. The first is to see whether the question 'Why should I always follow *that* principle?' makes clear sense. If it does make sense, this casts the fundamental nature of the principle into doubt. For the

[1] Darwall (1983), pp. 215–16; Nagel (1970), pp. 1–9; Gibbard (1990), esp. p. 49. See also Korsgaard (1996a), p. 104 and Smith (1994), pp. 150ff. Smith claims that it is all and only reasons which spring from the norms of rationality that make actions desirable. Of course there are other conceptions of rationality. Robert Nozick (1993), pp. 40, 117 is concerned with the human *faculty* of rationality, and is content to ask about its purpose or function.

principle is supposed to the most basic – the principle that stands behind all others. If the above question makes sense, then the putative fundamental principle certainly is not wearing its fundamental nature on its face. That is, it does not appear to be the end of the normative road. The second test is to see whether one could ever sensibly offer reasons for acting against the principle. If this is a real possibility then the principle cannot be the fundamental principle that tells us how we ought always to act. To illustrate these tests, it may be useful to use them to disqualify one possible fundamental normative principle: always act so as to maximize the satisfaction of your preferences.[2] Does it make sense to ask 'But why should I always act so as to maximize the satisfaction of my preferences?' Yes, it does. For one could elaborate the question in this way: 'Why should I always act so as to maximize the satisfaction of my preferences, if I know my preferences are the result of a brain defect that tends to produce self-destructive preferences?'[3] This failure to pass the first test is related to the way in which the proposed principle will also fail the second test. For one way of sensibly offering a reason to act against the principle is to say 'But if you follow this principle you will cause yourself a lot of pain, without any benefit.'

The second of the above tests is quite clearly one which a fundamental normative notion must pass. If there can be an adequate reason to act against a principle, that principle cannot be telling us how we ought always to be acting. The first test, however, is more slippery, and it may be useful to show how a principle may pass it without at first seeming to do so. Consider then the following:

> One should never perform an action that will harm oneself unless it will bring some compensating benefit to someone (perhaps oneself). All other actions are rationally permitted.

It seems obvious that one could sensibly ask 'Why should I always follow this principle?' One reason it seems obvious is that there seems to be an answer. For example, one might offer 'Because then one will avoid

[2] It is unclear if any contemporary philosophers hold such a simple version of this view. But the criticisms offered here also tell against more sophisticated versions of such principles. For a very clear presentation of these criticisms see Ripstein (2001).

[3] This is not the place to descend into arguments about the various ways in which one might patch up the suggested principle. But it is worth mentioning that the strategy of ruling out desires that are the result of, say, a brain defect, is not a simple one. For we use the notion of rational action in determining what counts as a brain *defect*, rather than (say) a statistically rare configuration of neurons.

suffering harms.' But in fact that is not an answer, since it is false that if one successfully follows this principle, one will necessarily avoid suffering harms. This is because the principle permits one to suffer harms in cases in which one will thereby produce compensating benefits for someone else. One might then suggest the following amended answer: 'Because one will avoid suffering harms, except in cases in which one will thereby produce compensating benefits for someone.' What is important to see is that with this amended answer one has ceased to offer a *further* reason to obey the principle. One has simply pointed out that by following the principle one follows the principle. Of course, this brief discussion has not shown that the above principle actually does pass the first test. It only shows one way in which a principle may misleadingly appear to fail it. Moreover, though the above principle may in fact pass this one particular test, it may be inadequate for other reasons.

This book is part of the tradition that seeks to discover and defend a fundamental normative principle applicable to action – of course by some means other than the production of a still more fundamental principle. That is, it seeks to provide an account of a principle that passes both of the tests mentioned above. It is devoted entirely to this principle, and not to its employment in arguing for further normative claims. In particular, no moral view is advocated, although it will be clear that the account has significant implications for the development of moral views.

In the phrase 'fundamental normative principle,' the word 'fundamental' should not be taken to mean 'most important.' For there are many other normative principles that, in different contexts, are likely to be more important and more salient than the principle that is the central topic of this book. Of course, we should never follow these more salient principles if they can be shown to violate the fundamental one: that *is* part of what it means for it to be fundamental. Another part of what it means is that if it is clear that an action does *not* violate the fundamental principle then there may be nothing we can say to dissuade even a rational agent from performing it – the agent may remain perfectly rational in resisting all our arguments. As will become clear later in the book, this means that the fundamental normative principle gives agents a very wide scope in making decisions about how to act. Because of the lack of guidance that the principle provides, some may be tempted to think that it cannot really be fundamental. But that is to confuse being fundamental with being *most generally useful*, or with being *salient*. It will turn out that, because we almost always act rationally without having to think

about it, the fundamental normative principle will very rarely be of much use in particular decisions. It will not tell us, for example, which career to choose, or whether to marry, or to have children, or whether to pursue wealth over enlightenment. As I will argue in various ways in what follows, it will not even tell us whether to take the high moral road, or the low. These questions we must answer for ourselves – they are *choices*, and it is futile to search for a basic principle that will authoritatively hand us the correct answer. In a limited number of cases I have found that when people are tempted to act against the fundamental normative principle, it is sometimes effective simply to point this out. This tends to bring the real source of the temptation into clearer focus, which helps in resisting it. But the primary usefulness of a clear view of the fundamental normative principle is not – at least directly – practical. Rather, it is *theoretical*: the principle will figure in an explanation of what it is for an action to be rational, in a sense that is closely connected with mental functioning. This notion, in turn, is often indispensable in restricting the scope of 'everyone' as it is used in philosophical theories (such as contractualism). The principle will also play a role in explaining why we should want to be rational, in that sense. And of course the fundamental principle will have many *indirect* practical implications, for very often such a principle plays an obvious and central role in the development of moral theory. And a moral theory, if it is clear, may have significant practical implications for people who care about morality.

RATIONALITY AND MENTAL FUNCTIONING

It seems fairly clear that whatever the fundamental normative notion might be, it will use the *facts* about one's situation in yielding its judgments. This is why, when we are trying to decide how to act, we do not simply rest content with our present beliefs or evidence about the consequences of our actions, but seek out additional relevant information. Seeking this information is part of the process of figuring out what to do. Sometimes, through no fault of our own, we may fail to get the correct information, or may form justified, but false, beliefs. Because of this we may often fail to discover what we ought to do, and consequently we may fail to *do* what we ought to do. In failing to act according to the fundamental normative principle in such cases, we are not to be blamed. Nothing has gone wrong in the mental processes that produced our action. We would not want to call such actions 'irrational,' if we were taking irrational action to count

against the rationality of the agent in a way that was relevant to questions of moral responsibility, competence to give consent, freedom of will, mental health, and so on. Similarly, we may sometimes perform an action that is permitted according to the fundamental normative principle, given the *facts*; however, given our *beliefs*, it may be that our performance is obviously the result of some mental malfunction. In such cases we may want to call the action 'irrational,' if we are concerned with these same questions of moral responsibility, competence, and so on.

Since there may often be adequate (but unknown) reasons to perform actions that would be irrational in this 'mental functioning' sense, it should be clear that the 'mental functioning' sense of rationality is not the fundamental normative sense. It fails the second test. Nevertheless, it should be equally clear that the two senses of rationality are very closely related. But there is an interesting puzzle that one encounters in trying to specify exactly how they are related. It is very tempting to think that the 'mental functioning' sense of rationality is nothing but the fundamental sense, relativized to the *beliefs* of the agent, in place of the *facts* of the case.[4] But this cannot be correct. For it may be that an action would be rational, in the fundamental sense, if the world were as my beliefs represent it, and yet it may still be that my performance of the action would be irrational in the 'mental functioning' sense. This may happen because I conspicuously lack a belief that I should definitely have: the belief that my action will cause me a great deal of suffering, for example. I may refuse to believe this, although I have more than enough evidence to believe it, because it may be that the suffering will be caused by someone I love, and I may deceive myself into thinking that the person would never hurt me. In such a case my action would be rational, in the fundamental sense, if my beliefs accurately represented the world. But it is nevertheless irrational in the 'mental functioning' sense. The next obvious strategy would be to define rationality in the 'mental functioning' sense in the following way: it is simply the same as the fundamental sense, but relativized to the beliefs that the agent *should* have, given the available evidence.[5] But this strategy also fails, perhaps even more spectacularly. For it may be that an action would be rational, in the fundamental sense, if the world were as I *should believe* it to be, and yet it may still be that my performance of the action would

[4] See Brandt (1979), pp. 72–73; Gibbard (1990), pp. 18–19; Harman (1982), p. 127; Raz (1999a), p. 22.

[5] This definition follows a pattern used by Rawls in defining what he calls 'subjective rationality' in relation to what he calls 'objective rationality'. See Rawls (1971), p. 417.

be irrational in the 'mental functioning' sense. How could this happen? It may be that, though I *should* believe that a certain unpleasant action will benefit me greatly in the long term, I do not *actually* believe it. In such a case, the fact that the action will benefit me (and that I should believe this) does nothing to mitigate the irrationality of performing it, if it would be irrational to do so in the absence of the future benefits. So two initially plausible accounts of the relation between the two senses of rationality are completely inadequate. And it is obvious that one cannot simply relativize to the set of beliefs that one *does or should* have, for this will typically be a set of inconsistent beliefs. Nor can one relativize to the beliefs one *does and should* have, for if one believes that a certain action will be quite painful, and will benefit no one, then it would be irrational to perform the action, even if one *should not* have this belief.[6] This book provides an account of the relation between the 'mental functioning' and 'fundamental' senses of rationality in a way that not only avoids counterexamples, but also explains why the above relativizing definitions fail, and why they fail in the particular ways they do.

The 'mental functioning' and 'fundamental' senses of rationality are often distinguished by calling the former 'subjective rationality' and the latter 'objective rationality,' and this is the terminology I will use in this book.[7] But quite often philosophers do not distinguish the two senses at all. And sometimes the fundamental sense is the only sense of rationality that is officially recognized, so that the phenomena captured by the 'mental functioning' sense end up being described with phrases such as 'rational, relative to the beliefs of the agent.'[8] In earlier writing I sometimes borrowed a piece of terminology from Allan Gibbard, who uses the term "advisable" as a label for "[w]hat it makes sense to do objectively, in light of all the facts" – that is, for what I am calling 'objectively rational' action.[9] Gibbard's terminology has the advantage of minimizing the risk of thinking that the objective notion has much to do with mental functioning *directly*. However, I now prefer the terms 'subjective rationality' and 'objective rationality' because, despite the fact that a perfectly (subjectively) rational person might often perform objectively irrational actions, it is uncontroversial that there is a *very* close connection between subjective and objective rationality. Using the two terms 'rationality' and 'advisability' wrongly lends an air of plausibility to objections that depend

[6] See Cullity and Gaut (1997), p. 2.
[7] See Rawls (1971), p. 417; compare Gibbard (1990), p. 89.
[8] Williams (1981), p. 103. See also Sobel (2001).
[9] Gibbard (1990), p. 89.

on the false premise that a fully informed agent, performing a rational action, might nevertheless be performing an inadvisable one. Also, the 'subjective/objective' terminology allows me to use phrases such as 'an account of rationality' in order to indicate an account both of subjective and objective rationality, and of the relation between the two. In what follows, when I use the words 'rational' or 'irrational' without any qualification, they should be understood in the subjective sense, which fits more with the everyday understanding of these words.

Although I will not provide a full account of the relation between subjective and objective rationality until chapter 7, some limited claims about their relation are independently plausible, and will be very useful in a number of earlier arguments. First, if an agent knows all the facts relevant to his action, then if that action is objectively irrational – that is, if it is prohibited by the fundamental normative principle – it is also subjectively irrational. This connection will allow us to move from the objective irrationality of an action to its subjective irrationality (and therefore from its subjective rationality to its objective rationality) in all cases in which it is permissible to stipulate that the agent has all relevant beliefs. This claim is very similar to one of Gibbard's: "in the special case in which I know all that bears on my choice, what is rational for me to do is what is advisable for me to do."[10] My claim is slightly weaker, however, for it does not entail that we can always move from objective rationality – what Gibbard calls 'advisability' – to subjective rationality (or, therefore, from subjective irrationality to objective irrationality) even in the case in which the agent is fully informed. As chapters 7 and 8 will explain, this move can be illegitimate when the agent does not care about the considerations that make his action objectively rational, or only performs that action because of failures of instrumental rationality: cases in which an agent does 'the right thing for the wrong reasons.' These are cases in which the etiology of the action is what makes it subjectively irrational, and this gives rise to the possibility that if the same action had been done for other reasons it would have been subjectively rational – and therefore it also gives rise to the possibility that an action can be objectively rational despite being subjectively irrational, even in a fully informed agent. Acknowledging this possibility, we can make the following claim: if the agent is fully informed, then if his action would have been subjectively irrational *no matter what its etiology*, it is also objectively irrational. Since, as has already been noted, if

[10] Gibbard (1990), p. 19.

the action of a fully informed agent is objectively irrational then it is also subjectively irrational, Gibbard's claim is very close to correct.[11]

Another interesting feature of subjective rationality is the following. It seems that if one is simply unmoved by awareness of the prospect of some significant harm for oneself – say, that one's action will cause one a great deal of pain, or will risk some nontrivial injury – this does nothing to the normative force of the reason that one is aware of. This is not to deny that one can be perfectly rational in willingly suffering such harms, if there are sufficient countervailing reasons. But the fact that one needs significant countervailing reasons shows that a rational person cannot be very indifferent to such harms for himself. On the other hand, relative indifference to the harms that one's actions may cause other people is not nearly as universally regarded as irrational in the 'mental functioning' sense that is relevant to questions of moral responsibility and so on. Rather, when we speak of such indifference, we use words such as 'callous,' 'selfish,' or 'mean.' No one denies that it is rationally *permissible* to be motivated by other-regarding reasons. But it does not seem to be rationally *required* to the same extent that it is rationally *required* that one avoid harms for oneself. Of course there are views of rationality according to which one is in fact required to be as strongly motivated by altruistic as by self-interested reasons. This introductory chapter is not the place to combat such views. Rather, it is the place to mention that such accounts need to make us comfortable with some apparently counterintuitive judgments as to whether certain actions are subjectively rational or not – rational in the sense that is relevant to questions of moral responsibility and so on. That

[11] It may be worth mentioning at the outset that subjective rationality, as understood here, is not to be confused with Thomas Scanlon's technical and restricted sense of rationality, according to which actions are rational or irrational depending solely on whether or not they are in line with the agent's *normative judgment* that he or she ought to perform the action. Scanlon's sense is inadequate if we want a general notion that captures the wide range of failures of practical mental functioning that are relevant to questions of mental illness, competence to give consent, moral responsibility, and so on. One reason for its inadequacy as such a general notion is that, as I argue in chapter 9, our actions very rarely involve the normative judgments that are presupposed when one calls an action rational or irrational in Scanlon's sense. This is not to criticize Scanlon's choice of terminology. One is free to use whatever terminology one wants, as long as one is clear about what one means. It is only a reminder that Scanlon himself recognizes many other failures in practical mental functioning: insensitivity to certain reasons, compulsions, and phobias, even when they are accompanied by rationalizing normative judgments (or no normative judgments), etc. It is to this more general class that I am applying the term 'subjectively irrational.' Of course, acting against one's considered judgment as to how one ought to act is *one species* of the sort of irrationality with which this book is concerned, and it will be captured by the account offered in chapter 7.

is, in order to succeed in convincing us that it *is* irrational, in this sense, to be indifferent to the harms one's actions will cause other people, they will have to account for the fact that we generally wish to hold extremely immoral people fully responsible for their sadistic actions.

But even if it is a mistake to defend the normative equivalence of self-interested and altruistic reasons, one cannot simply deny that there are such things as altruistic reasons. That is, even if it is only callous, and not irrational, to be indifferent to the harms one causes others, one should not therefore deny that it can be perfectly rational to make great sacrifices – even the ultimate sacrifice – for others. It would be a poor theory of rationality that insisted that it was irrational to sacrifice one's life to save a group of strangers, or that held that one's 'real' reason in such a case was essentially self-interested.[12] If an agent is strongly motivated to save a group of other people, and acts accordingly at the cost of his own life, this may be perfectly selfless, and perfectly rational. These two facts – that it is rationally permissible to be selfish, but also rationally permissible to make selfless sacrifices for others – seem to suggest that whether or not one has an altruistic reason depends upon whether or not one has a corresponding altruistic desire. And indeed there have been philosophers who explicitly claim that while one's objective *interests* or *needs* provide desire-independent reasons, there is another class of reasons, which includes altruistic reasons, that stem from one's *desires* or *values*.[13] The plausibility of such views derives entirely from their ability to capture some otherwise elusive phenomena. Moreover, such accounts will need some way of limiting the content of one's reason-giving desires or values, so that they do not end up claiming that one has a reason to drink paint simply in virtue of a desire to do so, or that one has a reason to exterminate some offending race of human beings because of one's racist values. It is one goal of the present book to provide an *explanation* for the differential impact of desire on the relevance of reasons to the subjective rationality of action. Moreover, this explanation will limit the importance of desires in such a way that one's desire to drink paint or to hurt someone else never rationally justifies one's action, while one's desire to help someone else can provide such justification.

Because the following arguments will be so much at odds with the Kantian view that moral requirements are also rational requirements, it may seem as though they must be concerned with a more stringent notion of

[12] It would be as poor as a theory that insisted that it would be irrational *not* to make such a sacrifice, because one would, by failing to act, cost more lives than one saved.

[13] See Copp (1995), pp. 172–85 and Foot (1978b), pp. 148–56.

subjective rationality: perhaps something closer to the colloquial notion of insanity. But this would be a misperception. On the view that will be put forward here, smoking generally counts as mildly irrational, as does postponing a trip to the dentist. The difference with a Kantian view is not a conceptual one, or a question of the severity of the charge of irrationality. Rather, it is a substantive disagreement about what really does count as a defect (large or small) in practical reasoning. Of course the notion of irrationality in play here is *related* to the notion of insanity. But it is unlikely that there is *any* plausible notion of irrationality, either practical or theoretical, such that an agent might do countless extremely irrational actions, or hold countless extremely irrational beliefs, and still avoid the charge of insanity.

RATIONALITY AND MORAL THEORIES

Very roughly speaking, contractualist moral theories hold that morality is the system of rules that people would agree to, under certain conditions. However, if 'people' is understood here as 'actual people' then it is unlikely that contractualism will yield very extensive or determinate results. After all, some people hate to agree with other people, other people are too stupid to agree to anything, and still others are simply self-destructive lunatics. So contractualist theories will need to restrict the scope of 'people' in some way. Intuitively, the kind of people we would like to exclude are *irrational* people. That is, contractualism is plausible as a moral theory if it claims that morality is the system of rules that *rational* people would agree to, under certain conditions. It should be clear, therefore, that the notion of rationality is likely to play a crucial role in such a moral theory. It should also be clear that being rational cannot simply be equated with being moral, for then contractualism would be trivial.

Now, one interesting question for contractualist moral theories is whether a person, if rational, would or could *actually act* on the set of rules that they *would have agreed to* under the relevant counterfactual conditions. We can grant the contractualist the plausibility of the claim that the way a rational person will act is *quite similar* to the way that such a person would advocate that others act, and that it is also *quite similar* to the way that such a person would agree to act, given that others also agreed to act in that way. That is, we can agree that there is liable to be very significant overlap in these differently specified classes of action. But suppose that there is *any* mismatch between the way a rational person would

actually act, and the rules that a rational person would agree to act on, *on the hypothesis* that others similarly situated would also agree. If there is *any* mismatch then there is the danger that it will sometimes be irrational to be moral, according to a contractualist moral theory. That would be a very bad consequence for such a theory, if rationality is taken in the objective sense. For it would mean that there could not, even in principle, be an adequate reason to do what was, in those situations, morally required. And it would not be much better if rationality were taken in the subjective, 'mental functioning' sense. For one thing, this would mean that someone would have to be a little 'wrong in the head' to act morally in those mismatched situations. And in any case, the best explanation for the subjective irrationality of such action is likely to be its objective irrationality.

Is this possibility of mismatch a real one, and if real, does it mean that contractualist theories are doomed? On many popular conceptions of rationality the answer seems to be 'yes.' If we understand rationality in maximizing terms then it is extremely plausible that contractualist moral theories will always yield at least a small class of actions that are morally required but rationally prohibited.[14] Maximizing views of rationality say that, in a given choice situation, there is one class of rationally permitted actions: those that maximize some measure. On some views this is a measure of preference satisfaction, on others a measure of pleasure, and on still others it is a weighted sum of a number of distinct goods. It is extremely unlikely that the class of actions that actually maximize the relevant measure in actual circumstances will always include the class of actions that are required by a set of rules that have the following feature: in the relevant counterfactual circumstances it would maximize the measure to advocate or to agree to those rules.

A similar sort of trouble will afflict consequentialist moral theories. Such theories, very roughly, claim that morality is a matter of acting in such a way as to bring about the best consequences.[15] This may be a matter of trying

[14] I focus on subjective rather than objective rationality in the following discussion because both morality and subjective rationality are plausibly regarded as somehow relative to the agent's epistemic situation. Thus no difference in this epistemic relativization is available to remove the sting from the possibility of mismatch between morality and subjective rationality, as there is in the case of a mismatch between morality and objective rationality. Moreover, maximizing views of subjective rationality typically go hand-in-hand with maximizing views of objective rationality.

[15] These may be the best *actual* consequences, or the best *foreseeable* consequences. If the former, then the theory will have to separate moral wrongness from blameworthiness. But the argument offered here will still apply, with 'objective rationality' substituted for 'subjective rationality.'

to bring about the best consequences each time one acts, or of acting on motives that tend to produce the best consequences, or of acting according to a system of rules that is best with regard to consequences. Since 'best consequences' means 'best for everyone' in this context, consequentialist theories inevitably run into a problem if they also advocate a maximizing theory of rationality. For unless the theory of rationality simply says that rational action is action that brings about the best consequences in exactly the same way as the moral theory says that moral action does, then there is always the possibility that the action that maximizes the measure relevant to morality will not be the action that maximizes the measure relevant to rationality: some morally required action will be irrational. In fact, Mill runs into a very similar problem in *Utilitarianism*. According to Mill, one ought always to act so as to bring about the greatest amount of overall happiness for those whom one's action will affect.[16] But Mill also holds that the way one actually *will* act is determined by how much happiness one believes one will get for *oneself*. Certainly these two sorts of actions will not always coincide. Because Mill sees this, he is explicit in his advocacy of an education that will bring people's ideas of personal happiness more in line with their ideas of overall happiness. But until someone is successfully educated in this way, it may be impossible – if Mill is right – for that person to act as they ought. This problem is not exactly that it is *irrational* to act as one ought. This is because Mill is a *psychological* egoist, and not a *rational* egoist. That is, he says we are actually psychologically set up to maximize our own perceived happiness, not that it is rationally required that we do so. But the problem is identical in form. In a nutshell it is the following: if one has a maximizing view of rationality, and a maximizing consequentialist view of morality, then unless the two views are really the same view, one will sometimes have to choose whether to be moral or rational.

One avenue of escape from the above problems is to embrace the idea that rationality and morality really do amount to the same thing. This claim is characteristic of Kantian moral theories. These theories of morality begin by offering some morally neutral characterization of what it is to

[16] I use the phrase 'ought to act' advisedly here, instead of 'is morally required to act,' for Mill's maximizing view applies directly to 'ought,' and only indirectly yields moral requirements as actions that *ought to be punished*. See Mill (1979), ch. 5. This actually results in a much more plausible moral theory than Mill is generally credited with, and one that is not maximizing at all. For this reason the important conflict discussed above must be cast in terms of Mill's maximizing nonmoral 'ought.'

be a rational being. For example, rational beings might be beings who act on universal laws, or who have the capacity to evaluate their desires before they act on them. These theories then go on to argue that any such being is somehow implicitly committed to acting morally. As in the case of contractualist theories, it is important for Kantian theories that the initial characterization of rationality be given in nonmoral terms. Otherwise the view becomes trivial or circular. Moreover, it is important for these views that the notion of rationality they make use of be one that is somehow inescapable. That is, when a Kantian offers her account of rationality, it would be bad for her if many people could sincerely say 'Oh, well, if *that's* what you mean by "rational," I really don't care much whether I generally act rationally.'[17] The notion of rationality should be one such that virtually no one would ever want to act in an irrational way: or at least, it should be one such that no one could think that she herself had an adequate reason for acting irrationally.[18] The problem with this avenue of escape is that identifying rational action with moral action also identifies immoral action with irrational action. But typically if a person habitually acts in significantly irrational ways, we think that it is appropriate to call the person himself irrational, in a sense that is supposed to be relevant to questions of moral responsibility. Thus, on Kantian views, the more egregiously immoral one is, the less morally responsible one would seem to be. In fact, it is also a contingent fact about many Kantian moral theories that they identify irrational action with action that is not completely free or autonomous. This also suggests that grossly immoral action, which is also *eo ipso* significantly irrational according to such views, is not free or autonomous. Why then do we feel justified in holding the perpetrators of such action morally responsible? Of course I am not the first to point out this inherent difficulty with Kantian moral theories. And it is equally true that Kantians have many ways of attempting to meet it. But it is a significant problem nonetheless. Kantian moral theories clearly make the relation between rationality and moral responsibility hard to understand.

[17] One inadequate reply to this dismissal is 'Whether you care or not, you are a rational being, and so you are inescapably set up to act this way.' For this response makes it impossible to understand how irrational (or, hence, immoral) action is even a possibility.

[18] One can think that there is an adequate reason for someone *else* to perform an action that would be irrational for that person. This can happen when one is aware of relevant facts of which the other person is ignorant. Obviously this cannot happen when one is considering one's own actions.

One way to avoid all the problems for contractualist and consequentialist moral theories, without giving up the intelligibility of the connection between rationality and moral responsibility, is the following. One could offer a theory of rationality that has sufficient latitude (a) to classify all morally required action as rationally *permissible*, but also (b) to classify much immoral action as rationally permissible as well. The first of these features would allow one to avoid the problems that contractualist and consequentialist moral theories encounter when they are offered in tandem with maximizing views of rationality. The second would allow one to escape the problems that Kantian moral theories encounter when dealing with questions of moral responsibility. How might one build a theory of rationality that incorporates this latitude? It will require a certain amount of coordination between one's moral theory and one's theory of rationality. In particular, it will require that when one is morally required to perform an action that goes against one's personal interests, there will always be considerations that one can cite that are sufficient to *rationally justify* acting against those interests. And this will require that there be considerations one can cite to show that the action is not objectively irrational – for if an action is objectively irrational from the point of view of the agent on whom morality is making its demand, then it is subjectively irrational. But for this strategy to provide a solution, it is necessary that we understand these justifying considerations as exactly that: rationally *justifying* considerations. That is, the theory of rationality should hold that those considerations are necessarily sufficient to make the action rationally *permissible* – not that they are necessarily sufficient to make the action rationally *required*. This is the sense of 'justify' in which one might say that considerations of self-defense can *morally justify* a person in killing someone who is attacking her. The fact that one must kill the person in order to preserve one's life is a fact that makes it morally *permissible* to do so. But it by no means makes it morally *required*. Rational justification, like moral justification, is a matter of changing the status of an action from prohibited to permissible.[19] In the case of moral justification, the change is from morally prohibited to morally permissible. In the case of rational justification, the change is from rationally prohibited to rationally permissible. A large part of this book, including chapters 2 through 5, is devoted to clarifying this concept of justification, and distinguishing it from the logically related but

[19] The very same consideration that rationally (or morally) justifies an action may also make the action rationally (or morally) required. But then the consideration is doing something *more* than merely justifying the action.

distinct concept of requirement.[20] The distinction between the justifying and requiring roles of practical reasons is, in my view, one of the most important features of such reasons, and also – once one begins to look at actual cases in a systematic way – one of the most obvious. And yet it has been completely overlooked by virtually every contemporary ethical theorist, yielding the sorts of troubles detailed above.

If one's accounts of rationality and morality are such that, when an action is morally required, it will always be possible to cite considerations that rationally *justify* it, then one has escaped all the problems described above for contractualism, consequentialism, and Kantianism. Although no particular account of morality will be offered in this book, the account of rationality that is offered will make it easy to construct plausible moral theories that necessarily include such rationally justifying features in all morally required action. For the account of rationality will include the claim that altruistic considerations provide reasons that can rationally *justify* personal sacrifices: that is, altruistic considerations can make personal sacrifices rationally permissible. This will mean that morally required action will be rationally permissible just so long as those sacrifices are not blindly offered at the altar of a dogmatic moral view that requires sacrifice even when such sacrifice produces no benefits – or only insignificant ones – for others.

SUMMARY OF ADEQUACY CONDITIONS

It may be worthwhile to summarize this chapter's suggestions regarding adequacy conditions on accounts of practical rationality. One should provide a fundamental principle that gives what I have called the objective rationality of an action. This principle should pass the following two tests.

T1 The question 'Why should I always act *that* way?' should not make clear sense.

T2 It should not be possible to offer adequate reasons for acting against the principle.

[20] There are, of course, other related senses of 'justification.' For example, one of the implications of the view offered in this book is that actions that fall in a certain class do not stand in need of rational justification. If one performs such an action, and is challenged to provide a justification, it is possible to do so simply by showing that it belongs to this class. Such a justification clearly neither changes the status of the action, nor cites a consideration that does so. But it is plausible that this sense of 'justification' is derivative.

Moreover:

T3 The account should distinguish between rationality in the fundamental normative sense – the sense relevant to tests T1 and T2 – from rationality in the 'mental functioning' sense that is more directly relevant to questions of moral responsibility, competence to give consent, and so on.

With regard to rationality in the 'mental functioning' sense:

T4 The account should explain the failures of attempts to define rationality in this sense as being essentially the same as rationality in the fundamental sense, but relativized to some set of beliefs of the agent (including beliefs that the agent *should* have).

T5 The account should explain the differential relevance of desires with regard to self-interested and altruistic reasons.

T6 The account should yield the verdict of 'irrational' for the kinds of actions that we really do take as counting against moral responsibility, competence to give consent, etc. And it should not yield the verdict of 'irrational' for the kinds of actions that we do not take as indicating mental malfunctioning of this sort. In particular, it should not automatically yield a verdict of 'irrational' for all immoral action.

Further, with regard to morality:

T7 The account should be consistent with the claim that, for agents who know all the relevant facts, no morally required action is ever irrational in the fundamental sense.

T8 The account should be consistent with the claim that no morally required action is necessarily irrational in the 'mental functioning' sense.

This book will offer an account of practical rationality that meets all of these conditions. Much of the work in providing such an account will depend on the distinction between the justifying and requiring roles of normative practical reasons, which is explained and defended in chapters 2 through 5. Because this distinction, though implicit in much commonsense and philosophical reasoning, is not explicitly recognized by philosophers who write about rationality and reasons, chapter 6 provides a way for many contemporary philosophers to fit this distinction into their own accounts

of practical reasons. Chapter 7 is perhaps the heart of the book, and it is where the actual account of objective rationality is offered, along with an explanation of its relation to subjective rationality. Chapters 8 and 9 draw out some implications of the view, with regard to some of the issues that have been at the center of ethical theory for the past thirty years. One of these is the so-called 'internalism/externalism debate,' which focuses on the relation between the motives of agents, and the reasons those agents have. The other concerns the role of normative judgments in the etiology of intentional human action.

It may seem strange that the explicit account of objective and subjective rationality comes so late in the book, especially since one of the points of the book is to suggest that a failure to distinguish sharply between these two concepts is the source of a great deal of confusion. But one of the reasons why philosophers have been able to conflate objective and subjective rationality is that, in most cases, any claim about the subjective rationality of an action will imply the same claim about its objective rationality, and vice versa. As was explained above, this is true whenever the context is such that there is no harm in stipulating that the agent is aware of all the relevant facts, and when it is not the specific etiology of the action that makes it subjectively irrational. Indeed, because of this connection, and because we have stronger intuitions about subjective rationality than about objective rationality, I will couch most of the early arguments in terms of the subjective notion, and leave it to the reader to make the inferences regarding the objective one. That such a strategy is unproblematic is of course not a reason, in itself, for deferring an explanation of the relation between objective and subjective rationality. But an easy and relatively complete explanation of this relation is impossible until the distinction between the justifying and requiring roles of practical reasons can be taken for granted, and that *is* a reason for deferring discussion of the relation. Once the justifying/requiring distinction is understood and appreciated, the full account of rationality should be very easy to understand. It is my hope that it will also seem compelling.

2

Practical rationality, morality, and purely justificatory reasons

Because the normative notions of practical and theoretical rationality seem, due to their respective names, to be species of one genus, it is often assumed that there should be a very strong parallel between the two notions.[1] In particular, it is often assumed that for practical rationality, the business of normative reasons is to count in favor of (or against) doing something, and that for theoretical rationality, the business of normative reasons is to count in favor of (or against) believing something.[2] And in both cases it is assumed that reasons do this by providing justification which either *is* requirement, or which would tend, if the reasons became stronger or more numerous, to mount in strength and become requirement.[3] A closely

[1] Because theoretical rationality is relative to the epistemic situation of the agent, the notion under discussion here should be taken to be subjective rather than objective practical rationality. Moreover, morality also exhibits the kind of relativity to the agent's epistemic situation that is missing in the case of objective practical rationality. Making this explicit removes one potential source of confusion when we are comparing the analogy between practical rationality and theoretical rationality with the analogy between practical rationality and morality.

[2] In the remainder of this book the qualification 'normative' will generally be dropped when talking about reasons. But it is always to be understood that the reasons being discussed are normative, and not (necessarily) explanatory. Explanatory reasons are causal or psychological entities that *explain* my actions or my beliefs. But such reasons may well fail to *justify* or *require* them in a normative sense: they may fail to be relevant to 'ought' claims about those actions or beliefs.

[3] This view is so widespread that many theorists do not seem to recognize that there is a position opposed to it. As a result, it is not often clearly stated. Nevertheless, for relatively clear endorsements, see Darwall (1983), pp. 19, 54; Korsgaard (1996a), pp. 225–26; Audi (1997), pp. 146–47; Scanlon (1998), pp. 18–23; Copp (1995), p. 42; Velleman (1996), pp. 705ff; Edgley (1965), pp. 182–88. Foley (1991), pp. 365–66 also favors a unified treatment of rationality, and expresses something very much like the above view, according to which "reasonability is a matter of the relative strength of your reasons" and "[t]he rational is that which is sufficiently reasonable." See also Foley (1992), p. 111. A slightly different analogy between practical and theoretical reason was, according to Darwall (1999), p. 9, popular with rationalists like Balguy and Clarke. See Balguy (1978), p. 45; Clarke (1978), p. 614. Their analogy, translated from faculty-language to norm-language, is equally undermined by the analogy presented in this chapter.

related position holds that if a belief is held for no reason, or if an action is done for no reason, then the respective belief or action is unjustified and irrational.[4] As more theoretical reasons are found for the belief, or as more practical reasons are found for the action, or as existing reasons become stronger, the belief or the action becomes increasingly *justified*. If the justification becomes strong enough, then the belief or the action is *required*.[5]

The above position, that sufficient justifying reasons will always yield requirement, is consistent with two interpretations. The first interpretation, (a), allows some actions and beliefs to be justified but not required. The second interpretation, (b), is one on which 'increasingly justified' means only 'closer to justified,' and on which any action or belief that is *actually* justified is also required. The essential point, shared by both (a) and (b), is that any reason that can *justify* can also *require*, if it is instantiated strongly enough or in sufficient numbers, or if countervailing reasons are weakened or removed. For example, those who adopt interpretation (a) may hold that though I am not *required* to believe a rumor of war heard from one source, I might be *justified* in believing it. But if I begin to hear the same reports from a great many sources of equal reliability, and if I have no reason to doubt that there is a war going on, eventually I will be required to believe it. Those who favor interpretation (b) may hold that if I have no reason to doubt my source, then an unopposed reason to believe in the war generates, on its own, a requirement to believe it. But even those who favor interpretation (b) are likely to say the following. If I *do* have reasons to doubt the existence of a war, then the rumor of war I hear from one lone source nevertheless provides *some* justification for believing that there is a war; it is just an *insufficient* justification.[6] But if I hear more reports from more and more sources, the justification will become great enough that I no longer have sufficient justification to *doubt* the existence of the war. At that point I am fully justified in believing – and required to believe – that the war is going on. So both interpretations endorse the view that sufficient justifying reasons will eventually yield requirement. This is the essential point that this book challenges *in the practical realm*. That is, no

[4] For an expression of this view in the practical realm, see Foot (1978a), p. 173.

[5] In all these views, and in what follows, 'required' should be taken to mean 'required, on pain of acting or believing irrationally.' The exception is for moral requirements, where 'required' should be taken to mean 'required, on pain of acting immorally.'

[6] If someone wants to maintain that such reasons provide *no* justification until they actually require, then that person has simply *identified* justification with requirement. The arguments of this chapter tell equally strongly against this extreme position.

matter how strong the *justification* for some *action* becomes, it never follows, simply in virtue of the strength of such a *justification*, that one is required to do the action. Therefore, despite differences between interpretations (a) and (b) that might be relevant in other contexts, both interpretations are grouped together here as species of one view.[7] This book neither challenges, nor endorses, this view of justification and requirement in the *theoretical realm*.

When one holds that justification and requirement are linked as the above views link them, one needs to address the question of when it is that justification becomes requirement. One might try to answer this question from the agent's own point of view, and hold that the point at which one actually commits oneself to a belief or action is also the point at which one takes one's justification to have become a theoretical or practical requirement. Let us call this 'the commitment view.' If one holds the commitment view, then one will also hold that if one takes oneself to have equally good (or the very same) reasons for beliefs p and q, then one must either believe both p and q, or neither p nor q.[8] Alternately, it is possible to hold that *any* nontrivial balance of justification generates a requirement. Let us call this 'the limiting view.' On the limiting view, as long as it is clear that there is *any* nontrivial reason for an action, and no reasons against it, then the action is both justified and required. Philosophers who hold a certain widely accepted version of practical reasons internalism – that any rational agent, simply in virtue of being rational, will always be motivated to some degree by any practical reason relevant to her choice of action – are committed to something practically indistinguishable from the limiting view. This is because this version of reasons internalism entails that all unopposed practical reasons, at least of a certain bare minimum strength, will produce action in a *rational* agent. In other, more explicitly normative words, anyone who fails to act on such an unopposed reason is, to some degree, irrational. And this means that all reasons, at least of a certain minimum strength, provide prima facie rational

[7] One interesting difference might be that the former view generates a class of possible actions which are justified but not required: those possible actions for which the reasons for and against are relatively close to balancing. But this class will in general not be very extensive. On the other hand, the view of practical rationality advocated in this book allows a very wide and interesting range of actions that are rationally justified but not required. The range includes all morally required action, and also all self-interested action.

[8] See Darwall (1983), p. 110, for an explicit endorsement of this view with regard to theoretical reasons, and the assumption that the same holds true of practical reasons.

requirements.[9] This last claim is also a logical consequence of the limiting view. The arguments in this chapter suggest that there are some practical reasons that do not provide prima facie rational requirements. Thus, the arguments are directed not only against the common acceptance of an overly strong parallel between practical and theoretical rationality, but they also oppose an even more ubiquitous acceptance of practical reasons internalism.[10] This opposition is articulated more explicitly in chapter 3, in the course of responding to an objection that takes internalism for granted. The general debate between internalists and externalists will be taken up explicitly in chapter 8.

The argument of this chapter is primarily aimed at philosophers who both take theoretical reasons to function roughly as the various above views hold, and who also take theoretical rationality to provide a good model, in the respects mentioned, for practical rationality. It may well be true that reasons for *belief* function in one of the uniform ways endorsed by the various views described above. That is, it may well be true that any reason that justifies some belief would also require it, if it were stronger, or if there were more reasons of the same sort, or in the absence of countervailing reasons. Or it may be true (as the limiting view holds) that *any* reasons that actually justify some belief also require it. Moreover, it seems perfectly plausible that *some* practical reasons both prima facie justify and prima facie require. This chapter does not attempt to call any of these claims into question. But now consider the following claims about practical rationality.

Suppose that one could save forty children from severe malnutrition by smuggling them food and medicine at high risk of injury and death to oneself. In such a case, there is a very strong reason *against* smuggling the supplies: that one risks injury and death. It would be seriously irrational to risk injury and death in the absence of countervailing reasons.[11] But

[9] Here and elsewhere I use 'prima facie requirement' to indicate not an *apparent* requirement, but a requirement that persists until countervailing reasons remove it.

[10] For an unargued acceptance of reasons internalism of this sort, see Cohon (1986), pp. 545–56, esp. p. 556; Smith (1994), esp. pp. 60–2 and 151–77. Smith does indeed provide an argument for *moral* reasons internalism, but it is based upon an undefended assumption of the internalist requirement on reasons generally. The same is true in Nagel (1970). In fact Nagel's arguments only support a weaker view in any case, although he assumes internalism about reasons at pp. 66–7 and elsewhere. See also Velleman (1996), pp. 700–4, where Velleman asserts that it is "trivial" that "rationality is a disposition to be influenced by reasons," and even takes the externalist to agree. Darwall (1983), p. 52 does the same.

[11] This is true regardless of the etiology of the action. Therefore, since it would remain subjectively irrational even if we stipulated that the agent was fully informed, we can conclude that such an action is also objectively irrational. See p. 8.

in the example there is of course also a very strong reason *in favor of* smuggling the food and medicine: that doing so will save many children from serious illness. In the example, the reason in favor of smuggling the food clearly seems at least as strong as the reason against it. This is why one is rationally justified in smuggling the food.[12] But if all practical reasons were comparable along one axis of strength, then it should be irrational *not* to act on the reasons for smuggling the food, unless there are quite strong reasons against it. But would it be seriously irrational, or irrational at all, to fail to act so as to save forty children from serious malnutrition, if there were no reasons, or only weak reasons, against saving them? For example, do we regard ourselves as acting irrationally if, instead of preventing malnutrition in forty children by (relatively) painlessly donating a hundred dollars to Oxfam, we spend the money on a good bottle of wine, or on nothing at all? If the answer is 'No,' then we have a case in which there are very strong reasons in favor of donating the money (strong enough to justify risking injury and death), and only weak reasons, or no reasons, against donating it, and yet our action is not irrational. Callous or selfish it may be, but it is not irrational. The interests of the forty children provide a very strong justification for actions that would otherwise be irrational, but their interests do not provide any rational requirement to act.

It is the point of this chapter to make it plausible that the above description is correct: that there are some reasons, relevant to the rationality of action, which can be very strong rational *justifiers*, but which do not rationally *require* at all. That is, no matter how strong or numerous these reasons become, it is never the case that we are *irrational* if we fail to act on them. Let us call such reasons 'purely justificatory reasons.' One large class of these reasons stems from the interests of others. It is not that purely justificatory reasons have an upper limit on their possible strength, and cannot ever become strong enough to require. For in one sense, to be explained, the power of purely justificatory reasons may increase without any practical limit. Rather, the point is that justification, as a function of the normative reasons relevant to rationality, is a function logically distinct from the function of requirement. Justification is a matter of making it rationally permissible to do something that, without justification, would be irrational. The arguments below will suggest that some reasons can

[12] Again, since this remains true even when we stipulate that the agent is fully informed, this means that such an action is objectively rational. In what follows I will generally omit to point out implications of the sort mentioned here and in the previous note, unless there is some special reason to do so.

justify without in any way tending to make it required to do the action that they justify.

Some may take issue with this definition of justification – specifically with the idea that justification is to be understood as relative to something that would otherwise be irrational or immoral. But in fact another way of understanding the central point of this chapter is to see it as a defense of this conception of justification, as against one that conflates the process of justifying with the process of requiring. That is, the claim is that the notion of justification gets its sense from contexts in which things *stand in need* of justification. In the theoretical realm it is plausible to hold that almost *all* beliefs would be irrational to hold (or at least that one ought not hold them) in the absence of *some* reason to believe them. Thus, it may be true that *every* belief *stands in need* of justification. This, together with the fact that reasons for belief *tend* to require that for which they are reasons, makes it very hard to distinguish justifying reasons from requiring ones. And this makes it easier to conflate justification, as a process, from requirement. In fact, it may be useful to distinguish justification from requirement in the theoretical realm also, although I will not explore such an idea in this book.

One might be tempted to think that the altruistic reason in the above example – that one can save forty children from malnutrition – is a *moral* reason, and not a 'generic' practical reason. Two things can be said about this. First, it is a mistake to equate altruistic reasons with moral reasons. Much of the most grossly immoral action is done completely selflessly, for the sake of others: children, spouses, professional colleagues. But the altruistic reasons in these cases generally provide as little moral justification as would a corresponding self-interested reason, had the agent performed the same sort of immoral action for her own sake. Second, a consideration can be both a moral *and* a 'generic' practical reason. Very roughly, if a consideration contributes systematically to the moral status of an action, then it is a moral reason, and if it contributes systematically to the rational status of an action, then it is a 'generic' practical reason. Altruistic reasons are therefore 'generic' practical reasons, since they systematically rationally justify actions that would otherwise be irrational.[13] The notion of 'systematic contribution' is discussed in more detail in chapter 4.

[13] It may be worth mentioning here that purely justificatory 'generic' practical reasons that also happen to be moral reasons need not be purely justificatory moral reasons. This is why some morally required behavior might not be rationally required.

It is important to keep in mind the distinction between altruistic and moral reasons, and the possibility that one and the same consideration might be both a moral reason and a 'generic' practical reason, because the 'generic' practical reasons that this chapter argues are purely justificatory also happen to be altruistic ones. If one conflates moral and altruistic reasons, and fails to see that moral reasons can also be 'generic' practical reasons, then one may wrongly conclude that this chapter succeeds only in showing that moral reasons can be purely justificatory. Of course egoists and desire-satisfaction theorists might deny that altruistic practical reasons exist, or that they *systematically* contribute to the rational status of action. This is not the place to argue against either of these views, but one example may at least suggest why they are both inadequate: if some habitually selfish and mean person suddenly decides to turn over a new leaf, and begins helping other people, no one would call such behavior irrational. The idea that we have a fixed set of desires, and that it makes sense to try to maximize their satisfaction, is as much of a fiction as the idea that we are only (rationally) motivated by our own interests. The rational options open to an agent are those that would be judged rational *if* they were chosen. And, for all we can actually know about any particular agent, altruistic sacrifice is always something that could be chosen.

It may seem that there are more uncontroversial cases of practical reasons that justify without requiring, even when they are strengthened. For example, suppose that I am in a stuffy room, and know that I could get fresh air either by opening one of the three windows, turning on the central air-conditioning, or opening the door to the porch. Its being stuffy seems to be a reason that would justify performing any of these actions. But just in virtue of their being so many justified options, none of them could be required. And of course, even as the reason stemming from the stuffiness becomes quite strong, none of the available actions become required. So this seems to be an example of a reason that could be indefinitely strengthened without yielding a requirement to do any of those particular actions that it justifies. But these sorts of 'multiple option' examples do not illustrate the same point as the examples given above. The belief that they do illustrate the same point rests on a conflation of basic and derivative reasons, and on using a wrong level of description. The practical reason supplied by the fact that it is stuffy, which justifies my either opening a window or the door, is dependent upon the fact that stuffiness is unpleasant and that fresh air reduces stuffiness. Barring other circumstances, there would be no reason to open the window or door, if it were not the case

that I find stuffiness unpleasant and that fresh air reduces stuffiness. But nothing similar can be said about the reason 'because it will make me feel less uncomfortable.' It is basic.[14] 'Because it will make me feel less uncomfortable' is also the sort of reason that can require action, given a description of the required action that is at the same level of generality as the basic reason. If one is feeling extremely uncomfortable, and can do something about it rather easily then one is required to *do something to make oneself less uncomfortable*. If, as in the stuffy room example, there are a variety of ways of doing it, one is of course not required to do *each* of them. But one is nevertheless required to *do something to make oneself less uncomfortable*. If one has to do something that is itself unpleasant, or risky, then this reason also *justifies* doing that thing (at least, if the risk is not too great, etc.). Unless one realizes that talk of practical requirement demands appropriate levels of description, it will never be true that *any* action is required, since there will *always* be irrelevant features of the action one is contemplating. That is, to take a standard moral example, I am required to *save the baby*, but not to *save the baby by diving into the water*, as opposed to *saving the baby by jumping into the water*.

Because of the above considerations, it is important to take as motivating examples those of the sort that are offered above, in which *saving other people from death and pain* justifies but does not require one to *risk pain, injury, and death*, and also justifies but does not require one to *forego a small pleasure*. For these cases of justification without requirement do not depend upon a description being at too high a level of detail. They are not 'multiple option' examples. Of course Kantians may simply reject these types of motivating examples. It is a standard Kantian claim that immoral action is also irrational. And some Kantians might also claim that if one chose to spend a hundred dollars on a bottle of good wine instead of donating that amount to Oxfam, then one would also be violating the categorical imperative and (therefore) acting irrationally. Admittedly, there are very strong reasons to forego the wine and send the money off to Afghanistan or Honduras. But despite agreement with this, virtually no nonphilosopher would claim that buying the wine showed the slightest problem in practical mental functioning, even in a fully informed agent. Again, it is not just that such selfish action is not *insane*. Smoking is not insane either, and yet many ordinary people (even smokers) admit that it is 'stupid' to smoke, and that

[14] One need not be a foundationalist to make this kind of point. Basicness may be context-relative. See Heath (1997).

the reasons they have for smoking are not really sufficient to justify the health risks. On the other hand, those who buy wine, cars, new clothes, CDs, and so on, do not need to provide a rational justification for doing so. If the existence of people in extreme need is pointed out to them, they may feel more or less pressure to provide a *moral* justification, but this is a different matter. Even Kant does not *start* from the claim that immoral action is irrational. Rather, his claim is the result of ingenious arguments that start with a nonmoral conception of the nature of rational agency, and *lead* to the claim that the actions we recognize as immoral would not be performed by someone who was completely rational. This is not the place to show exactly where these ingenious arguments go wrong. But it is worth noting that when one is arguing for some practical conclusion (such as the claim that it is irrational to lie, cheat, steal, etc.), one must use normative words like 'reason,' 'ought,' 'rational,' and their cognates, with the meanings one learned for them when one learned one's native language. There are no *more basic* normative notions or principles one can appeal to.[15] Thus, if it is clear that, according to the actual use of these words, some reasons (say, altruistic ones) are only (rationally) justificatory, and are not (rationally) requiring, then it is unclear how any *argument* could ever show anything different. It is part of the point of this book to make it easier to see that the normal usage of normative language does indeed include such things as purely justificatory reasons, and that altruistic reasons are among them.

Thus, despite Kantian objections, there are strong reasons to accept the claims about the examples above: that it is not irrational to fail to donate money, even though it need not be irrational to make great personal sacrifices to achieve the same ends. Nevertheless, in the face of these examples and claims, one might point out that such a difference in the structures of theoretical and practical rationality must be accounted for. In the absence of such an account, it might be urged, the parallel should lead us to conclude that practical reasons stemming from the interest of others really *do* require action, and that it is only human selfishness that blinds us to this theoretically transparent fact.[16] Or, like the egoist and desire-satisfaction

[15] It is true that I am arguing for a fundamental normative notion – objective rationality – which is not represented in an obvious way in the language. Why cannot the Kantian simply invent some similar notion? The answer is that any such notion must pass the two tests given in chapter 1, and these tests make use of familiar normative words that must be used in standard ways.

[16] Because the idea of purely justificatory reasons is not discussed, this type of objection is not found explicitly in the literature. Though it might not reflect their considered views,

theorists mentioned above, one might claim that such purported 'altruistic practical reasons' are not really reasons at all, and can neither justify *nor* require, if they do not engage the interests of the agent in some way. If either of these claims is true, then all practical reasons may well have some power to require. The following arguments attempt to motivate a different parallel for practical rationality: a parallel with morality. The argument draws attention to the plausibility and common acceptance of moral reasons that violate a moral version of practical reasons internalism: moral reasons that need not find a corresponding motivation even in a morally good person. These are purely justificatory moral reasons. The analogy is then intended to support the idea there are reasons that are purely justificatory with regard, not to morality, but to practical rationality. The initial case for this claim emerged from the above discussion of some examples of rational action. This discussion pointed out that it is generally considered rationally permissible to risk a great deal to save forty children from malnutrition, but it is also generally considered rationally permissible not to make even a small sacrifice to achieve the same end. If the following point-by-point parallel between morality and this view of practical rationality is at all compelling, then objections to these claims that are based on a putative parallel between theoretical and practical rationality should lose much of their force.

THE PARALLEL WITH MORALITY

The parallel with morality includes the following features: first, a notion of a prima facie immoral action; second, a notion of a moral justification for such an action; third, the fact that the considerations that morally justify are not identical with the considerations that make an action prima facie morally required; fourth, the relevance of the distinction between persons; and fifth, the fact that the existence of morally justifying conditions does not imply that even a morally good person must be inclined to act in the justified way. The following five subsections explain and defend each of these five points in turn.

I have encountered this objection in conversations with Robert Audi, John Deigh, Cairin Cronin, and Mylan Engel. Of course, the objection is natural if one holds the position that there *is* a strong analogy between practical and theoretical rationality, or if one is an internalist about practical reasons. But the possibility of purely justificatory reasons is a direct challenge to these views, and cannot be denied simply based upon its obvious conflict with them.

Prima facie immoral actions

According to many standard moral theories, there is a notion of a prima facie immoral action, and then there are certain types of reasons that can justify such actions. Typically, for example, the prima facie immoral actions are ones that harm other people. When, and only when, one believes (or ought to believe) that one's action will harm someone else, is one required to have a moral justification for that action. Robinson Crusoe, skipping stones in a lagoon, is not in need of any moral justification whatever for doing so. To say he is morally justified in skipping stones for the following reason – that it harms no one – is simply a very misleading way of saying that actions that harm no one are not in need of moral justification. The fact that an action harms no one is not a moral reason in any important sense. It cannot ever make any otherwise immoral action morally acceptable, as, for example, the fact that an action will alleviate someone else's pain can do.

According to these accounts, morality is primarily concerned with how people are to act when their actions affect *others*. And since we, or those we are concerned with, might be these very *others*, the types of acts we are primarily concerned with are those that might *harm* other people. This concern with how other people treat us is implicit in the reasoning, for example, that leads to the choice of principles in the various versions of Rawls's original position that have appeared in contemporary moral theories. The driving force behind the negotiations is a concern that one not end up in a terrible position as the result of the actions of others. Even Kant, whose morality seems as far removed from concern for self as any moral philosopher's, excludes some forms of behavior because an agent could not will that other people would behave that way towards *him*.[17] Hobbes, also, is clear that morality is a set of rules that we want *other* people to obey, and that the reason one gives up one's rights is to secure other people's compliance with moral rules.[18] Versions of morality that try to elaborate the Golden Rule, or that try to explain the force of the question 'How would you like it if people did that to you' also fit this pattern.[19] They imply at least that a good heuristic for the content of moral rules is whether one would want other people to follow them when one considers that their actions may be directed at oneself. This range of views agree that morality is primarily a set of rules by which people are to conduct

[17] See Kant (1988), pp. 51–2, where he discusses an example of an imperfect duty to others.
[18] See Hobbes (1994), Pt II, ch. xvii. [19] See Strang (1995), pp. 378–85.

themselves when their behavior will affect others. The point here is not to deny this. Rather, the point is that it is an extremely common view that some actions are morally acceptable *not* because there are sufficient moral reasons in their *favor*, but because there is nothing morally to be said *against* them.

Justifications for prima facie immoral actions

Virtually all moral theories make it prima facie immoral to kill someone, to cause someone pain, or to deprive them of freedom. And yet virtually all moral theories also allow that in some cases it is morally permissible do so. This characteristic cuts across the distinction between teleological and deontological views.

First let us consider utilitarianism. For a utilitarian it is certainly true that if all one knows about an action is that it will increase the risk of hurting someone, then the presumption is that one should not do it; it is prima facie *immoral*. And yet there is certainly no question that utilitarian views sometimes allow an action to be morally permissible even though it is virtually certain to hurt someone. Indeed, it is often seen as a flaw in utilitarian views – especially act-utilitarian views – that it is too easy to justify hurting someone, since the avoidance of a slightly greater harm for someone else (perhaps even for oneself) is generally sufficient to do so. Indeed, another problem with utilitarian views is that there is no easy means to distinguish such questionable justifications from still more questionable requirements. For if one is justified in causing an innocent stranger a great deal of pain in virtue of that action's being necessary to avoid the same pain for oneself *and* one's friend, this is in virtue of a certain utility calculation. And this same utility calculation will also *require* one to hurt the stranger.

A classic deontological moral view is Kurt Baier's. For Baier, it is certainly true that if all one knows about an action is that it will increase the risk of killing or hurting someone, then the presumption is that one should not do it; it is prima facie immoral. And yet, there is certainly no question that Baier sometimes allows an action to be morally permissible even though it is virtually certain to kill or hurt someone.[20] His way of allowing violations of moral rules is not, as in the case of utilitarianism, merely a question of weighing the goods and the evils. Rather, it is a question of there being a rule that allows such violations. In this way, Baier avoids an

[20] See Baier (1965), pp. 99–108.

unreasonably close connection between justification and requirement.[21] For the rules that allow violations – that is, the rules that provide moral justification – are not the same as the rules that make actions prima facie immoral. The rules that *allow* violations do not *require* one to make those violations. So Baier can explain the existence of a large class of actions that are morally justified but not required.

The set of justifying considerations is different from the set of requiring considerations

So far the argument has been concerned with prima facie immoral actions and justifications for them. A concept very closely related to prima facie immoral action is prima facie morally *required* action. For whenever one action is prima facie immoral, there is another action (namely *refraining* from the first action) that is prima facie required. And whenever an action is prima facie required, there is another action (again, *refraining* from that action) that is prima facie immoral. Therefore, many moral views claim that certain considerations place an action into the category of prima facie morally required; in fact these are the same considerations which place a *different* action into the category of prima facie immoral. Examples of such considerations, again, might be that an action will avoid causing pain to someone other than the agent, or that one promised to do some action. It is prima facie morally required to do actions when these considerations apply to them, and it is prima facie immoral to refrain from such actions.

The previous section pointed out that many moral theories also claim that certain considerations can provide justification for prima facie immoral actions. For many moral theories the set of justifying considerations is different from and wider than the set of considerations that make action prima facie required. That is, for many moral theories there is some consideration that can justify an action, but which cannot make an action prima facie required.

The above feature does not apply to as wide a range of moral views as the previous two features. For utilitarianism, it is in fact precisely the

[21] However, it is unclear whether Baier avoids an unreasonably close connection between immorality and irrationality. In Baier (1978), pp. 248–50 he attempts to deflect this charge by distinguishing irrational actions from actions which are merely 'contrary to reason.' But in fact his notion of 'contrary to reason' plays the same fundamental normative role as does the notion of objective irrationality in this book. Since there cannot be a sufficient justification for doing an action which is contrary to reason, on Baier's view, there also cannot be any such justification for doing something immoral.

considerations that make an action prima facie required that also serve to justify action. That is, if an action will avoid causing someone pain, it is prima facie required by utilitarianism. But now suppose that an action is prima facie immoral for some reason. It is also true, on a utilitarian view, that such an action might nevertheless be *justified* by the fact that the action will avoid causing someone pain. Thus utilitarianism does not have the characteristic asymmetry described in this section. But in fact it is also precisely this symmetry in utilitarianism that makes it objectionable. More particularly, if the way justification works is merely by a sort of balancing of pros and cons in a homogeneous field of considerations, then morality becomes amazingly demanding, and the distinction between persons is lost.[22] These are precisely the problems that have forced some to reject at least *act*-utilitarian views.[23] Whether or not one agrees with these criticisms, the point here is only that when a moral theory makes no distinction between what can justify and what can require, this is a notable departure from commonsense intuition.

For other moral views, the fact that an action will hurt *someone other than the agent* is a consideration that makes avoiding it prima facie required. And what makes it justified to hurt someone else is *not* merely that one's action will avoid hurting yet another person. Rather, it is often that the person has consented, in virtue of some benefit for that very person. For example, a plastic surgeon is justified in this way when she inflicts pain on her patients. Or, it might be that the action will *prevent a far greater* harm for someone. For example, I am justified in taking someone's car keys away from her if she is so drunk that if she drives she significantly increases the likelihood of seriously injuring or killing someone. Both of

[22] In fact, the latter objection only applies to consequentialist views according to which benefits and harms are always benefits and harms *for a particular agent*. Some distribution-sensitive consequentialist views may avoid this objection by ranking states of affairs partly based upon some index of the number of people who are independently pursuing their plans of life, and how successfully they are doing so. See Scheffler (1995), pp. 26–30 for a description of some of these ranking principles.

[23] See Rawls (1971), p. 27; Sen and Williams (1982), pp. 4–5; Smart (1991), p. 110. Smart gives an exceptionally clear statement of the second objection on behalf of nonutilitarians. This objection is different from the claim that utilitarianism is quite demanding, though it does lead very quickly to this objection. It is only the former objection that is peculiar to utilitarianism, for a Kantian view could easily be comparably demanding. But because of the Kantian duty to develop one's talents, and also to treat oneself as an end, there seems to be not only room for, but also a principled reason why one should spend a certain amount of one's time and resources on projects that are one's own. This may sometimes be a significant burden, but it does leave room for what is recognizably an individual human life.

these cases suggest views on which the set of characteristics of an action that make it prima facie required is not the same as the set of characteristics that could justify some action that stood in need of justification. Another sort of case, which leads into the next point, is one in which a person is justified in hurting someone because only by doing so can that person avoid a far greater hurt for himself. One example of this case might be knocking someone unconscious in defense of one's life from an attack by that very person. Another might be very roughly shoving someone out of the way (but not into harm's way) so that one will, oneself, avoid being hit by a car.

The distinction between persons is relevant to the distinction between justifying and requiring

Suppose one knows that an action will have certain harmful consequences and certain beneficial consequences for certain people. This knowledge alone is not always sufficient to determine the moral status of the action. For an action done by me, which has the *same relevant consequences for exactly the same people* as an action done by you, might have a very different moral status, because of the relation of the harms and benefits *to the agent*.[24]

For example, suppose that I am so violently allergic to eggs that should one break in my vicinity, I risk going into anaphylaxis. You and I have gone grocery shopping, and you have picked up a half-dozen of these dangerous objects. As we are unpacking, I take three eggs out of the carton, and show you how I have learned to juggle. I am pretty good at juggling, but am putting myself at risk of a significant harm. Many common views of morality will claim that, at least solely in virtue of the risk I am taking with my own life, I am doing nothing immoral. Of course it would be traumatic for you to see me die in your kitchen. But my claim is not that it is perfectly fine for me to juggle the eggs. My claim is only that it matters that it is *me* putting *myself* at risk. For suppose that, knowing of my allergy, *you* take out three eggs while unpacking, and show me how *you* have learned to juggle, putting me at equal risk. I maintain that you *are* doing something morally problematic just in virtue of putting me at risk. How is this difference between your action and mine to be explained?

[24] The suggestion is not that the mere addition of the knowledge of this relation will automatically suffice to determine the moral status of the action. Rather, it is that this information is *relevant* to such a determination.

Consent might be cited as a relevant factor. For it might be claimed that I cannot be taken to have any *objection* to my own juggling, but that I might have an objection to *your* juggling, and this difference in *consent* is what accounts for the moral difference. But suppose I assure you that I have absolutely no objection to your risking my life by juggling; suppose in fact that is I who bring up your recent efforts to learn juggling, and ask you to show me. There still is a significant moral difference in virtue of its being *you*, and not *me*, who is taking the risk with my life. If it is declared inconceivable that I should give such consent, it must be asked how it is conceivable that I should juggle the eggs myself. So we must locate the relevant difference elsewhere. The difference, as many moral theorists have urged, is that one needs a moral justification for risking harm to someone else, but does not need such a justification when risking harm to oneself.[25] It does seem right that if, just because I believe I will enjoy it, I hurt, disable, or deprive some person of freedom or pleasure, I have done something wrong. In order to do these things, I need a *moral justification*. But if, in order to give myself equal pleasure, I harm *myself* in these ways, I do not seem to require it. Perhaps I need *rational justification*. But that is not the issue here.

For those unpersuaded by this example, perhaps because of a belief that my consent to my friend's juggling the eggs really *does* make it morally unproblematic, there is the case of morally culpable negligence. If I simply do not give sufficient consideration to the obvious danger in which I place myself by some action, no one seriously believes I have done anything immoral thereby. Rather, my action is termed stupid or irrational. But if I fail to pay attention to the likely harmful consequences of my action on someone else, I am morally blamed. In neither of these cases is consent a relevant issue, since in neither case is the person being put at risk in a position to give consent.[26]

I am not arguing here that the correct moral view takes harm to others always to provide a prima facie moral requirement, while avoiding harm to self can only provide moral justification. All I am doing is trying to engage the intuitions of those who agree that it is often the case that the distinction between agent and other is relevant to the question of whether some

[25] See Baier (1954); Ross (1939), pp. 72–75 and 272–77; Stocker (1994), esp. p. 690; Darwall (1994), esp. p. 700.

[26] See Slote (1984), pp. 190–91 for more reasons to believe consent is not the relevant issue. In fact Slote (1992) argues that the asymmetry that he shows very clearly to be a part of commonsense morality counts *against* commonsense morality. But these latter arguments are far less persuasive.

consideration could require an action, or could only justify an otherwise immoral action. If one's agreement goes this far, then one should ask oneself why one takes the normative notion of practical rationality to be so different from that of morality. Consider that we do not generally take the Nazis to have been *irrational*, and hence free from blame.[27] But we do often consider irrational and free from moral responsibility those who harm *themselves* in equally significant ways for no reason. This suggests that avoiding harm to others does not provide a prima facie rational requirement to act, but that avoiding harm to oneself does. And yet we do not consider it irrational to harm oneself in order to avoid harm to others; we do not, generally, try to discourage children who want to be firefighters. This suggests that avoiding harm to others *does* provide a rational justification for otherwise irrational actions. When one thinks on this one should at the very least also begin to think that there is a good chance that in many respects morality provides a stronger analogy for practical rationality than does theoretical rationality. And one should grant that it is plausible that there is a relevant difference between practical reasons stemming from harm to the agent, and practical reasons stemming from harm to others. And, finally, one should see that it is arbitrary to take as the *default position* that there is no relevant difference between such practical reasons. We need arguments on either side before we embrace or deny this position.

Justifying considerations need not motivate even a morally good person

Suppose one is doing something that requires moral justification, such as taking money from someone. And suppose one is in fact justified. It is *not* typically true that the justification has any power whatsoever to *require* one to act on it. In many cases justifying considerations are ones that it would be morally better to *ignore*. For example, David Copperfield's enemy Uriah Heep is morally justified in confiscating a piece of property offered as security on a loan he has made, when the terms of

[27] The move from irrationality to lack of moral culpability is of course not a simple one: some irrational actions can be morally blameworthy. But if the psychological defect that makes one's action irrational is precisely what explains why one performed an action that, in a normal person, would have been morally blameworthy, then there is more reason to think that one's irrationality mitigates one's moral responsibility. And this would be the case for the Nazis if an indifference to or delight in the suffering of others were itself irrational. I discuss the connection between irrationality and moral responsibility at greater length towards the end of chapter 4.

the loan have been violated by a delay in payment. But it is emphatically *not* the case that that clammy fellow is morally *required* to confiscate the property. If he decides not to confiscate the property – not to act on a justification that is available to him – not only isn't the action immoral, and not only doesn't it count against his moral character, it counts *in favor* of his moral character. While this is not true of all moral justifications, the example shows that the reasons people offer in order to justify otherwise immoral actions are not reasons that they are always morally required to act on, or even to be motivated by, even in the absence of countervailing reasons.

Another example of the above phenomenon involves the justified infliction of pain on someone who is harming one. It is commonly accepted that it is morally justified to harm, or even to kill, in self-defense. But if one is a committed pacifist, one might accept any level of harm – even death – if the only alternative was to harm one's attacker. And in doing so, one would not be behaving immorally at all. No matter how strong this sort of justification for harming one's attacker became, it would never morally require one to harm him in the slightest way. Indeed, it is at least plausible that being completely unmotivated to act on this type of justificatory reason would count in *favor* of one's moral character.

The parallel here with practical rationality is that certain reasons, relevant to the rationality of an action, need not motivate even a *rational* person. These, of course, are *purely justificatory* practical reasons, among which, this book suggests, we should classify all altruistic reasons. If an agent is never motivated by such reasons, that agent can, nevertheless, always act rationally. For one is not rationally required to act on purely justificatory reasons. Someone who is habitually selfish, petty, and mean, need not ever do anything that would generally be regarded as irrational. Perhaps it is the truth of this claim that lends an initial plausibility to the claims, discussed above, of egoists and desire-satisfaction theorists.

CONCLUSION

This chapter has drawn a parallel between practical rationality and common moral views in order to shift the burden of proof on two issues. Many common moral views imply the existence of purely justificatory moral reasons: reasons that do not provide even prima facie moral requirements. And the dividing line between reasons that do, and reasons that do not, provide prima facie moral requirements is often drawn by means of the

distinction between the agent and other people. Attention to these features of commonly accepted moral views should make it easier for theorists to go against an overwhelmingly popular philosophical tendency to simplify the notion of practical rationality by crediting the existence of only one sort of reason: one that, prima facie, both justifies and requires. If one's intuitions about morality allow purely justificatory reasons and a relevant distinction between persons, then one should begin to ask why it should be different when the topic is practical rationality. The burden of proof should shift onto the shoulders of those who maintain that all reasons relevant to the rationality of action can both justify and require, and that there is no fundamental difference, relevant to the rationality of one's actions, between one's own interests and those of other people.

If the above line of argument is correct, all accounts of rationality that claim that one is rationally required to act on one's best reasons, or the balance of reasons, or the strongest reasons, or one's judgments about these 'bests' or 'balances,' must be wrong. For suppose, as this chapter has argued, that reasons can play two different roles in determining the rational status of an action: requiring action, and justifying action that would otherwise be irrational. And suppose that the only two reasons relevant to some particular action are a strong requiring reason in favor of the action, and a *very* strong purely justificatory reason against the action.[28] For example, suppose one is on holiday. One is scheduled, on the day of one's return, for an important job interview: one that cannot be rescheduled, and the likes of which one may never see again. But on the day before one's return, a terrible storm hits the underdeveloped town where one is vacationing, and one's special talents will save many villagers from harm. What is 'the rational thing to do'? On the view offered here, there is no answer to this wrongly-framed question, which presupposes a unique answer (or, at best, a tie for first place) that tells one what one is rationally *required* to do. It would indeed be irrational to miss the job interview for *no* reason, but one has a sufficient reason; indeed it is *more* than sufficient, for it would justify a sacrifice of a more significant good than a mere job interview. But the justifying reason is not a *rationally requiring* one, so it is not irrational to go home as scheduled. Nor would it be irrational to go home for some *less* strongly compelling reason, which shows that the rational permissibility of the former choice was not a matter of the equality of reasons for and

[28] The notion of strength here – indeed, in this chapter as a whole – is an intuitive one. It will be given a clear content in chapter 4.

against the action. Talking of 'bests' and 'balances' here is not helpful, because although the reason to stay and help may be a better justifier than the reason to return, it nevertheless does not require one to act on it. Nor can one say that purely justifying reasons are simply weaker. For it might not be rational to risk one's life for the job interview, while it would be rational to risk one's life to help the villagers. So in this sense the justifying reason is in fact stronger. In the face of such examples, and of the parallel between practical rationality and morality, one should admit that the logical role of some reasons may simply be to justify, and that this is a different logical role than the role of requiring.

One final remark on another possible parallel between morality and practical rationality may be helpful here, for those who cannot yet wholly embrace the position advocated in this chapter. Many Kantians claim that *all* immoral behavior is irrational. This is quite an extreme position, and though it may be argued for, is not overly plausible on its face. Nevertheless, even its face may contain a grain of truth. Many people do believe, after all, that certain *extremely* immoral behavior is irrational: performing medical experiments on prisoners in concentration camps, for example, or allowing a baby to drown because one is late for the movies. What is interesting about this belief is that it is controversial. Everyone of course agrees that such behavior is the height of immorality. But is it *irrational*? There are conflicting intuitions here. The parallel with morality is this: there are also conflicting intuitions about the *morality* of actions that involve extremely severe *self*-inflicted harms, such as suicide or drug abuse. Some people have a basic intuition that suicide and drug abuse are immoral *in virtue of the harm to the agent*. But other people hold that if an action will harm only the agent, then it cannot be immoral. The striking similarity in the structure of these controversial cases in morality and practical rationality strengthens the case for a parallel between these two notions. And, for those who wish, it may also provide the material for constructing a view that is very similar to, but slightly weaker than, the view advocated in this book. For it may be that harm to others does have some relatively *small* power to make actions rationally required, so that when the harms are very great indeed, they can generate a nontrivial prima facie rational requirement. And it may also be that, in a similar way, harm to self does have some relatively *small* power to make actions morally required. The proposed modification would do nothing to undermine the main point of this book. The parallel between practical and theoretical rationality

would remain just as strongly opposed – indeed, even more strongly. And even on the modified view, altruistic reasons would have very *different* requiring and justifying strengths; any particular altruistic reason would be able to justify much more than it would be able to require. Because of this difference in justifying and requiring strength there would still be many cases, like the job-interview case and the charity case, in which one would not be rationally required to act on the stronger (justifying) reason.

3

The criticism from internalism about practical reasons

The previous chapter argued for the existence of a class of practical reasons – purely justificatory practical reasons – that have no power to make actions rationally required. It also suggested that altruistic reasons form a significant part of this class. The strategy of that chapter was to point out that a wide range of moral views grant, either explicitly or implicitly, the existence of moral considerations that function in the same way. That is, they grant the existence of considerations, the presence of which can change an otherwise immoral action into a morally permissible one, and yet that do not seem to be the sort of considerations that must weigh, in a positive way, in the motivational economy of a virtuous person. On the strength of some examples, and of further points of analogy between morality and practical rationality, it was then suggested that practical rationality also includes considerations that play a similar normative role. But if practical rationality does include reasons that function in this purely justificatory way, then it seems that there could be actions, favored by such reasons, that even a rational agent might not be motivated to perform – and this could be true even when the agent knows that there are no countervailing reasons. This implication offends against a dogma that is virtually universally accepted by contemporary moral theorists: roughly, that rational agents will be motivated to some degree by any practical reasons of which they are aware, and which are relevant to their choice of action.[1] Even those who balk at such a stark statement of the view typically nevertheless accept the idea that in the absence of countervailing reasons, a rational agent must act on any reason of which she is aware.[2] It is the

[1] See Cullity and Gaut (1997), p. 3. It is worth noting that not one contributor to the Cullity and Gaut volume challenges the internalism requirement. See also Cohon (1986), esp. p. 556; Smith (1994), pp. 60–62 and 151–77; Nagel (1970), pp. 66–67; Velleman (1996), pp. 694–726; Darwall (1983), p. 52.

[2] Joseph Raz, for example, offers some compelling arguments against the bolder view, but accepts the weaker. See Raz (1999b), p. 99.

purpose of this chapter to rebut an objection to the idea of purely justificatory reasons that emerges from an uncritical acceptance of such a view. The strategy is borrowed from an argument that Christine Korsgaard has recently used in rebutting an objection to the Kantian view that practical reason can motivate action by itself. The objection she addresses comes from Hume and neo-Humeans such as Bernard Williams.

Hume and Williams, among many others, argue, against Kantians, that practical reason is incapable of generating motivation on its own.[3] This view is best interpreted as making the following normative claim: no particular basic motivations are, in themselves, rationally required or prohibited. Christine Korsgaard calls this view 'motivational skepticism,' and in "Skepticism about Practical Reason" she shows that, whether or not it is a correct view, it should not be taken as *fundamental*.[4] Rather, motivational skepticism must be based on a prior view about what the principles of rationality are. In particular, it must be based on a prior skepticism about the existence of rational principles that can classify an action as irrational based solely on a description of the action, and without additional information about the contingent motivations of the agent who is performing the action. This form of skepticism she calls 'content skepticism.' Another way of characterizing content skepticism is the following: no action type is simply irrational (or rational), regardless of what the agent wants. One example of a principle denied by a content skeptic would be:

> M It is irrational to refuse to take medicine that will save one from lingering death, and restore one to perfect health.[5]

Another much more general such principle will be offered later in this chapter.

[3] See Hume (1978), p. 415; Williams (1981), p. 105. Alfred Mele (1989) convincingly advances a distinct but related *causal* view: that practical *reasoning* cannot produce motivation underived from antecedent motivation. His view is compatible with the claim that, in order to count as rational, an agent must have certain basic motivations that cannot necessarily be reached by *reasoning* (such as aversions to pain, death, disability, etc.). Thus, as Mele rightly notes, his position is neutral as between Humeans, such as Williams, and their critics, such as Korsgaard. See Mele (1989), pp. 419, 432, and 436 n. 19.

[4] Korsgaard (1996b).

[5] See Williams (1981), p. 105, for a denial of this principle, based on a prior acceptance of motivational skepticism. Williams would deny this principle even if it were expanded to exclude the cases that make such a denial plausible: cases in which, for example, one's death produces great benefits for others, or in which one anticipates that one's rescue from death will only lead to a life of great sorrow.

It would be more accurate to say that motivational skepticism must be based on a prior view of what the principles of *objective rationality* are. But Korsgaard does not seem to distinguish between objective and subjective rationality. Indeed, because her 'source of normativity' is the agent's own practical identity, and because she takes reasons to be 'endorsed impulses' rather than facts, she seems to take the subjective notion of rationality to be the fundamental notion.[6] Nevertheless, this difference is not very important here, for the dependence of motivational skepticism on an account of subjective rationality can trivially be extended to a dependence on an account of objective rationality, if one admits that objective rationality is a distinct and more fundamental notion. The question of whether or not certain basic motivations are rationally required is a question about how a rational agent will act. A motivational skeptic would have to deny M, even if it is understood as a principle of objective rationality, since if it is objectively irrational to refuse the medication, then a fully informed rational agent would be motivated to take it. Indeed, any rational agent who believed that he was in a situation ruled out by M would be so motivated. Thus even if M is understood as a principle of objective rationality, it continues to entail that certain motives are rationally required. Because of this, although the following discussion will often proceed in terms of principles of subjective rationality – for this is the way Korsgaard presents her arguments against Hume and Williams – it can easily be understood in terms of objective rationality, simply by assuming that the agents who figure in the examples and arguments possess all the relevant information about their actions.

Korsgaard argues that until content skepticism has been established, it remains an open question whether practical reason can generate motivation without relying on contingent antecedent motivation. That is, it remains an open question whether motivational skepticism is true. And she does not think content skepticism has been established. This point – that content skepticism has not been established – coheres with the overall picture of reasons and rationality that this book is advocating. But this chapter argues that Korsgaard has not taken her arguments far enough, and that their consistent application undermines the axiomatic status of the dogma mentioned above: a dogma that Korsgaard herself, along with virtually all contemporary moral theorists, also accepts. Following

[6] See Korsgaard (1996a), pp. 94, 99 n. 8, 108.

Korsgaard, we can call this dogma 'the internalism requirement on practical reasons.'[7]

The internalism requirement is a motivational view about reasons that Korsgaard clarifies and endorses in the course of her argument. It holds that for a consideration *C* to be a reason for agent *A*, it must succeed in motivating *A*, given that *A* is rational, and that *A* is aware of *C*.[8] That is, according to the internalism requirement, if a putative reason fails to motivate an agent to whose action it is relevant, then either it isn't really a reason, or the agent is to some degree irrational. The internalism requirement demands more of a reason than that it *could* motivate a rational agent who had it. It requires that it *would* motivate a rational agent who had it.[9]

This chapter argues that just as Hume's and Williams's motivational skepticism depends upon a prior acceptance of content skepticism, the internalism requirement depends upon the prior acceptance of the following view.

The requirement view: All practical reasons are prima facie rational requirements. That is, if one acts against such a reason, then one is either acting irrationally, or one is acting on other countervailing practical reasons of at least equivalent strength.

The conclusion, similar to Korsgaard's, will be that until the requirement view has been established, it remains an open question whether the internalism requirement is valid. And the requirement view has not been established. Indeed, though it is sometimes stated, one would be hard-pressed to find any argument for it at all. One reason for this lack of argument is that the internalism requirement, which follows from the requirement view, has not been widely challenged, so that there is no felt

[7] Recently, Sigrún Svavarsdóttir has provided compelling arguments for a similar conclusion regarding a different but related view: moral judgment internalism. That is, she undermines attempts to use consonance with moral judgment internalism as an adequacy condition on moral theories. See Svavarsdóttir (1999), pp. 218–19.

[8] For the remainder of this chapter, the 'awareness' rider should be taken as understood.

[9] There are also certain types of externalists who adhere to the internalism requirement as it is here stated. The argument of this chapter is directed at such externalists as well. See, e.g., Parfit (1997), p. 101, and Brink (1986), p. 36. In this latter article, Brink argues for the conceptual possibility that the recognition of a moral requirement might fail to give an agent a reason to act, and this is partly what his externalism consists in. But Brink also equates the question of whether the recognition of a moral obligation gives an agent a reason for action with the question of whether an agent would be irrational subsequently to fail to care about such moral requirements. This equation depends upon the internalism requirement as it is here understood.

need to defend it. Of course there have been principled defenses of *other* forms of internalism, but these often rest on an unargued assumption of the internalism requirement on practical reasons, or of the requirement view.[10]

If, against the requirement view, some reasons are not prima facie rational requirements, then the internalism requirement will in fact be false. To make the parallel with Korsgaard's argument against Hume and Williams more obvious, it is useful to note that the internalism requirement is a *motivational* thesis about reasons, and that the requirement view is a thesis about what the *content* of a reason-claim is. In a sense, the argument offered here is the same argument as Korsgaard's. The difference is only that the present argument recognizes a wider variety of potential principles of rationality than Korsgaard does, and, hence, the potential for a more complex relation between reason-claims and claims about the rationality of actions.

KORSGAARD'S ARGUMENT: MOTIVATIONAL SKEPTICISM DEPENDS ON CONTENT SKEPTICISM

Although Korsgaard argues that motivational skepticism about practical reason always depends upon a prior acceptance of content skepticism, she does not directly attack content skepticism itself. And so she does not directly attack motivational skepticism either. Rather, she is concerned with the link between the two forms of skepticism. Her point is that philosophers who wish to argue that reason is unable, on its own, to motivate action, must *first* argue about which principles are to be admitted as rational principles. And, if she is right, they cannot reasonably use motivational skepticism as a premise in such an argument. Korsgaard uses Hume as an example of a philosopher who is sometimes interpreted in a way that commits him to this mistake. But she shows, convincingly, that his actual argument respects the priority of content skepticism.

[10] See, e.g., Smith (1994) for a defense of moral judgment internalism based on an undefended assumption of the internalism requirement on practical reasons. See Foot (1978b), p. 152 for an implicit argument for the internalism requirement on practical reasons, based on an undefended assumption of the requirement view. Derek Parfit makes the same move in Parfit (1997), pp. 101, 130. See also Broome (1999), pp. 400–1, for a relatively clear, but undefended statement of the requirement view. Broome states the position in terms of 'ought,' rather than 'is rationally required to,' but the position is essentially the same, since for Broome ought-claims are 'strict demands,' rather than *pro tanto* or prima facie ones.

Hume

On Hume's view reason only helps us choose efficient means to ends. For Hume, normative standards for the choice of those ends come from another source. Reason cannot even rank ends, or determine that we should satisfy the ends we regard as 'our greatest and most valuable enjoyments.'[11] Hume's argument is that reason is concerned only with abstract relations of ideas and relations between objects. When reason is concerned with the first of these types of relations, it is doing mathematics, which cannot give rise to motivation. And when reason is concerned with the second type of relation, it is involved in causal reasoning, which gives rise to motivation only from pre-existing motivation. Thus, by surveying the types of rational processes, we can see that neither of them can generate motivation on its own. As Korsgaard notes, this particular argument depends in an obvious way on presuppositions about what processes count as *rational* processes. That is, if we go so far with Hume as to grant that reason is answerable only to principles of mathematics and causality, then it is indeed quite plausible that reason cannot, on its own, direct an agent towards one action rather than another, independently of some contingent and antecedent motivation. If Hume could defend his notion of the *content* of practical reason – its limitation to math and science – he would therefore be able to defend his motivational skepticism.

What Korsgaard is rightly pointing out here is that it only *looks* as if Hume is arguing *from* motivational skepticism *to* a restriction on what counts as a rational process. But he is not. Rather, the argument goes in the other direction. Hume argues that reason's essential function is to judge truth and falsity. It does this by making determinations of the accuracy of representations. And he claims that actions and passions do not represent anything, and therefore cannot be true or false. Thus reason has nothing to say about them. For Hume, no rational principle rules in, or rules out, any particular action, and, correspondingly, no rational process, on its own, necessarily leads to a motivation to any particular action.[12] That is,

[11] Hume (1978), p. 416.

[12] The logical relation between rational principles and rational processes is a complex one, and no explicit theory of that relation will be offered in this book. But it should be clear that if it is a rational principle that one should take efficient means to one's ends, then the psychological processes by which one discovers efficient means, and by which motivation to take those means is generated, are rational processes. For an excellent discussion of the relation between the processes of practical reasoning and the principles of practical

after restricting the nature of reason, Hume goes on to demonstrate that it cannot produce motivation out of whole cloth. With this description of Hume's argument, Korsgaard has at least *illustrated* her point: she has shown us a philosopher whose motivational skepticism is clearly dependent on an antecedent content skepticism.

But there is another way in which Korsgaard exploits Hume, to reach a more general conclusion. Because of his view that rational principles can only judge of truth and falsity, Hume believes they have nothing to say about motives or actions, since neither motives nor actions can be true or false. Motives, and actions based on them, cannot, for Hume, be straightforwardly irrational. They can only be *derivatively* irrational. They are derivatively irrational when they are based on false beliefs about objects or about the efficacy of means. But Korsgaard points out that there is a way in which reason might be seen to bear on action in a less derivative way, even on a view very close to Hume's. *True* irrationality, as she calls it, would be to choose inefficient means in full knowledge: without any mistaken beliefs. Hume explicitly denies the existence of true irrationality.[13] That is, he denies that the principle 'choose efficient means to your ends' is a *normative* principle, for he denies it one of the necessary characteristics of a normative principle: the possibility of not being followed. By allowing the possibility of true irrationality, Korsgaard turns the principle into something normative.

Now, even if true irrationality is granted, by taking the principle 'choose efficient means to your ends' as a normative principle of reason, it need not be granted that reason can generate motivation without the existence of some antecedent end. Certainly the particular instrumental principle under discussion is incapable of doing it, simply because it is instrumental, and therefore depends upon contingent ends to supply it with content. But Korsgaard argues that once true irrationality is granted, our attitude towards *other* putative principles of reason should become more liberal. In particular, our attitude might become so liberal as to allow principles that do bear directly on particular actions. It is this possibility that Korsgaard wants to argue for, and so she spends some time arguing for the existence of true irrationality.

rationality, see Mele (1989). In light of Mele's paper, it should be borne in mind that the processes of practical reason*ing* may be a *subset* of rational processes. That is, the processes by which one acts rationally, and the failures of which can explain irrational behavior, may include more than merely processes of reasoning.

[13] Hume (1978), p. 416.

True irrationality

Korsgaard argues that it is possible that a person could engage in flawless means/end reasoning, and yet fail to be motivated to take recognized means to her acknowledged end. Her extremely plausible explanation for this is simply that there might be interference in the transmission of motive force from acknowledged ends to recognized means. There are things that prevent us from acting rationally: rage, depression, drugs, arrogance, aneurysm. Here the case of theoretical irrationality *does* provide a useful analogy.[14] We all admit that there might be *theoretical* reasons decisively in favor of *believing* some claim, and yet I might be unconvinced by them. This would not make them any the less *reasons*. It might only be evidence that I am *irrational*. So Korsgaard thinks Hume is wrong, as he surely is, in failing to acknowledge the existence of true practical irrationality.

Once it is admitted that it is possible to be truly irrational, Korsgaard points out that Hume's limitation of rational processes to mathematics and means/end reasoning is less compelling. She illustrates this point with the example of prudence.[15] Hume takes prudence to come from a passion that, should it disappear, would take with it both the motive *and* the reason for prudential behavior. In this respect the 'prudential' passion is for Hume just like any other contingent passion. It possesses no special rational authority. One can have it, or lack it, and one's lacking it has no bearing on one's rational status, or the rational status of one's actions. But if one admits, as Hume does not, that one can sometimes simply fail to be responsive to rational considerations – if one admits the possibility of true irrationality – then when one fails to take means to one's greater good, there are two ways of describing what has happened. The first is Hume's explanation: one simply, and in a rationally acceptable manner, did not have one's greater good as one's end. The second is Korsgaard's explanation: one was irrational for failing to have had one's

[14] In fact, the analogy with morality supports the same point. There may be moral considerations that count decisively in favor of some action: that I fail to perform the action does not alter that fact; it only means that I am morally imperfect.

[15] This choice of illustration by Korsgaard shows that she takes prudential motivation as a potential rational process. And yet the *reasoning* involved in being prudent may be of only the broadly instrumentalist and constitutive sort countenanced by motivational skeptics. This suggests that Korsgaard may agree that rational processes need not all be processes of *reasoning*. See note 12 above. Such a view allows one to reject the internalism requirement while still holding the view that Alfred Mele calls 'motivational internalism': roughly the view that arguments will generally only have a practical effect on an agent if that agent has the appropriate antecedent motivations. See Mele (1989), pp. 422–29.

greater good as one's end. One is precluded from taking the second explanation if one has, like Hume, limited practical reasoning to mathematics and means/end reasoning, but that limit needs defense. By admitting the possibility of true irrationality, one can admit that prudence has rational authority *even though* it sometimes fails to motivate. That is, the obvious fact that many people are not prudent no longer rules out, or even argues against, prudence as a rational principle. One's decision to admit prudence as a rational requirement will depend on arguments that cannot be refuted simply by showing that prudence sometimes fails to motivate. This way of dealing with putative rational principles on which people sometimes fail to act widens the field of principles for which philosophers can argue.

Williams

It is of course possible to acknowledge a very wide field of rational principles – to doubt that Hume's notion of rational processes is even remotely complete – and yet still maintain that motivation cannot come from reason alone. That is, one can maintain motivational skepticism even if one grants that rational principles involve much more than the principle that one should choose efficient means to one's acknowledged ends. This is what Bernard Williams does.[16] And it is Williams who is Korsgaard's next target. Williams is arguing for the view that reason-claims must imply a motive. That is, he claims that if I say that you have a reason to ϕ, I must be taken to mean you have some desire or goal that would be served by your ϕ-ing, or that you adhere to some principle that speaks in its favor, or that some other similarly motivational claim is true of you. Of course, you might have some other goals that yield reasons against ϕ-ing. Neither Williams nor Korsgaard are committed to the implausible view that one can only be said to have a reason to do an action if that action turns out, all things considered, to be favored by reason. Indeed, the notion of being so favored, all things considered, presupposes that there may be reasons pulling in different directions. Any given action will probably have reasons both for it *and* against it, and hence not all reasons will, or even could be, *acted* on, even by a perfectly rational agent. That is why the discussion is always cast in terms of *motivation*, and not of *action* or *intention*. It is at least *plausible* that all reasons provide motivation to

[16] Williams (1981).

rational agents who have them. It is *not* plausible that all reasons provide such agents with sufficient motivation to prompt actual action. To restrict talk of reasons in such a way that they always do provide rational agents with sufficient motivation to act is to distort the notion of a reason beyond recognition.

Williams argues that a reason-claim must imply a motive, because otherwise we could not use the reason to explain the action for which it was claimed to have been a reason. Therefore, the argument continues, unless a consideration would motivate an agent if that agent were fully rational, it cannot be a reason for that agent. This, again, is what Korsgaard calls 'the internalism requirement.' Though Korsgaard is arguing against Williams, she does think that the internalism requirement is "clearly correct."[17]

Korsgaard admits that different considerations have the capacity to motivate different individuals. It might seem therefore that the internalism requirement entails that we can only determine whether or not a consideration is a reason for an agent by looking at the particular motivational capacities of that agent.[18] This is in fact Williams's position. He claims that reasons are always relative to what he calls the agent's 'subjective motivational set' – a collection of motivational entities that includes desires, but that can also include principles one adheres to, projects one has, and other things not resembling desires except in being sources of motivation. Williams assumes that rational processes must start from something that can motivate: that they must start from something in one's subjective motivational set. But Williams does not take means/end reasoning to be the only rational process by which motivation for particular actions can be teased out of one's subjective motivational set. He even goes so far as to allow that imagination might be such a process: one imagines what it would be like to achieve some end, and a desire for that end might be created. But even in this case Williams still claims that the capacity for imagination to engender a desire is dependent on the contents of one's subjective motivational set.

But, Korsgaard argues, once we, like Williams, have abandoned the Humean view that logic and means/end reasoning are the sole rational processes, there *may* turn out to be processes of practical reason that can yield motivation on their own. Korsgaard's point is that until rational processes are exhaustively inventoried, it is unsettled whether they can motivate us on their own. For example, if all rational people could,

[17] Korsgaard (1996b), p. 329. [18] Korsgaard (1996b), p. 325.

in principle, be convinced to accept some particular practical principle by some yet-to-be-discovered argument, then, even on Williams's view, that principle would yields reasons for everyone. According to Korsgaard, Williams simply denies that there could be an argument like this. But this shows that Williams's motivational skepticism about practical reason *depends* on a denial that there are any *substantive* rational principles that could be shown, by argument, to be valid. Therefore Williams cannot be taken as *showing* this to be so. It only *looks* like Williams limits the principles of practical reason by means of the internalism requirement.

But, Korsgaard rightly argues, the internalism requirement cannot, by itself, limit the content of principles of practical reason. This is because of the possibility of true irrationality relative to any proposed principle. When someone fails to be motivated by the reasons that some putative principle supplies, we can always preserve the validity of the internalism requirement by claiming that this failure is sufficient to show that the person is acting irrationally. Therefore the internalism requirement cannot limit the content of principles of practical reason to ones that only generate reasons out of an agent's antecedent motivations. For example, consider the principle 'Death is to be avoided at whatever cost.' Though such a simplistic principle is certainly not valid, it is not ruled out *merely* by the internalism requirement. For we can consistently maintain both that it is a rational principle, and that the internalism requirement is true, by claiming that anyone who does not avoid death at all costs is acting irrationally.[19] Of course, our ability to make this claim does nothing to *favor* the position that the principle 'Avoid death at whatever cost' will provide motivation to every rational agent.[20] Rather, our ability to make this claim shows only that citing the internalism requirement cannot, by itself, *rule out* that principle.

[19] Elsewhere Korsgaard denies that she defends the internalism requirement in this way. See Korsgaard (1997), p. 219 n. 11. There she seems to advocate the position that even after one has shown that some principle presents an unconditional normative requirement, one must still show that rational agents will be motivated to act in accord with it. This suggests that, at least at that point, she is taking the notion of rationality, as applied to agents, as a purely descriptive notion. At the very least, it suggests that she is using the notion of rationality here in a way that is conceptually independent of the idea of complying with unconditional normative requirements.

[20] See Dreier (1990), pp. 12–13. Dreier points out that this way of preserving the internalism requirement, because it is available to the advocate of *any* rational principle, cannot be used to argue in favor of any *particular* rational principle.

CONSEQUENCES OF THE ARGUMENT: THE INTERNALISM REQUIREMENT IS TOO STRONG

According to Korsgaard, the question is open as to whether there might be some processes of practical reason that can motivate all rational agents regardless of their contingent antecedent motivations. This opening of the question can of course be done without bringing the internalism requirement itself into question: this is in fact what Korsgaard does. But the possibility of such principles of reason drastically changes the way we can explain the truth of the internalism requirement. Consequently it suggests a different understanding of the requirement itself.

Here is how someone like Williams might understand why the internalism requirement is true. Adherents of views like Williams's believe a reason-claim is true in virtue of the existence of some antecedent motivation from which rational processes could produce a motivation to do some particular action. The picture this creates is one of a sort of reservoir of motivational fuel, from which rational processes, like the processes of internal combustion and transmission, extract and direct energy in specific directions. Now, if an automobile has a supply of fuel, and is in gear and running, but does not have its wheels move, then there is no question but that there is something wrong with the transmission mechanism. And in the same way, if it is true that a person has a reason, and therefore has 'motivational fuel,' and is awake and aware of the reason, and yet that person is not motivated by that reason, it is also tempting to say that there must be a failure in rationality. This is a natural analogy when rational principles are conceived of as principles that essentially govern only the *transmission* of motivation. Thus, when one takes one's own ends as the ultimate source of all reasons, as Williams does, it is natural to claim that it is irrational not to be motivated, to some degree, by every reason one has.[21] That is, it is natural to adhere to the internalism requirement.

So, if one grants Williams his motivational skepticism, he might seem to provide an *explanation* of the truth of the internalism requirement. He might be seen to do this by saying that all reasons stem from antecedent motivation, so that an appropriate motivation always exists, at least in a

[21] One might point out that in some cases incompatible means ϕ and π are available. Then it would be plausible to say that even a rational agent either would not be motivated to do ϕ or would not be motivated to do π. But in this case it is still true that the *same reason* favors both ϕ and π. That reason *is* therefore providing motivation, no matter which option the agent picks. So this kind of situation (which is the typical one) does not undermine the temptation of the internalism requirement.

fully rational agent, whenever a reason does. The nature of this explanation makes it tempting to say that a reason *must* motivate an agent, insofar as that agent is fully rational and aware of the reason. This happens because of the 'transmission of motivation' explanation of the capacity of reasons to motivate. When Korsgaard undermines this explanation of the motivational capacity of reasons, she removes one temptation to embrace the internalism requirement. Of course there are other temptations. The remainder of this chapter attempts to show how they too might be resisted.

THE DENIAL OF THE INTERNALISM REQUIREMENT

In criticizing the fundamental nature of the internalism requirement, we can use an argument *formally* identical to the one Korsgaard makes first against Hume and then against Williams. Hume and Williams *seem* to be arguing *from* a motivational requirement on principles of reason, *to* conclusions about what to admit as rational principles. Actually, Korsgaard shows, they are each *first* claiming (or assuming) something about what rational principles there are, and *then* drawing the motivational principle from their respective claims. Hume's initial claim is that reason judges solely of truth and falsity. He then concludes that principles of reason cannot produce motivation on their own. Williams's assumption is that no rational principles exist that could form the basis of a compelling argument showing some *substantive* practical principle to be valid. He concludes that one's acceptance of substantive practical principles will therefore always depend on one's antecedent motivations, and hence that the reasons for action that stem from one's acceptance of such principles will, like all of one's other reasons, also depend on those antecedent motivations.

But Korsgaard herself is assuming something nontrivial about the content of rational principles and reasons, and it is these assumptions that lead her to a motivational conclusion about them. Her undefended assumption is that reasons are prima facie rational requirements.[22] That is, she assumes that in order to act against a reason in a rationally permissible way, one always needs some countervailing reason. This assumption, which sounds very plausible in the abstract (where most philosophical discussion takes place), and which is almost universally held, leads her to the motivational conclusion she calls 'the internalism requirement.' Now, it is almost

[22] For one relatively clear statement of this assumption, see Korsgaard (1996a), pp. 225–26.

certainly true that *some* reasons are prima facie rational requirements. But, according to the view on offer in this book, not *all* reasons are. How might some reasons *not* be prima facie rational requirements? Consider the following possible principle of rationality:

> P It is irrational to do anything that one believes will cause one harm, unless one also believes that someone (perhaps oneself) will thereby be spared at least as significant a harm, or that someone (perhaps oneself) will thereby receive at least as significant a benefit.[23]

P would be made more plausible, but more complicated than present purposes require, if it were couched in terms of *likelihoods* of harms and benefits. Moreover, P stands in need of some account of what counts as a harm or benefit, and how it is determined that harms or benefits are to be measured as compensating for each other. Another modification to P might be to allow that *causing harm to others* also stands in need of rational justification, though *refraining from benefiting* them does not. All these issues will be dealt with in chapter 7. For present purposes, such modifications, clarifications, and additions are not required. Present purposes are served by the recognition that a principle *structurally* similar to P *could* be a rational principle.[24]

Structurally, P is usefully understood as having two parts. The first part specifies a class of actions that are potentially irrational: those that the agent believes will cause him harm. The second part specifies the sort of considerations that would make it rationally *permissible* to perform such an action: when anyone, possibly even the agent, will thereby avoid at least as significant a harm or gain at least as significant a benefit. The unless-clause in P is to be read as involving an inclusive 'or,' so that it yields a rational *permission*, rather than specifying conditions under which the agent is rationally *required* to do something that he knows will bring him harm. Some examples will help illustrate this.[25] P makes all of the following actions rationally permissible:

[23] The corresponding principle of objective rationality, obviously, simply removes the relativization to the beliefs of the agent. In fact, the relativization in P is too simple to make P a plausible principle of rationality, but the ways in which it oversimplifies are not relevant to the present discussion. See chapter 7 for a fuller discussion.

[24] B. Gert (1998), esp. chs. 2–4 defends a view with a structure similar to P.

[25] All these examples assume that no other significant reasons bear on the case.

(1) *Avoiding harm even though one could prevent more significant harm to someone else*. It is rationally permissible, according to P, to refrain from giving a hundred dollars to famine relief, even though one would not miss the money much if one did, and even if one knew that one's donation would probably prevent serious illness for a handful of impoverished children.

(2) *Doing things that do not harm oneself, just because one feels like it*. P claims that it is not irrational to pick up a stone and throw it into the woods, just because one feels like it.[26]

(3) *Suffering harm in order to avoid equivalent or greater suffering by someone else*. That is, P claims that all the following instances of altruistic behavior are rationally permissible: throwing oneself on a grenade to save one's fellows, giving away one's last bit of food to someone who is obviously more hungry, giving up one's seat on a bus to someone who seems tired, etc.

(4) *Suffering harm in order to provide someone else with a benefit that is at least as significant*. Suppose one believes that one of one's colleagues would benefit from detailed comments on a recent draft of a paper, but that one realizes that giving such detailed comments will be painfully boring, will involve a lot of time and effort, and will delay one's own research. P claims that nevertheless it is rationally permissible to volunteer to give detailed comments on the draft.

Of course, given the rational permissibility of actions such as those in example 1, none of the actions specified in examples 3 or 4 are rationally *required*. This is true even though the harms one would thereby prevent, or the benefits one would thereby produce, might be more significant than the harms one would thereby suffer. But though P thus classifies a very wide range of actions as rationally permissible, it by no means classifies *every* action as rationally permissible. Unless someone, perhaps oneself, will avoid a harm that is at least as significant, or will receive a benefit that is at least as significant, P classifies as irrational all actions that one knows will bring one some harm. Two examples will suffice.

[26] 'Just because one feels like it' is equivalent to 'for no reason.' It is not itself a reason. For, as Derek Parfit points out, when there is a reason to do an action, it is also a reason to want to do it. Thus, 'because it will give me pleasure' is a reason both to eat chocolate, and to want to eat chocolate. But 'just because I feel like it' is *not* a reason to *want* to throw a stone into the woods – it is just a restatement that one does want to. See Parfit (1997), pp. 127–28.

(5) *Suffering harm merely to avoid a less significant harm for oneself*. For example, if one has a bad toothache that promises to get worse without treatment, then P will classify an unnecessary delay in a trip to the dentist as irrational.
(6) *Suffering harm merely to get a benefit that is not equally significant*. For example, if one is beginning to have respiratory problems, then P will classify smoking as irrational, even if one gets pleasure from smoking.

Many philosophical theories of rationality cannot class all of 1 through 4 as rationally permissible.[27] And theories that can classify them as permissible often only are able to do so at the cost of wrongly classifying 5 and 6 as permissible also.[28] But common sense classifies all of these cases as P does, and this counts in favor of P. P classifies most of our everyday activities as rational, whether they are selfish, goodhearted, spontaneous, or carefully deliberated. And yet P does not do this simply by placing no limits on the object of rational desire. P classifies much addictive, compulsive, and phobic behavior as irrational, as well as self-destructive actions done out of rage, stupidity, lust, and so on. And even if an agent is acting on her coherent, informed, and considered preferences, P is still able to classify her action as irrational, if those preferences are self-destructive and do not involve compensating benefits for anyone else.

How does P manage to classify actions as it does? It does it by implicitly defining two logically distinct roles for normative reasons: justifying and requiring, which were more informally introduced in the previous chapter. We can define these roles more formally as follows:

[27] For example, the Thomas Nagel of Nagel (1970) would have difficulties classifying 1 as rationally permissible. So would any other theorist who does not allow that there is an *essential* distinction between reasons that involve one's own interests and reasons that involve someone else's. 'Essential' here means 'depending on more than one's *generally* knowing the content of one's preferences better than those of other people, one's *contingent* desire to do things for oneself, the *likelihood* of things going wrong if one tried to satisfy other people's desires, etc.'

[28] Bernard Williams would allow that 1 through 4 might be rationally permissible, depending upon the agent's subjective motivational set. But then he is also committed to allowing the same of 5 and 6, for people with unusual motivational makeups. Hume, of course, is in the same position. The same may appear to be true of Scanlon, for the rationality of action for Scanlon depends only on whether or not the agent acts in accord with, or against, that agent's own (possibly quite drastically wrong) judgment regarding how he ought to act. In some moods Korsgaard is very close to Williams and Hume here, allowing that the rationality of one's actions is totally dependent upon one's contingent practical identity, even if this happens to be the identity of a wanton or a Mafioso. See Korsgaard (1996a), pp. 256–57. But she also holds (p. 99 n. 8) that it is impossible for humans to act without a reason, so that 2 is not a possibility.

> The *requiring* role: the role of making it rationally required to do (i.e., to make it irrational to fail to do) actions to which it is relevant.
>
> The *justifying* role: the role of making it rationally permissible to do actions that would otherwise be irrational.

According to P, the only reasons that play a requiring role – the only reasons that are prima facie rational requirements – are those that involve avoiding harm to the agent. That is, it is only the presence of such a reason that can rule out an action as irrational. But all reasons can play the *justifying* role. That is, if an action would otherwise be irrational (in virtue of the fact that it will harm the agent) it can be made rationally permissible by the fact that it will avoid at least an equally significant harm to *anyone*, or will provide at least an equally significant benefit to *anyone*.

Korsgaard, along with most other philosophers, does not even see the possibility that some reasons may serve only to justify, and not to require. If there are reasons that serve only to justify, and not to require, then one is *never* rationally *required* to act on them. Such reasons are formally similar to excuses or canceling conditions in moral theory, except that they also provide common (but not universal) human *motives*. As will be discussed in chapter 4, this link to common human motives is part of the reason why it makes sense to offer them as reasons when one in engaging in practical argument. But, in any given instance, one can be fully rational and have no tendency to act on such a reason – no motivation. The existence of such reasons therefore is incompatible with the internalism requirement.

At this point no argument is being offered that P is indeed a principle of practical rationality. All that is being urged is that it might be. And if it might be, then it is logically possible that some reasons serve only to justify, and not to require. Given the logical possibility of such reasons – reasons that would falsify the internalism requirement – why does Korsgaard hold that all reasons are prima facie rational requirements? One explanation is that Korsgaard adheres to a view of causation as involving universal, non-stochastic laws of force, and upon the Kantian view of the will as a rational causality that has this same universal, law-following character.[29] That is, just as a force impressed upon an object will cause it to move in the line of force, unless there are other forces at work, so too (on Korsgaard's view) will a *rational agent* act as any given reason directs, unless there are other

[29] Korsgaard (1996a), pp. 225–28.

reasons that also bear on the choice. Thus, the following is Korsgaard's view about the *content* of the claim that a certain type of consideration, *x*, provides an agent *A* with a reason: such a claim always says (among other things) '*A* is prima facie *rationally required* to act on considerations of type *x*.' A logical consequence of this view of the *content* of rational principles is the *motivational* view that, in a rational agent, a reason always supplies some motivation. And this just is the internalism requirement. It is a logical consequence of the requirement view because if, as the requirement view holds, reasons are prima facie rational requirements, then a *rational* agent *must always act* on a reason when no countervailing reasons are present.[30] When no opposing rational motivation is present, or when such opposing motivation is removed, a rational agent cannot help but act as the unopposed prima facie requirement requires: that is what it is for it to be a prima facie requirement. Thus, in a rational agent there is by definition a disposition to act on any reason she has, when countervailing reasons are weakened or removed. This disposition to act constitutes the motivation that, in a rational agent, the internalism requirement states must exist.

As Korsgaard has shown, Hume's and Williams's motivational view that rational principles cannot motivate on their own is only as plausible as their antecedent claims about the content of those principles. So too is the internalism requirement only as plausible as the antecedent claim that reasons always specify prima facie rational requirements. How plausible is this view? This chapter has of course not shown that it is false, just as Korsgaard's argument did not show that Williams's claims, or Hume's, were false. As far as Korgaard's argument goes, Williams may still be right that there are no substantive principles of rationality that can be shown, by argument, to be valid. And it is equally true that, as far as the argument of this chapter goes, Korsgaard might be right that all reasons are prima facie rational requirements. That is, for all that has been said in this chapter, she might be right that all reasons are ones on which we *must* act, in the absence of countervailing rational considerations, on pain of acting irrationally. But Williams cannot use the claim that it is impossible for rational principles to provide motivation in support of his claim that there are no valid substantive rational principles. And in the same way, neither Korsgaard, nor anyone else, can use the internalism requirement to argue that all reasons are prima facie rational requirements. For the internalism requirement is a *consequence*

[30] For a similar argument in favor of the internalism requirement, see Tilley (1997), p. 113.

of the requirement view. And so neither Korsgaard, nor anyone else, can use the internalism requirement to argue against the existence of reasons that have only a justificatory role, or against principles of rationality that have the same logical form as principle P. That is, they cannot use the internalism requirement to argue against the conclusion of chapter 2.

The argument between those who accept the internalism requirement and those who reject it must concern itself with the question of whether there are purely justificatory reasons or not. According to principle P, or to any other principle of a similar formal structure, there are such reasons. It is up to the adherents of the internalism requirement to explain why the principles of rationality cannot take such a form. Although it is not the primary purpose of this chapter to descend into these particular arguments, it may be useful to flag one objection to the possibility of purely justificatory reasons at this point, if only to defer a full answer until later. This objection claims (a) that if a putative justifying reason does not find some corresponding motivation in an agent – if, for example, the agent does not *care* that her action will benefit some third party – then it cannot rationally justify that agent in making any sacrifice. So any justifying reason will in fact find a corresponding motivation in the agent. The objection then goes on to assert (b) that if the reason does have such a corresponding motivation – if the agent *does* care – then the reason provides a prima facie rational requirement: after all, isn't it irrational to act against one's preferences without some countervailing reason?[31] Thus, the objection seems to show that any justifying reason will be a requiring reason as well.

It is important to flag this objection here because it makes use of an interesting kind of situation: a situation in which even the stipulation of full information may not bring the objective and subjective rationality of an action into agreement, because the agent is not acting *for the reasons that make the action objectively rational*. In such situations there is a risk of error if we take our intuitions about subjective rationality to give us direct insight into objective rationality and, therefore, into the nature of reasons.[32] Both of the premises in the above objection fall into error in this way. First, (b) makes use of the idea that it is irrational to act against one's own desires. But such action is – at worst – only *subjectively* irrational. As chapters 4, 7, and 8 will argue, this may show very little about the reasons that favor

[31] For purposes of this objection, we can treat caring about, desiring, and preferring as relevantly similar.

[32] Or if, like Korsgaard, we take subjective rationality to be the end of the normative road.

or oppose such action, or whether or not the agent ought to perform it. It is easy to see this if one imagines a person with the irrational desire to smash his own head open.[33] And (a) is problematic for similar reasons. For there is certainly something plausible in the claim that someone who willingly performs an action that he knows will bring him a significant harm would not be justified in such an action, even if it had very significant foreseeable benefits for others, if that agent regarded those benefits with indifference or antipathy. As chapter 7 will explain, such an action may well be subjectively irrational, and it may be so *because* the agent lacked a certain motive. But that is not enough to rob the altruistic reason of justificatory power, or to justify calling the motive 'a reason' when it is present. For, as chapter 4 will argue, and as chapter 7 will continue to explain, it is to objective rationality that reasons are *directly* relevant.

Moreover, even when we are considering subjective rationality, it is important to note that the claims made in support of the plausibility of (a) are only appropriate for an action that has *actually* been performed. But our judgments of even the subjective rationality of actions have a much wider domain than this. We talk about what *would have been* rational to have done in the past, and what *would be* rational to do in the future. And when we discuss these actions, our descriptions of them generally do not include the motives that produce them. This is why we can talk about the *same* action being performed from *different* motives. Our judgments of the rationality of actions that were not (or have not yet been) performed are judgments of whether they would have been (or would be) rational *if they were chosen for the relevant reasons*. This is why it is plausible to say of essentially anyone, regardless of their current spiteful contempt for humanity, that it would be rational – in the sense of rationally permissible – for them to do the morally obligatory thing. For it is plausible that there are always sufficient agent-neutral *justifying* practical reasons in favor of morally obligatory action.[34] These reasons rationally *justify* such action, even if they cannot rationally *require* it. Any agent, no matter how consistently selfish and mean he has been in the past, might, for all we can ever know, choose to act on those rationally justifying reasons. Those who argue for desire-satisfaction views by using examples involving altruistic or misanthropic agents, or agents who are stipulated to have some other motivational setup,

[33] It is useless to object that it is *rational* desires that provide reasons, since it is the presence or absence of reasons *behind* a desire that determine whether or not the desire is rational.

[34] At least, it is hard to see how morality could require one to make a sacrifice if *no one at all* was going to be benefited or spared a harm.

are relying on a picture of a fixed set of desires or preferences that determine our choices. This picture is a psychological fiction of the tempting sort that Wittgenstein, for example, consistently challenges. For *even if there were such fixed sets of desires*, we would have no way of knowing enough about them for that knowledge to be what is behind our teaching of the concepts 'rational,' 'reason,' and so on. Rather, the teaching of these concepts, and therefore the relevant meanings of the corresponding words, is more plausibly taken to be based on the foreseeable consequences of actions.

Those who persist in holding that, despite a certain unavoidable ignorance of their precise content, we do indeed have such fixed sets of desires, may be tempted to wield some version of the 'ought-implies-can' principle against the idea that there are desire-independent purely justificatory reasons. But the existence of such reasons would not violate even the strongest 'ought-implies-can' principle. For there is no 'ought' in play when the question is only one of the rational *permissibility* to which justificatory reasons are relevant. But worse than this, for such an objector, is the following: the necessary form of 'ought-implies-can' is not available in the domain of rationality. When 'ought to ϕ' means, roughly, 'is rationally required to ϕ,' it simply cannot plausibly be taken to imply 'is psychologically able to ϕ.' Some people suffer from mental illnesses (addictions, phobias, etc.) that essentially *compel* them to do things that they rationally ought not do, or *prohibit* them from doing things that they rationally ought to do: their irrational actions are precisely the result of the fact that they are psychologically unable to act otherwise. Perhaps such mental illnesses erase the *moral* 'ought,' but they leave the rational 'ought' intact.

CONCLUSION

This chapter defended the thesis that there may be such things as purely justificatory reasons, and the consequent thesis that the internalism requirement on reasons may be false. The first part of the argument was formally the same as Korsgaard's own argument in "Skepticism about Practical Reason." Both Korsgaard and I conclude that when one is trying to justify a principle of reason, whether that principle looks like the categorical imperative or like principle P, no real work can be done merely by citing some motivational requirement. In her own argument, the motivational requirement shown to be insufficient was motivational skepticism: the view that principles of practical reason cannot motivate on their own. In

this chapter, the motivational requirement shown to be insufficient was the internalism requirement on reasons. The first part of the argument thus cleared the way for an unprejudiced examination of a putative rational principle, P, that specifies some reasons that have no necessary power to motivate even rational agents. These reasons have only *justificatory* power, and never *require* action. The existence of such reasons is not to be denied simply by pointing out that they are incompatible with the internalism requirement. That is, one cannot simply insist that a view of rationality must be false if it implies the existence of normative reasons that need not motivate a rational agent to whose choice of action it applies. The possibility of principles of rationality formally similar to principle P must be ruled out on other grounds.

As in the previous chapter, there is room to accommodate those who do not wish to go so far as to embrace *purely* justificatory reasons. Some may have the intuition that extremely cruel behavior, even when it involves no risk of harm to the agent (which could only happen under rather peculiar circumstances, for it is very hard to be certain that one's extremely cruel behavior will not open one up to revenge or punishment), is irrational *in virtue of the altruistic reasons against it*. An adherent to such a view can still agree that principle P correctly classifies actions 1 through 6 above. For a principle similar to P – call it Q – according to which altruistic reasons have *relatively little* requiring power, would also classify those actions as P does. It is true that Q would not imply the existence of *purely* justificatory reasons. According to Q *all* reasons would have *some* power to require. Because of this, all reasons would motivate a rational agent, if they were unopposed by countervailing reasons. So a version of the internalism requirement would remain true. But this will be of very little consolation to those who wish to use the internalism requirement against the central theses offered in this book. For even according to Q it will remain true that we will need to distinguish the justifying and requiring roles of reasons when we are determining the rational status of particular actions. And it will remain true (as examples 1, 3, and 4 above illustrate) that a rational agent need not be motivated to act on the strongest of two opposed reasons, whether we take 'strongest' to mean 'strongest in the justifying role' or 'strongest in the requiring role.'

4

A functional role analysis of reasons

In the previous two chapters a great deal of attention was paid to the notion of a practical reason. In chapter 2, the justifying and requiring roles of these sorts of reasons were distinguished, and in chapter 3 this distinction was further developed. But even in chapter 2, one might have noticed that the roles of justifying and requiring were explained in terms of an antecedent notion of rationality. For example, the justifying role is the role of making an action rational, when otherwise it would be irrational. Unless we have some idea what 'rational' and 'irrational' mean, this claim will make little sense. What emerged more or less explicitly in chapter 3 was that the need to distinguish between these two roles for practical reasons depends in an important way on what principles of rationality we are willing to recognize. Principles that have the form of P imply that some reasons can play a justifying role without being able to play a requiring role at all. And principles similar to Q allow that some reasons might be able play a very significant justifying role without being able to play much of a requiring role. So, perhaps despite appearances, the basic normative notion that has been used so far in this book has *not* been the notion of a reason for action. Rather, it has been the notion of an action's wholesale rational status – although, since the notions of subjective and objective rationality are so closely related, it may not yet be entirely clear which of these is most basic in explaining the justifying and requiring roles of reasons.

It is of course possible to begin an analysis of normativity by explicitly taking the notion of 'a reason' as basic, rather than the notion of an action's wholesale rational status. This is Thomas Scanlon's strategy.[1] But if one does this, then although one can say that reasons 'count in favor' of things, when one asks Scanlon's question '*How* do reasons count in favor of actions?' the only response available is his disappointing 'as a

[1] Scanlon (1998), p. 17.

reason.'[2] This is the result of denying that there is any normative concept more basic than that of a reason, and of defining 'reasonable,' 'rational,' and so on, *in terms of* reasons. This chapter argues that this is a crucial methodological mistake. Rather, for purposes of philosophical analysis, *wholesale rational status* should be taken as more basic than reasons. To make a rough analogy, the strategy of taking reasons to be the basic units of normativity is a mistake of the same sort as taking individual words to be the basic units of meaning in language. Just as it is more profitable to take the *sentence* as the basic unit of meaning, it is more profitable to take wholesale rational status as the basic unit of normative assessment. And just as we can identify the various meanings of individual words with the systematic contributions that they make to the meanings of sentences, we can understand the various normative roles of reasons by understanding how they contribute systematically to wholesale rational status. This sort of analysis of normative reasons in terms of wholesale rational status allows for much more interesting and informative answers to the question 'How do reasons count in favor of actions?' than reason-based accounts can offer. Such an analysis also provides insight into features of reasons that go beyond their bare normativity.

TAKING THE NOTION OF 'A REASON' AS BASIC

How is a normative reason relevant to an action? If one takes the notion of 'a reason' as basic – inexplicable in terms of more basic normative notions – then it is very tempting to answer 'by counting in favor of it, or against it, to a greater or lesser degree.' If one does take this as an answer, the question of importance when deciding what to do will be 'What are the reasons relevant to my choices, and what are their relative strengths?' And when one takes this as the important question, then one's basic method for assessing how one ought to act will almost certainly be either a maximizing or a satisficing one.[3] That is, one will take it as a truism that one should always act in the way that the relevant reasons support to the highest, or

[2] While Scanlon's fundamental reliance on the notion of 'a reason' is by no means idiosyncratic, he has the virtue of seeing, and admitting, this implication. See also Skorupski (1999), pp. 436–37 for a similarly clear reliance on the notion of 'a reason' as the basis of an analysis of the normative.

[3] One way of avoiding this result is by claiming that the strengths of reasons are often incommensurable. Joseph Raz embraces such a view, which I discuss and criticize in chapter 5. Scanlon himself, despite taking reasons as basic, seems to want to resist a maximizing view by claiming that the ways in which reasons combine are very complex. See Scanlon (1998), p. 32. But this means that his 'counting in favor' relation is also extremely complex. Unless

to a sufficiently high, degree. Stated in the abstract, without the resistance provided by concrete examples, this sort of account of objective rationality (or advisability, or all-things-considered oughtness, or reasonableness, or whatever one wants to call one's most basic normative domain) is very plausible-sounding and attractive. Indeed, if one does take the notion of 'a reason' as normatively basic, it becomes almost inconceivable that reasons could function in any other way.[4]

But there are problems with such accounts. Consider: what are the reasons that favor a two hundred dollar donation to public radio? Let us suppose that the primary reasons are that by doing so, roughly a hundred people will receive some pleasure and edification that they would not otherwise get.[5] And what are the primary reasons in favor of a two hundred dollar donation to UNICEF? Let us suppose that the primary reasons are that such a donation would save three or four children from serious illness and possible death. Now, according to reason-based accounts, the rational status of an action is a function of the reasons that apply to it. And it is to be assumed that the same reasons can reappear in different contexts, retaining their essential normative capacities. Thus, it makes sense to ask whether it would be irrational to risk, say, the loss of one's arm in service of the same reasons that favor donating two hundred dollars to public radio (if the arm-risking method were the only available one). And the answer to this question is 'Yes, it would be irrational to take such a risk for these reasons.' But it would not be irrational to risk the loss of one's arm in service of the same reasons that favor donating two hundred dollars to famine relief (again, if that were the only way to produce the same benefits). Thus there is an obvious and intuitive sense in which giving ten minutes of pleasure and edification to a few hundred people provides reasons that are clearly weaker than the reasons provided by saving three or four people from serious illness and possible death. This sense is that the latter reasons *would make it rationally permissible to act against some reasons that the former would not*.

he explains this relation, however, he must be viewed as simply relying on a heterogeneous collection of particular normative verdicts about wholesale rational status, and the theoretical appeal he makes to the 'basic' notion of 'a reason' must be regarded as illusory.

[4] For one self-proclaimed expression of this inability, see Kagan (1989), pp. 378–80.

[5] The specific details of these examples are not important. If the reader disagrees with the assessments of the relevant reasons, any other example may be substituted in which the reasons in favor of one action are primarily pleasure and edification for a few people, while the reasons in favor of another action are primarily the prevention of significant suffering and death for a few people.

But if there are clearly stronger reasons in favor of donating to charity than in favor of donating to public radio, wouldn't it be *irrational* to choose to donate to public radio over donating to charity?[6] When one takes the notion of a reason as basic, so that reasons can only contribute to the rationality of action in one way ('as a reason'), then one is almost forced into this counterintuitive assessment of the rational status of the options. One initially attractive escape is to move from a maximizing conception of rationality to a satisficing one. That is, one could hold that as long as no alternative action is a favored by reasons that are *much stronger*, then one is rationally permitted to do as one pleases, and that this is why it is rationally permissible to donate to public radio, despite the stronger reason in favor of donating to charity. It would be hard to provide a formal argument against this sort of claim.[7] Rather, one must produce and motivate a better overall theory, as this book is attempting to do. But in any case, the move to satisficing seems unlikely to save our intuitions in the present case. For, when one bothers to list them, the reasons in favor of donating to charity seem *very considerably* stronger than the reasons in favor of donating to public radio. And yet, prior to the corrupting influence of an overly simple philosophical theory, our intuitions are that both donations are rationally permissible.

A FUNCTIONAL ROLE ANALYSIS OF REASONS

In the preceding section, the reasons that favored donating to public radio were compared in strength with the reasons that favored donating to charity. The reasons that favored donating to charity were judged stronger in the following sense: they *would make it rationally permissible to act against some reasons that the reasons in favor of donating to public radio would not*. That is, it would be rationally permissible to risk, say, imprisonment, if that were the only way to save three or four people from the evils of severe malnutrition. But it would be irrational to take such a risk only to fund ten minutes

[6] Of course, for neo-Humeans who advocate desire-dependent views of rationality there is a simple answer to this question: 'No, it isn't irrational; it all depends on what you care about.' This objection has already been dealt with in chapter 3, but for additional powerful arguments against desire-dependent views of rationality, see Quinn (1995), p. 195; Dancy (2000), pp. 35–38; Raz (1999b), pp. 50–64; Scanlon (1998), pp. 35–42.

[7] Although chapter 5 contains a formal argument against the idea that the rational permissibility of an option is the result of its being within a certain 'margin of practical indifference,' in terms of the strengths of the reasons that favor it, of the best option.

of public radio. In light of these comparisons, we can offer the following criterion of strength for reasons.

C1 Given two reasons, R_1 and R_2, R_1 is stronger than R_2 iff (= if and only if):
(i) R_1 would make it rationally permissible to do anything that R_2 would make it rationally permissible to do.
(ii) R_1 would make it rationally permissible do some things that R_2 would *not* make it rationally permissible to do.[8]

'To make it rationally permissible' means 'to make it rationally permissible, when without the reason it would *not* be rationally permissible.' Here is an illustrative example. If R_1 would make it rationally permissible to suffer the loss of one's finger, but R_2 would only make it rationally permissible to suffer a day's mild nausea, then we can say that according to C1, R_1 is stronger than R_2.[9] Given this analysis, not only can we give sense to the notion of one reason being stronger than another, but we also have one informative answer to Scanlon's question 'How do reasons count in favor of actions?' This answer is 'By being able to make it rationally permissible to perform actions, in cases in which, without them, it would be irrational.'[10] This is an analysis of normative reasons that picks out a functional role that reasons play, relative to the wholesale rational status of actions. So we have the following functional role analysis of normative reasons.

FA1 A consideration is a reason if it can make it rationally permissible to perform actions that would be irrational without it.

[8] It is true that, if C1 is to preserve the general transitivity of 'stronger than' among reasons, we must deny the following possibility: that there could be two reasons, R_1 and R_2, such that R_1 could make it rationally permissible to suffer H_1 but not to suffer H_2 while R_2 could make it rationally permissible to suffer H_2 but not to suffer H_1. This seems a benign assumption. At least, if we deny it, then we must deny that reasons can be generally compared in strength. Of course, a number of philosophers hold exactly this view, claiming that reasons may sometimes be incommensurable. I argue in chapter 5 that appeals to incommensurability are the result of a failure to distinguish the two distinct kinds of normative strength explained here.

[9] Again, we assume here that R_1 would *also* make it rationally permissible to suffer a day's mild nausea. See note 8 above.

[10] In fact, a more accurate answer would be: 'By counting towards their rational permissibility by either: (1) actually making them rationally permissible, (2) reducing the number of additional reasons it would take to make them rationally permissible, or (3) making it so that they would continue to be rationally permissible, even if there were more reasons against them, when, without those supporting reasons, they would become irrational.' This more involved answer is required by the *pro tanto* or prima facie nature of reasons.

In previous chapters we have been calling the role given by FA1 'justifying,' for when actions would be irrational without some reason, then those actions stand in need of *justification*, and it is reasons that provide this justification. Of course, if FA1 is to be useful in identifying justifying reasons, it presupposes that we have some way of determining the wholesale rational status of actions. But we should not let this worry us at this point. The question at issue is whether it is better to take wholesale rational status as basic, or individual reasons. Whichever choice we make, some story will have to be told: either a story about how we come to know that an action is irrational, or a story about how we come to know that a certain fact provides a reason for an action.

FA1 is of course not the whole of the story. FA1 only gives the positive, justifying role of reasons. But reasons can only play this positive role relative to actions that would otherwise be irrational. And we have already seen instances in which reasons *rule out* actions that otherwise would have been rationally permissible. These are instances in which reasons are playing what we have been calling the 'requiring' role. So we should add the following to our functional role analysis.

> FA2 A consideration is a reason if it can make it irrational to do something that would, without that consideration, be rationally permissible.

FA2 gives sense to the claim that reasons can count *against* actions. It also allows a second and distinct answer to Scanlon's question 'How do reasons count in *favor* of actions?' This second answer is 'By being able to make it irrational to fail to do those actions' or, equivalently, 'By rationally requiring those actions.' For example, it is generally rationally permissible to sit in one's room, reading a book. But if the room is on fire, so that one will burn to death if one continues to read, then it is irrational to fail to get up and leave. Thus, the fact that one can only avoid a fiery death by getting up and leaving counts as a reason in favor of getting up and leaving. But it is a reason that favors getting up in a way that is distinct from making it merely rationally permissible to get up. Rather, this reason makes it rationally *required* to get up: required, on pain of irrationality.

As in the case of FA1, if we have some independent way of determining the rational status of actions, we can use FA2 to determine which considerations are reasons. FA2 also suggests a second criterion of strength for reasons, although it will be strength *in a second sense*, and cannot be assumed to correspond to strength in the first sense.

C2 Given two reasons, R_1 and R_2, R_1 is stronger than R_2 iff:
(i) R_1 would make it irrational to do anything that R_2 would make it irrational to do.
(ii) R_1 would make it irrational do some things that R_2 would *not* make it irrational to do.

For example, that one will suffer some painful scrapes is a reason against action. But this reason is not as strong as the reason provided by the prospect of losing an arm. This follows from C2. For while it would be irrational to reach into a lion's cage to pick up a fallen hundred-dollar bill, it would not be irrational to reach into a prickly holly-bush. Again, we determine the strengths of these two reasons based on wholesale judgments of the rational status of actions. We do not need an *antecedent* notion of reasons being stronger or weaker than one another.

FA1 and FA2 pick out two logically distinct roles for reasons. Thus the final answer to Scanlon's question is disjunctive. That is, reasons count in favor of actions either by being the kind of consideration that can make it irrational to fail to perform them (which is the same as making it rationally required to perform them), or by being the kind of consideration that can make it rationally permissible to perform them, when otherwise it would have been irrational.[11] If one adheres to the idea of reasons as basic, it may initially be tempting to claim that these two functional roles are simply alternate aspects of the same underlying normative property. One might even be tempted to make the stronger logical claim that the proposition that one reason is stronger than another in the justifying role necessarily implies that it is also stronger in the requiring role. But this logical claim is demonstrably false. It is very easy to provide consistent descriptions of cases – and this is all that is required to refute the logical claim – in which two reasons have the same strength in the justifying role but very different strengths in the requiring role. We have already seen many such examples in the previous two chapters. Here is a further reason for holding that those examples were completely coherent. It is plausible that, with regard to justifying strength, it does not matter whose interests are involved in a reason: the agent's, or someone else's. That is, any sacrifice that would be made rationally permissible on account of the benefits it was likely

[11] The words 'kind' and 'can' are used here, because when other reasons are involved, a requiring reason might not *actually* require an action, and a justifying reason might not *actually* justify an action, all-things-considered. Reasons always provide prima facie requirements or justifications.

to produce for the agent would also be rationally permissible if it were undertaken in order to produce those very same benefits for someone else instead. But reasons involving the interests of the agent seem to have significantly more requiring strength than reasons involving the interests of others: it would be irrational to hurt oneself significantly for a few hundred dollars, but hired killers are not irrational; they are immoral.[12] Of course the general normative judgments involved in these claims might be considered controversial. But one need not agree with those judgments in order to see that they are coherent. And their bare coherence demonstrates the logical distinction between justifying and requiring.

Again, it is not that some reasons play a 'merely' justifying role because they are comparatively weak, and that if they were strengthened they might be able to play a requiring role. To see this more clearly, it may be worth turning our attention again to the same distinction as it appears in many moral views. On many moral views, self-preservation can morally justify a very great deal, but cannot morally require anything, while the avoidance of some comparatively small harm to someone other than the agent morally justifies far less, but can nevertheless morally require more than self-preservation would. This shows that, in the moral realm, requiring strength is not merely a higher degree of justifying strength. The same is true for reasons as they are relevant to the *rational* status of actions.

REASONS AND TWO CONCEPTS OF RATIONALITY

Although the functional role account provided above allows for an informative answer to Scanlon's question, so far it remains too crude. FA2 seems to classify certain things as reasons that we do not want to classify that way. For example, John Broome has distinguished between reasons, on the one hand, and what he calls 'normative requirements' on the other.[13] FA2 cannot account for this distinction yet. The distinction is the following.

[12] Again, neo-Humeans might explain this by reference to the differences in the contingent desires of hired killers and normal people, but the argument of this chapter is not directed at those who hold such views. See note 6 above. Against Kantians who wish to assert that hired killers are in fact irrational, one might urge the following. Such a view is the result of combining a correct rejection of desire-dependent views of rationality, with a failure to note the distinction between the justifying and requiring roles of reasons. For if one regards reasons as a matter of desire-independent fact, and realizes that the interests of others provide reasons, and if one can see only the requiring role of reasons, then one is very likely to conclude that immoral behavior is irrational.

[13] Broome (1999), pp. 398–419. It is worth noting that Broome himself sees only one role for normative reasons: requiring. See ch. 3 n. 10.

As Broome rightly notes, it not always true that if one believes that one ought to ϕ, then one actually ought to ϕ. For, of course, one might *wrongly* believe that one ought to ϕ, without actually having any reason to ϕ at all. Simply believing that one ought to ϕ does not, by itself, provide a reason to ϕ. And yet, there does seem to be something irrational in one's believing that one ought to ϕ, and yet not ϕ-ing.[14] The situation is even starker if one wrongly believes that one is *rationally required* to ϕ. If one fails to ϕ while one has this stronger belief, it seems appropriate to say that one is acting irrationally. In Broome's terminology, the belief that one is rationally required to ϕ *normatively requires* that one ϕ. Here is a concrete example of how FA2 will conflate such normative requirements with reasons. Suppose that *without* the belief that she is rationally required to get up early and begin working on an urgent project, it would be rationally permissible for Joanna to sleep in. It might nevertheless be irrational for Joanna to sleep in, if we add to the description of the situation, the fact that she believes that she is rationally required to get up early. Since this belief makes it irrational to do what, without the belief, would be rationally permissible, FA2 will wrongly classify it as a reason against sleeping in.

How can we avoid labeling the belief that one is rationally required to ϕ 'a reason'? First, note that in order to make the example more telling, we might sensibly have stipulated that Joanna *ought* to sleep in, because she is too tired to work productively, and because she is entirely unjustified in her belief that the project is due soon. That we can stipulate that Joanna ought to sleep in, even though her sleeping in would be irrational in a certain sense, suggests that the two senses of rationality – the objective and the subjective – have come apart in a way in which they have not yet come apart in many of our earlier examples, when we were allowing the stipulation of full information. In fact this is right. Objective rationality is relative to the facts of the case, regardless of whether they are known to the agent.[15] It is this sense of rationality that stands behind our claim that Joanna ought to sleep in. Subjective rationality has a more intimate connection with freedom of the will, mental illness, moral responsibility, and so on. When one calls an action subjectively irrational, one is committed to the claim that something has gone wrong in the practical mental functioning

[14] Indeed, this seems to be the only type of action Thomas Scanlon is willing to call 'irrational.' See Scanlon (1998), pp. 25–27.

[15] See Brandt (1979), pp. 72–73; Gibbard (1990), pp. 18–19; Raz (1999a), p. 22. See also Cullity and Gaut (1997), p. 2.

of the agent. But to call an action irrational in this sense is consistent with the claim that there are facts, unknown to the agent, that support the claim that, on the whole, she ought do the action. Reasons are directly relevant to this latter sort of claim, and are only indirectly relevant to claims about subjective rationality. This is why objective rationality is sometimes referred to as 'what we have most reason to do.'[16] The kind of irrationality involved in Joanna's sleeping in, when she believes that she is rationally required to get up, is subjective rationality. The plausibility that reasons are directly relevant to what we ought to do, whether or not we know of those reasons, and the fact that FA2 wrongly classifies the belief that one ought to ϕ as a reason to ϕ if we take the relevant sense of rationality to be subjective, both argue that the sense of rationality that figures in FA1 and FA2 should be taken to be objective rationality. Since Joanna's belief that she is rationally required to get up does not have any impact on the objective rationality of her sleeping in (or getting up), it is not a reason.

A second and related reason why we should not take Joanna's belief as a reason to get up is that it appears to be unnaturally strong, if it is taken as a reason. This is because, *no matter what other reasons in favor of an action there might be*, it seems irrational to do the action if one believes that the action is irrational. It is as if the putative reason simply couldn't be outweighed by other reasons, no matter how strong they were. But that is very odd. It suggests that one's belief that a certain action is irrational is a stronger reason against the action than is the fact that it will save twenty people from dying horrible fiery deaths. The putative reason here does not seem to lend itself to any sort of weighing or balancing. This second reason why we should not regard Joanna's belief as a reason is related to the first in the following way. The reason why her belief makes it irrational to sleep in, regardless of the 'opposing' reasons for sleeping in, is that there are again two senses of rationality in play. The reasons in favor of sleeping in are relevant to the objective rationality of Joanna's action, and they support the claim that Joanna *ought* to sleep in. Joanna's (false) belief that it would be irrational to sleep in does not make it irrational to sleep in by *outweighing*

[16] See Parfit (1997) and Scanlon (1998), p. 30. This characterization correctly reflects the primary relevance of reasons to objective rationality. But the word 'most' smuggles in the assumption that reasons play only one normative role, and has the unfortunate consequence that the relevant reasons typically pick out one action as uniquely favored. Joseph Raz argues persuasively against this view in "Explaining Normativity: Reason and the Will" in Raz (1999b), pp. 100–2.

these reasons. Rather, it makes it irrational in a different sense altogether: it makes it subjectively irrational.[17]

So far, we have seen a consideration that FA2 would wrongly classify as a reason if the wrong sense of 'irrational' is being used. That is, Joanna's belief would be classified as a reason by FA2 because it turns an otherwise subjectively rationally permissible action into a subjectively irrational one. Are there also considerations that FA1 would wrongly classify if we understand 'rational' in the subjective sense? That is, are there considerations that can change an otherwise subjectively irrational action into a subjectively rationally permissible one, but that we would not want to classify as reasons? There are. Consider ignorance, as it functions in the following example. Suppose we see our friend Bob about to remove a wasps' nest from the corner of his garage with his bare hands. We ask why he is doing that, and he answers that it is ugly, and he wants the garage to look nicer. If Bob knows what we all generally know about wasps, this is a subjectively irrational action: it shows that something has gone wrong with Bob's mental functioning. But if we add to our description of Bob's action, the fact that he is completely, and (somehow) excusably, ignorant about wasps and wasps' nests, then it may be subjectively rationally permissible (but extremely unfortunate) for him. Thus, ignorance turns a subjectively irrational action into a subjectively rationally permissible one. If FA1 is understood in terms of subjective rational status, ignorance will turn out to be a reason. But ignorance is not a reason in favor of an action.

REASONS AND MOTIVES

There is another reason why we would not want to call ignorance a reason for action. It does not seem that anyone could ever act *for* this reason. Here we find the grain of truth in Bernard Williams's explanatory requirement on practical reasons.[18] Williams holds that if a consideration is a normative reason, then it must be that people sometimes act for that reason. Unfortunately, he takes this to mean that it must be psychologically possible for *any agent* who has a reason to act on it, merely by going through some broadly instrumental rational processes. Because of this, Williams is committed to

[17] See Stampe (1987), p. 344 for an argument that desires provide reasons that is based entirely on this error. Stampe even notes the 'extraordinary authority' of desire as a reason in this connection, but does not see this as a sign that something has gone wrong.

[18] See Williams (1981). For a more moderate view, compatible with the position offered in this chapter, see Raz (1999b), pp. 100–2.

the view that if an agent (perhaps because of a severe chemical depression) simply has no desires that would motivate him to take some medicine that would cure him, and if this

> is not the product of false belief; and he could not reach any such motive from motives he has by the kind of deliberative processes we have discussed; then I think we do have to say that . . . he indeed has no reason to pursue these things.[19]

Many people have thought that this is too strong a conclusion to draw from the fact that normative reasons are the kinds of things that are often cited in explanations of action.[20] But it does seem true that unless a consideration is the *kind of thing that people sometimes act for*, then it is not a reason. People do act in order to get pleasure, to avoid pain, to help other people get or avoid these things, etc., and these things provide normative reasons. But ignorance simply cannot play this role. It is true that ignorance can be cited in explanations, but people cannot act *for* ignorance in the way they can act *for* reasons. When people act for reasons, then those reasons are their motives. Current ignorance is in the wrong category to be a motive for anyone.[21]

WEIGHING REASONS

A final formal condition on reasons is more complicated. It takes its cue from a remark of John Broome's, that "weighing is just what reasons are made for."[22] Briefly, the condition is that the systematic contribution that a reason makes to the rational status of action must lend itself to representation in terms of strength values.[23] In order to explain what this amounts

[19] Williams (1981), p. 105. See also Johnson (1999). For my own interpretation of the explanatory requirement, see J. Gert (2002b).

[20] See, e.g., Heath (1997), p. 454; Parfit (1997), pp. 111–14.

[21] A person could act *in order to become or remain ignorant* – perhaps of some painful fact. But in such a case it is more plausible to say that the real reason for her action was to avoid pain.

[22] Broome (1999), p. 412.

[23] Joseph Raz (1999a), p. 43 makes this more explicit than Broome, writing that "all reasons are comparable with regard to strength . . . and that this is their only feature relevant to the outcome of practical inferences." Unfortunately, Raz assumes that reasons play only one normative role, and therefore have only one strength value. As a result of this assumption, and in order to capture the full range of rationally permissible action, he is forced to invent the notion of 'exclusionary permissions,' which have the effect of allowing, but not requiring, one to omit certain reasons from one's calculations of what to do. This device approximates the effect of some reasons having more strength in the justificatory role than in the requiring role. Raz's suggestion is discussed at greater length in chapter 5.

to, it will be useful to examine another way in which a consideration might contribute to the normative status of an action, without being a *reason* for it in this sense. Because the rationality that figures in FA1 and FA2 is very plausibly exclusively a matter of the reasons relevant to an action, we will have to look elsewhere for clear examples. Morality will provide fertile ground.

First consider a simple act-utilitarian account of morality according to which the sole good is pleasure, and the sole evil pain. On such a view, the way to determine the moral status of an action is the following. One surveys all the possibilities, and isolates the relevant consequences, which consist only in increases and decreases in pleasure and pain. One then calculates some utility score for each option. The morally correct choice is the option with the highest score. In making the relevant calculations, each increase or decrease in pain or pleasure makes a constant contribution in the calculation of the total utility score for a possible choice. That is, if someone will suffer a bitter disappointment as a result of the agent's choosing option A, this counts against A in exactly the same way that it would count against B, if that same disappointment were to be a result of choosing B. Thus, we can say that the pain of the disappointment makes a constant contribution to the moral status of any given option. Because of this it makes sense to say that the disappointment of the person is a moral reason against an option (even if it is outweighed by other moral reasons). And thus it also makes sense to call such a moral view a reasons-based morality, and to say that moral reasons are provided by increases and decreases of pleasure and pain for the people affected by a possible action. For what one does, in order to decide which option to choose, is to list the reasons for and against each possible option, and see which option is favored by the balance of these reasons.

Now consider a certain sort of rule-utilitarian account of morality, again according to which pleasure is the sole good, and pain the sole evil. This account holds that morality can be understood as given by a group of rules and severities of punishment that have the following joint feature: if people knew that violations of these rules were liable to the specified levels of punishment, then the consequences would be better (in terms of total net pleasure and pain) than with any other set of rules and punishments. What kinds of things might count as moral reasons on such a view? There are two obvious options. The first is increases and decreases in pleasure and pain: after all, the view is still a utilitarian one. The second option is

the fact that an action breaks one of the rules of the system: for example, that it is an act of deception.

Are either, or both, of these kinds of considerations moral reasons, within the framework of the rule-utilitarian morality we are considering? It is very tempting to answer 'Yes, of course.' Nor is that answer wrong, in a sense. One reason why it is tempting to answer 'Yes' is that both the fact that an action will hurt someone and the fact that it will deceive someone, are considerations that fulfill the moral analogue of FA2. That is, they are considerations that can change a *morally permissible* action into an *immoral* one. Moreover, these considerations also fulfill the motivational condition on reasons. Namely, moral agents are commonly motivated to avoid hurting or deceiving people: they often act *for* these reasons.

But is it possible to determine the moral status of an action by listing these reasons for and against the available options, taking their strength values into account, and determining which actions come out with acceptable comparative scores? Certainly one cannot do this if one takes increases in pleasure and pain as the sole reasons. Indeed, this is part of the advantage that rule-utilitarian views have over act-utilitarian views: there are actions that actually have, overall, the best consequences, but that we nevertheless regard as highly immoral. To take one standard example, it may be that the surreptitious killing of a selfish patient who came to the hospital for a wart removal could save the lives of five other extremely benevolent people. And it is well known that the same sorts of problems arise even if one takes the breaking of a moral rule as a reason with constant weight. For when one does this, it becomes morally permissible to deceive one person in order to prevent five other people engaging in similar acts of deception, or to kill an innocent person to prevent five other such murders.

But couldn't we regard such a view as reason-based in any case, by regarding moral reasons as having *variable* or *context-dependent* weights? Although it is tempting to say so, the answer is 'No.' The problem is that one could not know these variable weights until one had already completed some other procedure that yielded the wholesale moral status of the action.[24] These variable weights could then be *read back into* the

[24] See Philips (1987), pp. 367–75. Philips argues in a similar way against what he calls 'the constancy assumption' in moral theory: the assumption that the weight of a moral reason is constant and does not vary from context to context. His argument is similar to the one offered here, in that he asserts the conceptual priority of moral *principles* over that of

consequences, but they would then obviously be of little additional use in determining the moral status of the action. On the rule-utilitarian moral theory being considered here, the determination of the moral status of an action is not a matter of weighing moral reasons for and against it. This should come as no surprise. Rule-utilitarianism is a rule-based, and not a reasons-based account of the moral status of particular actions. And this is true despite the fact that in determining which *rules* are best, we may well assume a constant weight for the reasons for and against them. That is, arguments about the rules may be reason-based, but arguments about the status of particular actions will be in terms of the rules themselves, and not the reasons that support the rules.

Now we can state the final formal condition on a consideration's being a reason: it must be possible to characterize the consideration's normative significance in a way that does not rely upon a prior determination of the wholesale status *of the particular action* to which it is relevant. For reasons are supposed to be useful in determining that very status. The simplest way in which a reason can do this is by having one constant value (its strength), which either counts in favor or against any action to which it is relevant. The view offered in this book is that two values suffice: the strength of the reason in its justifying role, and the strength of the reason in its requiring role. But the point here is only that there is a philosophically important sense of 'reason' according to which considerations are reasons only if the way they contribute to the status of action is sufficiently systematic that one can use the reasons to determine that status. On the rule-utilitarian view described above, the fact that an action is deceptive is not a moral *reason* in this sense, though it is of course of moral significance. It is not a reason, in this sense, because (again, on the rule-utilitarian view we are assuming) there is no systematic way in which acts of deception, even of the same severity and about the same subject matter, contribute to the moral status of actions. Rather, in order to see that the deception is relatively important (or unimportant) one first has to determine the wholesale moral status of the action.

There is another reason why it is tempting to count the fact that an action will deceive someone as a moral reason against the action. Or, if one likes, there is a *sense* in which the fact that an action will deceive

moral *reasons*. But where Philips goes slightly wrong is in thinking that one can calculate the variations in the weights in moral reasons, based on teleological considerations to which the context is relevant. In fact, Philip's 'teleological considerations' are just rule-consequentialist calculations of wholesale moral status.

someone *is* a moral reason. For it is perfectly correct to say of a particular action that *the reason why it is immoral* is that it involves deception. This is an explanatory sense of 'reason.' Its use indicates that the deception will figure in an explanation of why the action counts as immoral. On the rule-utilitarian view outlined above, this means that the deception would be part of an explanation of why it would be bad if that sort of behavior were not liable to punishment. But whether or not this justifies counting the fact that an action will deceive someone as a moral reason *in some sense*, it should be clear that it is a reason in a sense that does not fit into the framework of weighing.

The requirement that one must be able to use reasons to determine the rational status of actions does not contradict the claim that wholesale rational status is theoretically more basic than the concept of a reason. After all, we typically determine the meanings of particular sentences based on our prior understandings of the words they contain, and this does nothing to undermine the theoretical position that it is better to take the sentence as the basic unit of meaning. True, the relation between *particular* reasons and the rational status of a *particular* action gives the reasons, rather than the rational status, an explanatory priority. But the relation between the *concept* of a reason and the *concept* of wholesale rational status gives the concept of rational status an explanatory priority. And when we are explaining what reasons are, it is this latter relation we are concerned with.

BASIC REASONS

'But surely,' it might be objected, 'even when we are considering the *rational* status of actions, rather than their moral status, there are considerations that are obviously reasons for and against actions that do not have constant weights, even within one particular normative role. For instance, sometimes the fact that a particular person is waiting for me might be a reason *for* going to the place where she is waiting, and sometimes the fact that the same person is waiting there might be a reason *against* going there. This is true even when the sense of rationality at issue is the relevant objective one. Moreover, this reason provides intelligible motivation, given appropriate circumstances, both for going to the place where the person is waiting, and for avoiding that place. That is, someone can act *for* this reason. Intuitively, it *seems* like a reason. We *call* it "a reason." And it meets all the criteria you offer, except for the obvious and drastic failure to

meet the "constant weight" criterion. Surely this shows that the "constant weight" criterion should be abandoned.'

In a sense, this objection is right. But for an important class of reasons – the class directly relevant to a fundamental normative notion – it is wrong. Let us flesh out the example in the objection. How might it be a reason *for* going to a certain place, that a certain someone is there? Well, it may be that I know that this person will be happy to see me. Of course, there may be reasons against going to the place. It may be that it would prevent me from getting important work done. But, if we hypothesize the prospect of the person's pleasure in a meeting, it is hard to deny that the fact that I know the person will be in the place provides me with a reason for going there. Now, how might it be a reason *against* going to a certain place, that that very same person is there? Well, it may be that the circumstances of our relationship have changed, and that I anticipate nothing but unpleasantness from a meeting. Of course there may be other reasons *for* going to the place. It may be that I need to go there in order to get some work done. But, if we hypothesize the prospect of great unpleasantness in a meeting, it is hard to deny that the fact that I know the person will be in the place provides me with a reason against going there.

The answer to this objection lies in distinguishing basic from derivative reasons. Using Mill's methods of difference, we can see that what determines whether or not I have a reason to go to the place *in these cases* is not whether the person is there, but whether it is likely that going to her location will produce pleasure or pain for her.[25] If an action is likely to give someone pleasure, this is *always* a reason in its favor, and it is this reason that stands behind the fact that the person's presence at the place provides me with a reason for going there. And if an action is likely to produce pain, this is *always* a reason against it, and it is this reason that is behind the fact that the person's presence at the place provides me with a reason against going there. This is not to deny that, given the proper circumstances, it *is* a reason to go to a certain place, that a certain person is there. But in order for this kind of fact ('that such-and-such person is waiting there') to be a reason, there must be another sort of reason standing behind it. And this supporting reason ('that I will cause such-and-such person to have some pleasure if I go there') does *not* need anything to stand behind

[25] This is not meant to suggest that pleasure and pain are the only reason-giving considerations. Rather, they are the basic relevant considerations *in the example*.

it. It does not even need the support of my desire to cause that person to have pleasure, for it a reason whether or not I desire this.[26] We can call these latter reasons – reasons that do not need any other considerations to stand behind them – 'basic reasons.' And we can call the former 'derivative reasons.'[27] When we are using reasons to calculate the rational status of an action, it is basic reasons that we are concerned with. And these reasons do make systematic contributions to the rational status of actions.

Although philosophers often use examples of reasons that are nonbasic, it should be clear that anyone who takes seriously the idea of balancing or weighing reasons is committed to there being a preferred or basic level of description for reasons. Otherwise one will end up adding the same reason into the calculation under a variety of descriptions. That is, in the example above, one would end up having the following two reasons to go to a certain spot, where one should have only one: (1) 'that *P* will get pleasure if I go there,' and (2) 'that *P* will be there.' It is basic reasons that should lend themselves to weighing, and it is therefore basic reasons that should make systematic contributions to the rational status of actions. The objection relies on the failure of derivative reasons to make such systematic contributions.

THE FINAL ACCOUNT

Here then is the final formal account of normative reasons.

FA3 In the sense of 'rational' that has to do with objective rationality, a consideration is a *basic reason* if and only if:
(1) it corresponds to an intelligible object of human motivation

[26] This is easiest to see when one keeps firmly in mind that the type of rationality at issue is not the rationality of proper mental functioning. Rather, it is the sense of rationality that is related to claims about whether or not anyone could sincerely recommend the action to the agent, based on the likelihoods of its various consequences.

[27] Another way of trying to achieve the same effect as the distinction between basic and derivative reasons is with a distinction between complete and incomplete reasons. See Raz (1999a), pp. 18–35. This strategy does avoid the problem of there ever failing to be a reason whenever some particular complete reason obtains, even if other circumstances change. But it is extremely plausible that any one of Raz's complete reasons becomes complete precisely in virtue of implying that there is some *basic* reason to do the action: that the action will, for example, avoid pain for someone, or save their lives, or give them pleasure.

(2) it plays at least one of the functional roles (i) or (ii), and has constant strengths, and is comparable to all other reasons, within and across these roles[28]
 (i) making it rationally permissible to do actions that would, without it, be irrational, or
 (ii) making it rationally required to do actions that would, without it, be rationally permissible to omit.
If a reason can fulfill role (i), then it is said to have *justifying* strength. If a reason can fulfill role (ii), then it is said to have *requiring* strength.

Condition (1), admittedly, makes FA3 less than purely formal. But it is also quite plausible that it is eliminable. For it may well be that only intelligible objects of human motivation will meet condition (2); it is certainly hard to think of a counterexample to this hypothesis. (1) is included in FA3 primarily as a reminder of the considerations offered at p. 72. One reason to eliminate (1), in addition to making the account more formal, would be to make it more plausible that a similar account might also apply to reasons for anger, hope, and so on, and especially to reasons for belief. On the other hand, it may be that on a functional role account of reasons for belief, there would be some claim parallel to (1), expressing a claim about reasons for belief corresponding to possible objects of human perception or credence.

Of course, since FA3 is so formal, it does not tell us which substantive considerations actually are reasons for action. But this is as it should be for a functional role analysis. In order for FA3 to help us here, we need some independent way to determine the wholesale rational status of actions.[29] But FA3 will take a substantive account of wholesale rational

[28] C1 and C2 above explain how reasons can be compared with respect to strength within each role. But the relation between justifying and requiring is such that it is also possible to compare the strengths of reasons across roles. For example, we can say the following:

C3 Given two reasons, R_1 and R_2, the requiring strength of R_1 is greater than the justifying strength of R_2 iff it would be irrational to perform any action against which there was a reason with the requiring strength of R_1, and in favor of which there was a reason with the justifying strength of R_2, and to which no other reasons were relevant.

[29] Internalist full-information accounts of normative reasons might also easily be modified to become accounts of wholesale status instead. Modified in such a way, these accounts become much more plausible. For it is *not* plausible that a fully rational agent would have a desire corresponding to *each and every reason* applying to his choice, as these accounts must assume. See Smith (1996). For adherents of such views who continue to want to

status, whatever shape such an account might take, and yield a substantive account of basic reasons. And, prima facie, the prospects seem better for the production of an independent account of wholesale rational status than for the production of an independent account of reasons. For reasons are *pro tanto* in nature, and may sometimes elicit no noticeable behavioral or phenomenological response at all. This may happen, for example, when other relevant considerations provide reasons that are much more important, whether those reasons support or oppose the weaker reason. But with regard to the wholesale rational status of actions, there are characteristic motivational and behavioral responses. At least there are such responses to the status of an action as *irrational*, and this is all that is needed to define 'rationally required' and 'rationally optional.'[30] The 'observational' advantage of starting with wholesale rational status is parallel to an advantage, in accounts of linguistic meaning, of starting with sentences instead of words. For it is sentences that do things. Single words can do similar things only in the degenerate cases ('Run!') in which they form a sentence on their own. Similarly, single reasons can sometimes provide rational requirements or prohibitions. But just as it would be a mistake to base a linguistic theory exclusively on one-word sentences, it is a mistake to examine actions to which only one reason is relevant. Such a restricted view makes it impossible to get a clear view of the justifying role of reasons. For this role is interestingly manifested only when reasons justify acting *in the face of* other reasons.[31] Unfortunately, a casual survey of discussions of normativity reveals an overwhelming preponderance of oversimplified examples.

IMPLICATIONS

It would be hard to overestimate the significance, for contemporary ethical theory, of a general appreciation of the distinction between the justifying

regard them as accounts of reasons, and not wholesale rational status, chapter 6 offers one suggestion that will still allow such theorists to distinguish the requiring and justifying strengths of reasons.

[30] The perceptive reader will notice that with this sentence the sense of 'rationality' seems to have changed from objective to subjective. In fact, what there is a reliable motivational and behavioral response to is neither objective nor subjective irrationality, but something that might be called *apparent objective irrationality*. This is enough, however, to give content to a notion of *actual objective irrationality*, as chapter 7 explains.

[31] In fact, one requires three distinct reasons, yielding three actions corresponding to each of the three possible opposing pairs, in order to produce examples that demonstrate how justifying and requiring strength need not co-vary.

and requiring roles of practical reasons. For it is this distinction which allows us to formulate a view of practical rationality that is consistent with the following two claims: (1) morally required action is always rationally permissible, and (2) not all immoral action is irrational. It is obviously desirable to be able to hold the first of these claims. For if one concludes that, based on all the relevant reasons, an action is not rationally permissible, then nothing remains that one could adduce in favor of performing it. Indeed, one would have to admit that it ought not be performed. These would be unpleasant things to have to say about a morally required action. The reason to hold (2) is that otherwise it seems that we should regard people who perform immoral actions as less than fully rational – and the more egregiously immoral, the more severely irrational. Then we will either have to absolve such people, at least partially, from responsibility for their immoral actions, or we will have to sever or attenuate the connection between rationality and moral responsibility. Neither of these is an attractive option. But if we acknowledge the distinction between the justifying and requiring roles of practical reasons, then it is open to us to construct a moral theory according to which the reasons that favor any morally required action are always sufficient to rationally *justify* it, even though they may not be the sort of reasons that could make it rationally *required*.

It is true that the attraction of (2) depends on the view that it would be a bad idea to sever or attenuate the connection between rationality and moral responsibility. It is possible to challenge this view, or to hold that the connection is already rather more attenuated than I represent it as being. Certainly, the bare fact that an action is irrational is not sufficient to absolve the agent who performs it from moral responsibility. But if someone performs an action *because* of a phobia or a compulsion, we do tend either to excuse her or lessen the degree to which we hold her morally responsible, should that action be one that would have drawn significant moral condemnation if it had been performed by someone without that mental illness. This is because if a desire or aversion is sufficiently strong to cause one to act irrationally – and this is a plausible account of when desires or aversions qualify as compulsions or phobias – it is reasonable to regard them as, in a certain sense, 'irresistible.'[32] Now, if we hold that

[32] Of course not all of the desires that stand behind such illnesses are *literally* irresistible – perhaps *none* are. A compulsive hand-washer could probably be persuaded to refrain from washing his hands if his life was threatened, and an addict could probably be persuaded to defer an injection by similar means. 'Irresistible' should be understood, in such contexts,

strong altruistic reasons – ones that can justify a great deal – also *require* a great deal, then it will be quite severely irrational to act on a desire to harm or kill someone. But if one does harm or kill someone, just for fun, or for some small profit, then what seems to explain this is the lack of a rationally required motive, or the possession of a rationally prohibited one. That is, what explains one's action is the very defect in virtue of which it was irrational. It is hard to see how it would be fair to hold someone morally responsible in such a case.

I am sure that the above argument will leave some readers unpersuaded. But in fact I also think that the best argument for the claim that immoral action is not always irrational is simply a direct appeal to intuitions about rationality, and need not proceed indirectly, through intuitions about moral responsibility. Immoral actions that do not *also* qualify as irrational based on the harms they are likely to cause the agent simply do not seem irrational. Clever low-risk embezzling schemes, if they really are obviously low-risk, do not seem irrational. The raw exercise of power for personal gain, such as that exhibited, for example, by Carlos Menem during his tenure as president of Argentina, does not seem irrational, especially because by stacking the courts with political allies he eliminated any real chance of a criminal conviction in later years. In the face of the fact that so many people who gain such power end up acting in immoral ways, the appropriate response is to admit that what keeps us moral is, *to some degree*, our own self-interested stake in behaving morally – the possibility of punishment, censure, loneliness, lack of friends, and so on. It is much less plausible, as an explanation of the phenomena, to claim that the power to act with impunity either tends to come to those who are already irrational, or that such power tends to produce irrationality.

I do not think that those who hold that all immoral action is irrational adopt that view because it seems correct on its face. Rather, I think they are drawn to it on the basis of theoretical claims that are hard to avoid if one does not acknowledge the distinction between the justifying and requiring roles of reasons. For it is clear that altruistic reasons can indeed be very powerful. If one doesn't notice that the intuitively unproblematic examples that demonstrate this power are exclusively examples that demonstrate *justifying* strength, one is likely to think that altruistic reasons

as 'sufficiently strong that in some cases one's awareness of reasons that would make it irrational to act on the desire would be psychologically incapable of dissuading one from acting.' For an account of disabilities of the will in these terms, see Duggan and B. Gert (1967, 1979).

are very powerful *simpliciter*. And it is a short step from this view to the idea that immoral behavior is irrational. Views that do not acknowledge both functional roles of practical reasons tend to hold that there is generally a unique action that one has *most reason* to perform.[33] But if this is right, then moral theory (as opposed to a theory of practical rationality) is in danger of losing all its practical importance. After all, once one has determined which action one has most reason to perform, what more could one wish to know, for purposes of deciding how to act?[34] Perhaps in cases in which a number of actions were tied for first place, it would be interesting to know which of the actions was morally required. But even in such a case, there would be no special reason to perform the morally required action *instead* of one of the others. If, on the other hand, morality is, as G. A. Cohen has put it, "a choice within rationality," then when all of the rationally permissible options are laid before us, those of us with a concern to behave morally will have some genuine use for moral theory.[35]

[33] As has been mentioned, there are ways of avoiding this by retreating to a satisficing view of rationality, or by appeal to the notion of incommensurability. Chapter 5 argues that satisficing views cannot capture some fairly uncontroversial intuitions about the rational status of particular actions, and that incommensurability cannot capture others.

[34] There is a related danger for views that identify rationality and morality.

[35] See his reply to Korsgaard in Korsgaard (1996a), p. 173.

5

Accounting for our actual normative judgments

In arguing for the theoretical need to distinguish the justifying and requiring roles of practical reasons, some of the work in the previous three chapters was devoted to arguing that we can take certain commonsense normative claims at face value. Among those claims were, for example, that being immoral is not necessarily irrational, but that making sacrifices for other people is not irrational either; that it is irrational to refuse to take medicine that will restore one to perfect health and a happy life, regardless of one's indifference to that prospect; that simply having a strong desire for things like pain or disability does not, by itself, give one the slightest reason to pursue them. This chapter will take these sorts of claims for granted. The point here is to argue in a more formal way against views that attribute a single strength value to practical reasons, and that go on to claim that the rational status of an action is dependent only on the strengths of the reasons that favor and oppose it.[1] Once this is demonstrated, the second half of the chapter will then go on to argue that the justifying/requiring distinction explains the relevant phenomena better than do two other proposals: incommensurability of reasons, and a technical device called an 'exclusionary permission.' In this chapter, as in chapters 2 and 3, the arguments will be presented in terms of subjective rationality. But they can generally be taken to demonstrate formally similar points regarding objective rationality, since the claims will remain true even if we stipulate that the agent is fully informed. And indeed, in the examples I make use of in this chapter, I will make the simplifying assumption that all agents are aware of the relevant reasons, and that they do not falsely believe that there are any other relevant reasons. The possibilities of ignorance and mistake have important normative implications, as we will see in chapter 7, but these possibilities can be bracketed for current purposes.

[1] These strength values could be ordinal or cardinal, vague or determinate; it will make no difference to the arguments presented here.

SINGLE-VALUE VIEWS

It is a common view that:

> R1 If there is a reason in favor of an action, and no reasons against it, then one is required, on pain of irrationality, to perform the action.[2]

On the surface, of course, this is an extremely plausible view. Suppose, for example, that there is only one action that will help one to avoid an immediate painful injury: perhaps one is in the path of a thrown rock. In such a case, and in the absence of any other reasons to stay put, one is rationally required to step out of the way of the rock. Similarly, it is a common view that:

> R2 If there is a conflict between the only two reasons relevant to a choice, and one reason is stronger than the other, then one is required, on pain of irrationality, to perform the action favored by the stronger reason.[3]

In fact, R2 is sometimes taken as explaining what it means for one reason to be stronger than another.[4] Again, this view has considerable intuitive appeal. Consider, for example, the following modification of the above example. It is still the case that one will be painfully injured by a thrown rock, but one also knows that if one allows this to happen, then one will receive a huge amount of money. Perhaps the rock-thrower is a drunken tycoon who has invariably settled such cases in the past. In this case there is a reason in favor of staying put, and also a reason in favor of getting out of the way.[5] To the degree that we regard the former reason as clearly

[2] See, e.g., Nagel (1970), pp. 50–51; Smith (1994), pp. 148, 174–75, and 177; Korsgaard (1996a), pp. 225–26; and Foot (1978b), p. 152. In fact, Nagel is committed to this view only if he means by 'sufficient,' 'sufficient to cause action in a rational agent,' but this does appear to be what he means. And Smith is more concerned with the rational requirement to act when one *believes* there is a reason. But the position this book is defending tells equally strongly against such a view.

[3] See Raz (1999a), pp. 25–26. In fact, Raz is one of the few theorists who, although committed to this view, explicitly recognizes that this cannot be the end of the story (1999a), p. 35. Raz supplements his account by adding second-order reasons and other second-order reason-affecting entities, and by committing himself to widespread incommensurability of reasons and values. See Raz (1999b), p. 46. I argue for the superiority of my solution below.

[4] What R2 actually defines is the circumstance of one reason having sufficient requiring strength that the justifying strength of the other reason is insufficient to justify acting against it.

[5] Of course, this could also be called a reason against staying put.

stronger than the latter, it is plausible to regard jumping out of the way as a case of irrationality – of 'weakness of will' caused by immediate fear of pain. Finally, it is a common view that:

> R3 Rational action in general is action based on the best reasons: 'the rational alternative is the one supported by a preponderance of reasons.'[6]

This view, too, has a great deal of plausibility on its face. Suppose some decision is to be made, and that it is sufficiently complex that a group of people are assembled to research the various options. At the end of the information-gathering and of the assessment of the reasons in favor and against all the options, it would be strange indeed if the head of the committee decided on an option that everyone (including the committee head herself) agrees has less in favor of it than some other option.

The three views listed above increase in strength: R1 is implied by R2, and R2 is implied by R3. To see that R3 implies R2, consider any case in which there are only two reasons relevant to a choice, these two reasons favor different options, and one of the reasons is stronger than the other. This is the type of case about which R2 is making its claim. Now assume that R3 is true – that rational action is action based on the best reasons. Then it follows that in the representative two-reason case, the rational action is the action based on the stronger reason. That is, it follows that R2 is true. To see that R1 is implied by R2, consider a case in which there is only one reason in favor of an action, and no reason against it or in favor of any other action. Such a case is plausibly regarded as a limiting instance of two-reason cases. That is, such a case is plausibly regarded as a case in which there is a reason of *some* strength in favor of an action, and a reason of *zero* strength (no reason) against it. Assume that R2 is true – that when two reasons conflict, the rational action is the action on the stronger of the two. Then it follows that in the one-reason case, the rational action is the action based on the lone relevant reason. That is, it follows that R1 is true.

This chapter argues explicitly against R2: that one is always rationally required to act on a stronger reason. Hence, by implication, R3 will also

[6] See Gibbard (1990), p. 160 and Parfit (1997), p. 99. See also Herman (1993), pp. 166–68. There Herman indicates the common nature of this view by explaining an interesting way in which it might be argued that Kant does not hold it. But her 'defense' of Kant suggests that for all practical purposes he holds both this view and the view that the stronger reason always generates a requirement.

be denied: the view that one is always rationally required to act on the balance or preponderance of reasons. R1, which amounts to the claim that all reasons create prima facie rational requirements, this chapter neither affirms nor denies. The official position of this book is that R1 is false, but as has been mentioned at the conclusions of chapters 2 and 3, a small modification to the official view could easily accommodate the truth of R1 without giving up any significant points. In particular, it is possible to concede R1 while still holding that the requiring and justifying strengths of reasons can come apart. That is, it is possible to concede R1 without conceding R2. And the motivated denial of R2 is a view with very significant philosophical implications.

A MOTIVATING EXAMPLE

Suppose that I am thinking of donating two hundred dollars to a certain charity. The money I donate will provide food for forty children for four months, preventing them, at least for that period, from suffering from serious malnutrition and the attendant risk of illness and death. It is clear that I am rationally allowed to donate the money if it is reasonable to believe that the money will be used for this purpose, and if there are no other significant reasons bearing on the case. But I am not rationally *required* to donate the money just because there is a reason in its favor that is sufficient to justify doing so. Even though the reason in favor of donating the money is sufficient to *justify* doing so, it is not sufficient to *require* it. One point of the justifying/requiring distinction is that it allows us to say that an altruistic reason's insufficiency to require action is not a result of its being too *weak* to generate a requirement. For in one important and intuitive way the fact that an action will prevent serious malnutrition in forty children for forty days is a very strong reason: it can rationally justify a great deal. In this sense, such a reason is in fact much stronger than the reasons against the donation. For the mere prospect of saving two hundred dollars cannot justify nearly as much. And yet it is not irrational to refrain from donating just because one wants to keep the money for one's own indefinite future purposes.

The inability of the altruistic reason to make it rationally required to donate the money is not the result of that reason's weakness. Rather, the insufficiency of this altruistic reason to rationally *require* the donation is the result of its being of a certain type: a type, the function of which is to rationally *justify*, and not to rationally *require*. Even if I could help the

people a great deal more, for the same money, or if I could help them an equal amount at less expense, it might still not be irrational to fail to do so. Indeed, virtually every reader is in this circumstance, for virtually every reader has at least two hundred dollars that she could donate to charity, and that she is not going to donate. Even if a reader has recently donated a large amount, she is still in this position, for unless her donation has actually impacted her finances in a significant way, she still has two hundred additional dollars that could be disposed of relatively painlessly.[7] And the reason in favor of donating those additional two hundred dollars is the same powerful reason: the donation could prevent malnutrition for forty children for four months. For most readers, the loss of two hundred dollars would not make any significant impact on their finances, happiness, or opportunities. As a result, for most readers there is only a very weak reason against giving the money away. And yet I, and they, may rationally decide simply to spend the money on a new coat, or tickets to the opera, even if we know that by donating it we could do a great deal of good for others.

It is the official position of this book that altruistic reasons can *never*, in themselves, rationally require action, even though such reasons can justify actions that stand in need of justification. But neither the motivating examples nor the arguments in this chapter depend upon this view. Rather, the examples simply depend upon the existence of a gap between the justifying strength and the requiring strength of altruistic reasons. It is a matter of indifference to the arguments presented below whether this gap is the gap between *some* and *none*, as the official view holds, or between *more* and *less*.

Again, even though the altruistic reasons in favor of a donation need not rationally require me to donate two hundred dollars, these altruistic reasons are not *weak*. And even though one might be rationally required to spend the same amount of money for some other reason – for example, to avoid the loss of one's index finger – this does not mean that the altruistic reasons in the donation example are *weaker* than the reasons provided by the prospect of saving one's finger.[8] Indeed, in a very important and intuitive sense, given by FA1 in chapter 4, the reason provided by the fact that one's action can save forty children from malnutrition is a stronger reason than the reason provided by the prospect of saving one's finger. For

[7] Graduate students and adjunct professors should modify the amount in these examples.

[8] Of course, one is rationally required to save one's finger at this price only in the context of affluent societies, in which the sum of two hundred dollars is relatively easy to come by.

the very same altruistic reason would justify acting in ways which would be ridiculously irrational otherwise, and which would *not* be regarded as justified by the mere prospect of saving one's finger. For example, suppose that it is wartime, and that one is the sole adult in charge of an abandoned orphanage of forty 'enemy' children. One has gone to the supply base to get the next forty days' rations of food and medicine. There, a report comes that much of the route back to the orphanage is now within shelling and sniper range. Orders come that one is not to risk one's life bringing food to these children. Would it be irrational to disobey these orders, risking one's life and one's career, in order to save these children from starvation, sickness, and serious risk of death? No. Indeed, it would not be irrational to help these children even if one were virtually certain that it would cost one one's life. And yet to risk the same thing merely to save one's index finger would be the height of irrationality. Nor is this case made special by the context of wartime. In cases in which one can save forty children from serious malnutrition for four months, if there is no way of doing so without risking death, it is rationally permissible to risk one's life in order to save the children, and it is rationally permissible not to risk one's life.

A naive reading of the above example, in line with a single-value view of the strength of reasons, would suggest that saving forty children provides a reason of roughly the same strength as does saving one's career and life. But this cannot be the whole of the story, if it is also true that one is rationally allowed either to donate, or not to donate, two hundred dollars in order to save the same number of children from the same kinds of harms. For one is certainly rationally required to spend two hundred dollars to save one's career and life. Indeed one is rationally required to spend two hundred dollars to save one's index finger. The naive view that the rational status of an action is a matter of the balance of reasons – a view expressed by R2 and R3 above – leads to the following claims.

(1) Saving forty children is roughly as important as saving one's career and life, for in a choice between the two, either option is rationally permissible.
(2) Saving forty children is roughly as important as saving two hundred dollars, for in a choice between the two, either option is rationally permissible.
(3) Saving one's career and life is clearly more important than saving two hundred dollars, for in a choice between the two, one is rationally required to spend the two hundred dollars.

Admittedly, these three claims do not involve any formal contradiction. But they do suggest that there may be a contradiction lurking somewhere, for there is a significant failure in transitivity in the relation 'is roughly as important as.' At pp. 94–98 I will try to show where that contradiction lies.

A more sophisticated reading of the above cases suggests that the determination of the rationality of actions is not simply a matter of weighing the univocal strengths of the relevant reasons for and against the action. Rather, the principle that unifies our judgments of all the cases is something more like principle P of chapter 3:

> P It is irrational to do anything that you believe will cause you harm, unless you also believe that someone (perhaps yourself) will thereby be spared at least as significant a harm, or that someone (perhaps yourself) will thereby receive at least as significant a benefit.

Surely many will find P inadequate. Indeed, chapter 7 will point out a number of significant problems with it. One reason some philosophers will object to P is that they wish to give altruistic reasons some measure of requiring strength, and so would favor something more like Q. But for most everyday cases, P is an adequate approximation of Q. For it remains true that it is only when harms to others are disproportionately great that it becomes *irrational* to cause them (or to fail to prevent them) without some reason. In less extreme cases, actions that harm others for some negligible benefit to the agent are not irrational. Rather they are, depending on the magnitude of the harms involved, thoughtless, selfish, mean, cruel, or heinous. On the other hand, if an action involves *any* foreseeable nontrivial harm to the agent, then that action requires a rational justification. And the rational justification of such agent-harming actions must involve the avoidance of harms, or the getting of benefits, of at least a comparable magnitude, either for the agent or for someone else.[9] P allows one to see, in a bold relief that Q may obscure, the logical difference between the power of a reason to require action, and the power of a reason to

[9] This may seem to make it irrational, for example, for parents to make comparatively large sacrifices so that their children will receive comparatively smaller benefits. But real life cases of this kind are very complex. It would be a mistake to think that a parent who spent ten thousand dollars in order to get three thousand dollars to her adult daughter was acting irrationally simply because ten is greater than three. There are many more reasons involved in such cases than simple economic ones, and even economic reasons are not best measured in dollars.

justify actions that stand in need of justification. But if one moderates P to accommodate the intuitions that may make it seem too extreme, this difference in logical space remains. For even Q implies that the justificatory strength of altruistic reasons greatly exceeds their requiring strength. And this is enough to falsify the claim that one is rationally required to act on the stronger reason, no matter which sense of 'stronger' one chooses.

In the examples here, the reasons provided by the prospect of saving forty children from serious malnutrition have a great deal of justifying strength, but no (or not much) requiring strength. This is why one would be rationally justified in undertaking a suicide mission aimed at saving children from these harms, but why one would not be rationally required to do so, and why one would not even be required to donate two hundred dollars to produce the same effects. In contrast to this, the reasons provided by the prospect of saving two hundred dollars have a small justifying strength (they cannot justify risking the loss of one's index finger, much less the risking of one's career and life). But the prospect of saving two hundred dollars also has nontrivial requiring strength, since it is irrational to throw two hundred dollars away without any reason. The justifying strength of the prospect of saving two hundred dollars is in fact at least as great as the *requiring* strength of the altruistic reasons in favor of donating two hundred dollars to a good charity. This is why it is rationally permissible to keep the money. But the justifying strength of the altruistic reasons is far greater than the justifying strength of the reasons to save the money. This is why it is rationally permissible to risk one's life for these altruistic reasons, but it is not rationally permissible to risk one's life, or even (if the risk is genuine) one's finger, to save two hundred dollars.

TWO ARGUMENTS AGAINST SINGLE-VALUE VIEWS

This section provides two technical arguments that the strength of a reason cannot be represented by a single value. Although such a single-value view is not typically explicitly endorsed by philosophers, it is implicit in all *current* accounts that identify reasons with the desires that an agent would have under some sort of ideal conditions.[10] Stephen Darwall makes this point explicitly about the deflationist informed desire account of normative reasons, writing that:

[10] The qualification 'current' is necessary, because, as chapter 6 will argue, such views could accommodate the distinction between justification and requirement by interpreting the theoretically important counterfactual in a more reasonable way.

as a deflationist view, it holds that the normative force of reasons is fully constituted by the motivational pull a consideration exerts when considered in light of knowledge and experience.[11]

In fact, the same point goes just as well for non-deflationist ideal desire accounts of normative reasons. In Darwall's words, such accounts hold that

> p is a reason for S to do A if, and only if, were S to consider p in the right way he would be given some motivation to do A.[12]

The only difference between such an 'ideal desire' view and the deflationist view is that the phrase 'in the right way' in the ideal desire view is replaced, in the deflationist view, by conditions that can be specified in nonnormative terms. This difference, though important in other contexts, is not relevant to the question of whether such views imply a single strength value for any given normative reason: both views have this implication. Therefore, if Darwall is right in his claim about deflationary informed desire accounts, and if the suggestion is right that the point goes just as well for nondeflationary ideal desire accounts, then the single-value view of the strength of reasons is extremely widespread.

Against the single-value view, this chapter presents further arguments that at least two values will be required to characterize the normative capacities of practical reasons. Until one has done the math it is very tempting to suggest that, with the aid of a 'fudge-factor' X – let us call it 'the margin of practical indifference'[13] – a single strength value can do the required work. Or one may think that the job can be done by appeals to vagueness in the measure of the strength of reasons. That is, it is tempting to suggest that the above examples only show that the particular altruistic reasons discussed are strong enough to put their respective actions into the 'rationally permissible' category, but are not strong enough to put them into the 'rationally required' category. Recall the troubling claims that a naive interpretation of some of the previous examples seemed to commit us to:

[11] Darwall (1990), p. 262. The view of reasons as related to ideal desires explains the naturalness of using the word 'force' to describe the normative capacities of reasons. But this terminology lends itself too easily to an interpretation of the normative capacities of reasons based on an analogy with physical forces, and the consequent idea that, given the relevant reasons, there is only one rational way to act. Esther Gert's preference for the term 'force' here seems to me a rare misfiring of her philosophical intuitions.

[12] *Ibid.* [13] This label comes from a suggestion by Paul McNamara.

(1) Saving one's career and life is roughly as important as saving forty children.
(2) Saving forty children is roughly as important as saving two hundred dollars.
(3) Saving one's career and life is clearly more important than saving two hundred dollars.

These claims were troubling partly because of a striking failure in transitivity in the relation 'is roughly as important as.' But failures of transitivity are a commonplace where thresholds or vague concepts are at issue. And surely the strengths of reasons do not admit of precise measurement. Because of this, advocates of a single-value view may believe that appeals to a threshold like the margin of practical indifference, or to vagueness, will be able to explain away any apparent inconsistencies in their view. In order to dispel this illusion, this section offers two arguments showing that a single-value theory cannot, even with the help of a margin of practical indifference, capture the strength of a reason without systematically going against our intuitions about the rationality of certain types of actions. The simplifying assumptions made throughout these arguments are, admittedly, not completely insignificant. But even the schematic arguments provided here should convince an open-minded reader that it is unlikely that a more nuanced single-value theory of normative strength will fare any better.

The two-gap argument

First some remarks about notation. Let us use uppercase letters to label action types. We will call the reasons in favor of an action of type A 'R_A', the reasons in favor of an action of type B 'R_B', and so forth. And let us abbreviate 'the strength of reasons R_A' with the symbol '$S(R_A)$'. For our purposes, action types are individuated only by the number and type of reasons that favor or oppose them.

The single-value theory of reasons and rationality holds that an action is rationally required if and only if the *balance* of reasons in favor of the action exceeds the strength threshold X – the margin of practical indifference. The single-value theory holds that if the balance of reasons favor a given choice, but favor it by a gap smaller than X, then the choice is rationally permissible but not rationally required. In such a case other options are also rationally permissible, as long as the accumulated strength of the reasons in favor of them is within the margin X of the strength of the reasons in favor

of doing the 'best' action. What we need, in order to raise troubles for the single-value view of the strength of reasons is a set of either/or decisions that have the following properties. In deciding between an action of type A and an action of type B, one is rationally required to choose the action of type B. In deciding between an action of type B and an action of type C, one is rationally required to choose the action of type C. In a choice between an action of type A and an action of type D, either is permitted. And in a choice between an action of type C and an action of type D, either is permitted. If we can find action types A, B, C, and D, as described above, then the following will be true.

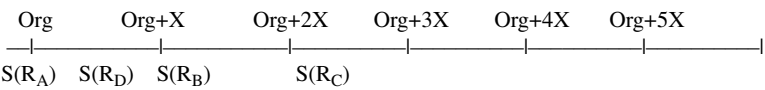
$$S(R_A) + X < S(R_B). \quad (1)$$

$$S(R_B) + X < S(R_C). \quad (2)$$

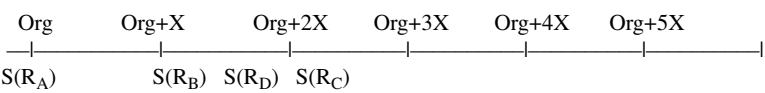
$$|S(R_A) - S(R_D)| \leq X. \quad (3)$$

$$|S(R_D) - S(R_C)| \leq X. \quad (4)$$

We can diagram 1 through 3 as follows:



And we can diagram 1, 2, and 4 as follows:



By inspection of the diagrams, we can see the problem. 3 and 4 assert that $S(R_D)$ must be within the margin of practical indifference, X, both of $S(R_A)$ and of $S(R_C)$. This means (because of 1), that $S(R_D)$ must be less than $S(R_B)$, while (because of 2) $S(R_D)$ must be greater than $S(R_B)$. But $S(R_D)$ cannot be both greater and less than $S(R_B)$.

Can we find actual action types A, B, C, and D that have the properties required to generate this contradiction? We can find many such examples. What we need are action types A, B, and C so that $S(R_A)$, $S(R_B)$, and $S(R_C)$ are clearly at sufficiently wide intervals so that it would be irrational to choose an action of type A over an action of type B, or an action of type B over an action of type C. And then we need an action type D, such that it is rationally permissible, both in a choice between an action

of type A and an action of type D, and in a choice between an action of type C and an action of type D, to choose either. Suppose then that the main consequence of choosing an action of type B over an action of type A will be the loss of two hundred dollars, and that the main consequence of choosing an action of type A over an action of type B will be a month of very bad depression.[14] If one has a reasonable income, it would be irrational to suffer such a depression merely to save two hundred dollars. Thus $S(R_A) + X < S(R_B)$. Now suppose that the choice is not between saving money and being depressed, but between being depressed and dying rather painfully. The strength of the reasons in favor of avoiding depression is still $S(R_B)$. Let us call the action that avoids the painful death an action of type C. It is irrational to die painfully merely to avoid a month's depression. Thus $S(R_B) + X < S(R_C)$. Now for the action type, D. Let us suppose the reasons in favor of an action of type D are that by doing so, one will prevent serious malnutrition, with attendant risk of sickness and death, in forty children for four months. In a choice between saving the children and saving oneself from a painful death, it is rationally permitted to decide either way. Thus, $|S(R_D) - S(R_C)| \leq X$. But in a choice to save the children or to save two hundred dollars – a choice which most readers are now in a position to make – it is also rationally permitted to decide either way. Thus, $|S(R_A) - S(R_D)| \leq X$. And that, plus the assumptions of a single-value view of the strength of reasons, is enough to generate the contradiction.

One of the assumptions I have made on behalf of an adherent of the single-value view is that R_A (the reasons in favor of or against doing an action of type A) can be 'separated' from an action of type A, so that R_A could also be the reasons in favor of or against some other action. In such a case, the strength of the reasons in favor of the two actions would be, in both cases, $S(R_A)$. There are, for example, many different actions that will cause one to lose one's arm. All of these actions have, as a reason against them, that one will lose one's arm. And that reason is always quite strong, although of course there may sometimes be other reasons that outweigh it (for example, in the case in which one's arm is gangrenous). That the same reason can occur in many different contexts, and that it always bears the same weight, seems true to me, and I offered some arguments in favor of such a view in chapter 4. But whether it is true or not, it is an indispensable premise for someone who holds the single-value view. Unless the same

[14] Not, however, so bad as to make suicide likely, or the loss of one's job. In such a case, additional reasons, which I do not want to consider, would favor B over A.

reason, R_A, can occur in different contexts, where it has the same strength, it makes no sense to talk about $S(R_A)$ at all. The notion of *the strength* of a reason is like the notion of *the weight* of an object, or *the length* of a stick. If the regularities in the behavior of objects placed on a scale did not allow one to assign weight values to objects – values that remained constant from occasion to occasion – then we would not have the concept of 'the weight' of an object, and could not make appeal to the weights of objects in explaining why one side of a scale went up, and the other went down.[15] Similarly, the single-value view of the strength of reasons presupposes both that one can re-identify the reason when it occurs in different contexts, and that it bears the same relations of 'weaker than' and 'stronger than' to all the other reasons with which it can be compared. This is why the transitivity of the strengths of reasons can also be assumed in criticism of the single-value view.[16]

In the above example, we chose action types A, B, and C so that $S(R_A)$, $S(R_B)$, and $S(R_C)$ were clearly at sufficiently wide intervals so that it would be irrational to choose an action of type A over an action of type B, or an action of type B over an action of type C. That is, there is clearly a very great difference between the strength of the reason in favor of an action that will save one two hundred dollars, and the strength of the reason in favor of an action that will save one from a month of extreme depression. Because we wanted the intervals between $S(R_A)$, $S(R_B)$, and $S(R_C)$ to be so great, we chose an action A with relatively weak reasons in favor of it. Because of the weakness of $S(R_A)$ it is *more* plausible (but not, I think, *actually* plausible) that one is in fact irrational to choose an action of type A over an action of type D. But given that it is irrational to spend two hundred dollars today *simply* to save one hundred dollars tomorrow, it seems that any plausible X must be rather smaller than the difference between $S(R_B)$ and $S(R_A)$ in the example. This suggests that we could have chosen action types A, B, and C so that $S(R_A)$, $S(R_B)$, and $S(R_C)$ were much closer together. This would have allowed $S(R_A)$ to be much higher, making it still more plausible that it would be rationally allowed to choose an action of type A over an action of type D.

Of course one can still deny the normative judgments made in the above example. One could claim that we are all acting irrationally when we decide to save two hundred dollars rather than donate it to a charity that we believe will use it to prevent serious malnutrition in a large number

[15] See Wittgenstein (1953), §142.
[16] See Darwall (1983), pp. 67–73 for a different sort of argument for the same conclusion.

of children. I do not think very many people seriously hold this belief. But the point of the two-gap argument is not to argue against this belief. Rather, it is to show that the single-value view cannot accommodate our normal intuitions about the rational status of many actions, even using a margin of practical indifference, X, which allows many actions to be labeled 'rationally permissible.' Of course one is always free to deny that our normal intuitions are correct. But one should realize when one is committed to doing so. And holding the single-value view does so commit one.

The equal justification argument

Suppose we have action types A and C. And suppose that in an either/or choice between an action of type A and an action of type C, it is rationally permissible to choose the action of type A. The equal justification argument starts from the premise that if the choice of the action of type A is permissible in virtue of self-interested reasons R_A, then it is possible to construct an action type B, which is also rationally permissible, but in virtue of altruistic reasons, R_B. These reasons R_B are to be understood as involving exactly the same types and quantities of harms or benefits as do reasons R_A, except that while the harms or benefits involved in R_A are harms and benefits exclusively for the agent, the harms and benefits involved in R_B are harms and benefits exclusively for someone other than the agent. Call this premise – that R_A and R_B have the same justifying strength – 'the agent-neutrality of justification.' Here is an example of what the agent-neutrality of justification amounts to. Suppose it is admitted that, in situation S, it is rationally permissible to risk a terrible injury in order to try to save one's life. Then, according to the agent-neutrality of justification, it would also be rationally permissible, in a suitably similar situation S*, to risk the same injury in order to try to save someone *else's* life. Here 'suitably similar' only means that no reasons other than R_B are introduced by the change in situation, and no reasons other than R_A are eliminated.

Recall that the single-value theory of reasons and rationality holds that if, in an either/or situation, the *balance* of reasons favor C_1 over C_2 by a margin in excess of X (the margin of practical indifference) then it is rationally required to choose C_1 over C_2. Further, the single-value view holds that if the balance of reasons favors C_1 over C_2, but only by a gap smaller than X, then either choice is rationally permissible. The equal

justification argument will first show that reasons of equal justificatory strength must, on the single-value view, have equal requiring strength. This should come as no surprise. If it were false, then the single-value view would in fact be allowing the central point for which this chapter is arguing: the logical separability of the requiring and justifying strength of reasons. Then the equal justification argument simply points out that there are reasons that we do not think of as having the same requiring strength, but which the agent-neutrality of justification asserts must have the same justifying (and, hence, according to single-value views, requiring) strength.

In order to show that the single-value view of reasons and rationality cannot use the margin of practical indifference, X, to explain our intuitive judgments, we will need action types A, B, and C that have the following properties. Actions of type A and actions of type B must be favored by reasons R_A and R_B that differ only in the following respect. The harms and benefits involved in R_A accrue to the agent, while the harms and benefits involved in R_B accrue to someone other than the agent. But it must be irrational to choose an action of type C over an action of type A, while it is not irrational to choose an action of type B over an action of type C. If we can find such action types, then the following will be true.

$$(\forall y)((S(R_A) + X > S(R_y)) \supset (S(R_B) + X > S(R_y))). \quad (1)$$

1 simply asserts that whenever it is rationally permissible to choose an action of type A over an action of type y, it is also rationally permissible to choose an action of type B over an action of type y. This is true in virtue of the agent-neutrality of justification, and our stipulation that R_A and R_B differ only in respect of the person to whom benefits and harms accrue. But we also have, because of our stipulation about actions of type C, the following.

$$S(R_C) + X < S(R_A). \quad (2)$$

$$S(R_C) + X > S(R_B). \quad (3)$$

2 claims that it is irrational to choose an action of type C over an action of type A, while 3 claims that it is rationally permissible to choose an action of type C over an action of type B.

Now choose action type D, such that

$$S(R_D) = S(R_C) + 2X. \quad (4)$$

Solving for $S(R_C)$, we get

$$S(R_C) = S(R_D) - 2X. \tag{5}$$

From 1 we have

$$(S(R_A) + X > S(R_D)) \supset (S(R_B) + X > S(R_D)). \tag{6}$$

From 2 and 5 we have

$$(S(R_D) - 2X) + X < S(R_A). \tag{7}$$

Simplifying, we get

$$S(R_D) - X < S(R_A). \tag{8}$$

Adding X to both sides we get

$$S(R_D) < S(R_A) + X. \tag{9}$$

This is the same as

$$S(R_A) + X > S(R_D). \tag{10}$$

From 10 and 6 we get

$$S(R_B) + X > S(R_D). \tag{11}$$

But from 3 and 5 we have

$$(S(R_D) - 2X) + X > S(R_B). \tag{12}$$

Simplifying, we get

$$S(R_D) - X > S(R_B). \tag{13}$$

Adding X to both sides we get

$$S(R_D) > S(R_B) + X. \tag{14}$$

This is the same as

$$S(R_B) + X < S(R_D). \tag{15}$$

But this contradicts 11.

As in the two-gap argument, this contradiction will only cause troubles for the single-value view of reasons and rationality if we can find actual action types A, B, and C that have the required properties. But again, we can find many such examples. Suppose that actions of type A involve saving my own index finger, and actions of type B involve saving someone

else's index finger, and actions of type C involve saving myself the now familiar sum of two hundred dollars. If the agent-neutrality of justification is true, then premise 1 is true. If it is irrational to sacrifice one's index finger *merely* to save two hundred dollars, then premise 2 is true. And if it would not be irrational to refuse to spend two hundred dollars to save a stranger's index finger, then premise 3 is true. Again, even with the aid of the margin of practical indifference, X, the single-value view of reasons and rationality cannot accommodate our intuitions here, as long as one grants the premise I have called 'agent-neutrality of justification.' In assessing the independent plausibility of the agent-neutrality of justification, one should not be misled by the name. It is not the strong premise that it makes no difference to any question about the rationality of an action, whether the relevant benefits or harms are going to accrue to the agent or to someone else. That is a much more controversial claim, and it should already be obvious that according to the view on offer here, it is false. Rather, the agent-neutrality of justification only involves the following claim: that if it is rationally permissible to make a certain sacrifice for one's own sake, it would also be rationally permissible to make that same sacrifice for someone else's sake. That does not mean that, rationally speaking, one *should* make the sacrifice. In only means that if one chose to do so, one would not be choosing irrationally.

As with the two-gap argument, one can avoid the conclusion of the equal justification argument if one is willing to reject one or more of the premises. Thus, if one is enamored of a single-value view of reasons and rationality, one can claim that the agent-neutrality of justification is false. Or one can hold that any cost that I would be rationally required to pay in order to save my own index finger I would also be rationally required to pay in the interest of saving the index finger of a complete stranger. But if one does not want to make either of these concessions, then one cannot continue to hold that the strength of a reason can be represented by a single value, and that rational action is action favored (within some threshold) by the balance of reasons.

OTHER SOLUTIONS

Even if the preceding arguments have successfully demonstrated the inadequacy of single-value views, this by itself does not provide very strong support for the justifying/requiring distinction. For there may be better ways to account for the intuitions that caused trouble for single-value

views. This section will briefly present and criticize two alternate solutions, both defended by Joseph Raz: widespread incommensurability, and exclusionary permissions. In the case of incommensurability, it will even turn out that the requiring/justifying distinction can help explain how certain normative phenomena might have led Raz and others to embrace such a troublesome and paradoxical concept.

Widespread incommensurability

Joseph Raz explicitly endorses the commonsense view that the normal case of decision-making is one in which the relevant reasons make a number of options eligible, but do not require any of them. He calls this claim 'the basic belief,' and he is sufficiently strongly committed to its truth to hold that, despite difficulties in seeing the theoretical position it expresses, it should be maintained unless it can be shown to be incoherent.[17] One way in which Raz expresses the basic belief is by saying that the primary function of practical reasons is to render options eligible, rather than to require them. This sounds very much like the claim that the primary function of reasons is to justify actions, and not to require them. Indeed, at points one might even think that Raz endorses the position defended in this chapter. But although Raz is one of the very few philosophers who have explicitly considered a justifying/requiring distinction, he ultimately rejects such a view because he does not see the possibility that the very same reason might have different justifying and requiring strengths.[18] Rather, he only considers the possibility that some reasons are exclusively justifying, while others are exclusively requiring.

Raz's strategy for explaining the basic belief is to claim that the reasons that favor incompatible actions are often incommensurable. His interpretation of incommensurability is clearest in the two-option case, and what it means there is that neither of the opposed reasons defeats the other, but that they are also not of equal strength. Because of this, Raz holds that reason cannot adjudicate between them, or between the actions they favor. Raz also holds that as long as one acts on an undefeated reason, one is acting rationally. This claim, together with Raz's interpretation of incommensurability, explains why action on either reason, in the

[17] Raz (1999b), pp. 100–1.

[18] Raz (1999b), pp. 101–2. Raz puts the distinction in terms of enticing and requiring *reasons*, rather than in terms of enticing and requiring *roles* or *strengths*, and this contributes to his rejection of it.

two-option case, would be rationally permissible.[19] Such an explanation allows Raz to make all the normative judgments expressed in the two-gap and equal justification arguments, without falling into the inconsistencies that undermined single-value views. For example, in the two-gap case he could assert the incommensurability of the relevant altruistic and self-regarding reasons. This would explain the rational permissibility of choosing either A or D when presented with those two options, as well as the rational permissibility of choosing either C or D when presented with those two options. And if he continued to hold that the reasons involved in actions of type A, B, and C were all commensurable, he could continue to hold that it is rationally required to choose actions of type B over actions of type A, and actions of type C over actions of type B. Similar reasoning would apply to the examples used in the equal justification argument.

One might think that the preceding brief presentation of Raz's view of incommensurability must be unfair. Surely he must provide some *explanation* for the paradoxical claim that there can be pairs of reasons, neither of which is at least as weighty as the other.[20] But Raz gives no positive explanation. Rather, he shifts the burden of proof, so that the challenge becomes 'How could reasons involving apples be compared in weightiness with reasons involving oranges (or aches)?' Difficulties in explaining how such comparisons could be made or defended, combined with a need to explain the 'basic belief,' push Raz to endorse the thesis of widespread incommensurability of reasons without much explanation of how it could be true. So in arguing against Raz here, it will be useful to do two things. The first is to mention a problem with the notion of incommensurability that undermines its usefulness in accounting for our intuitions in cases like those used in the two-gap and equal justification arguments. And the second is to try to dispel some of the air of mystery surrounding the ability to compare reasons that have very different substantive content.[21]

One of the most powerful objections to incommensurability has been called 'the nominal-notable objection.'[22] Before presenting it, it will be useful to give a quick argument that seems to support the idea that there can be incommensurable reasons. Suppose, then, that I am trying to choose

[19] Raz (1999b), pp. 102–3.

[20] For this alternate expression of the mutual nondefeating nature of incommensurable reasons, in terms of weight, see Raz (1999b), p. 182.

[21] In fact, Raz himself provides one way in which mixed values, such as romance novels, can be compared, even though they involve a number of distinct goods. See Raz (1999a), pp. 182–201.

[22] See Ruth Chang's introduction to Chang (1997).

between two options, A and B, and that the only relevant differences between them is that choosing A will spare me a month of annoying but not debilitating knee pain but cost me four hundred dollars, while choosing B will save me the four hundred dollars, but will result in the month of knee pain. In this case, let us grant that it would be rationally permissible to choose either A or B. This could be true because the reasons favoring each option are exactly equal in strength. But this explanation loses appeal once one grants the plausible claim that even if one increased the cost of A to five hundred dollars, both choices would remain rationally permissible. This suggests, in favor of Raz's position, that the permissibility of the two original options was not the result of their being favored by reasons of precisely the same strength, but was the result of their being favored by incommensurable reasons. However, this same strategy of changing the cost can be used against an advocate of incommensurability. For suppose that we now reduce the cost of A to some nominal value – say, three dollars – or that we increase the duration of the knee pain to some quite notable degree – say, five years. These changes do not alter the nature of the values underlying the reasons, and should therefore not alter either the fact of incommensurability or the consequent rational permissibility of either choice.[23] But they do, since it is irrational to refuse to spend three dollars to avoid a month of annoying but not debilitating knee pain. That is the nominal-notable objection.

An advocate of the justifying/requiring distinction can agree with Raz that the rational permissibility of either option in the first choice was not the result of a precise (or even a rough) balance of reasons. But it was not the result of incommensurability either. Rather, it was a result of the fact that each reason, in the original case, possessed greater justifying strength than the requiring strength of the other. This response is not open to the nominal-notable objection. For when the cost of the cure is reduced to three dollars, the justifying strength of the economic reason is greatly reduced and can no longer justify suffering any significant amount of annoyance or pain. Thus it becomes irrational to refuse to spend the money.

It now remains to suggest how it is possible to compare reasons that involve diverse values. A full solution to this problem is beyond the

[23] If the reader believes that changes in these quantities of pain and money bring with them other normatively relevant changes, other examples can be chosen that involve less complex goods. Despite any quarrels with the details of the example, the point should remain clear.

scope of this book. But it is worth noting that one consideration that has led philosophers to doubt that such reasons can be compared is that in conflicts of such reasons there is no uniquely rational choice. The requiring/justifying distinction accommodates this fact. This said, here is a suggestion as to how to solve the comparison problem. It was argued in chapter 4 that the basic unit of normative assessment should not be 'a reason.' Rather, the notion of wholesale rational status should be taken as more basic than that of a reason. In arguing for this conclusion, appeal was made to an analogy with language: just as it is more profitable to take the *sentence* as the basic unit of meaning, it is more profitable to take *wholesale rational status* as the basic unit of normative assessment relevant to actions. And just as we can identify the various meanings of individual words with the systematic contributions that they make to the meanings of sentences, we can understand the various normative roles of reasons, and their strengths in those roles, by understanding how they contribute systematically to the wholesale normative statuses of actions. In this way, the ability to compare apples and aches is not a *prerequisite* for the ability to judge it rationally permissible to give up an apple to spare oneself an ache. Rather, the ability to make wholesale normative judgments about the rationality of action – an unsurprising ability to find in normal human beings – is a prerequisite to learning about the relative normative capacities of reasons involving apples and aches.

Why have philosophers been led to embrace a notion as paradoxical as the sort of incommensurability described above? The justifying/requiring distinction provides an explanation for this fact. Philosophers have typically assumed that if reasons have comparable strength values, then those values will be characterizable with single values, whether cardinal or ordinal, rough or precise. But if the justifying/requiring distinction is valid, and if reasons with equal justifying strengths often have very different requiring strengths, then attempts to characterize the relative strengths of reasons with a single value will often fail: two values will be required. The thesis of incommensurability may be a response to the perception of this systematic failure for particular pairs of reasons. In essence, the thesis of incommensurability is an inference from the true premise 'the strengths of these reasons cannot be compared as if they were single values' to the false conclusion 'the strengths of these reasons cannot be compared at all.' Put in this way, the inference is obviously formally invalid. But it is easy to understand how one might make this inference if one did not see that reasons might have more than one dimension of strength.

Exclusionary permissions

A second proposal by Raz for accommodating the relevant phenomena agrees with single-value views in holding that first-order reasons possess a single strength value. Moreover, this proposal also agrees that the strength of a first-order reason is a measure of its power to override other first-order reasons. But Raz goes on to claim that it is sometimes rationally permissible, but not required, to exclude some first-order reasons from one's deliberations. By making this additional claim, Raz can account for our intuitions in the two-gap and equal justification arguments. The name that Raz gives to the normative entity that makes it rationally permissible to exclude some first-order reasons is 'an exclusionary permission.'

Raz uses exclusionary permissions to explain the phenomenon of supererogation.[24] Here is how the explanation works. Suppose that an agent has, as one of the options open to him, a morally supererogatory action. Raz assumes that since such action is praiseworthy there must be reasons in its favor outweighing all conflicting reasons. Indeed he thinks of such actions as in a sense required by reason.[25] These are themselves problematic assumptions, but let them stand. That is, let us assume that first-order reasons unambiguously favor the supererogatory action. Now, since the action is supererogatory, we are also assuming that it is neither immoral nor irrational to refrain from it. How can all of these claims be true? How can all the relevant reasons unambiguously favor the action, and yet not make it rationally required? Raz's answer is that in cases of supererogation there is an exclusionary permission that allows us to

[24] See Raz (1999a), pp. 89–95 and Raz (1975). It is worth noting that cases of moral supererogation only form a subset (albeit an important one) of the set of actions whose rationally optional status is explained by the justifying/requiring distinction. This is partly because altruistic reasons, which, on my account, have much more justifying strength than requiring strength, can just as easily be reasons for immoral action as for morally good action. An altruistic reason can rationally justify a risky immoral action, undertaken for the sake of one's family, friends, or colleagues. Accounts that attribute rationally optional status *only* to morally supererogatory action are inadequate accounts of practical rationality.

[25] Raz (1999a), pp. 91, 94, and Raz (1975), p. 165. From the claim that it is always wrong to say one ought not perform a supererogatory act, Raz (1975), pp. 165–6 infers that such acts are always favored by conclusive reasons. This suggests that he only allows for two sorts of actions: those one ought to perform, and those one ought not perform. An alternate view claims that there are actions of which it is false that one ought to perform them, and also false that one ought not. I suspect that most actions fall into this class, and that Raz's notion of incommensurability also commits him, against this 1975 argument, to such a view.

exclude some of the first-order reasons that would otherwise make the action rationally required. Presumably these excludable reasons include all or some of the altruistic ones. In effect, an exclusionary permission generates rational options by allowing two different calculations based on the relevant first-order reasons: one that uses all the first-order reasons, and one that uses only a subset of them. Since an exclusionary permission only says that one *may* exclude a certain subset of first-order reasons, but does not *require* one to do so, either of the calculations is permitted. Since the calculations favor different actions, both actions are also permitted.

Here is how an exclusionary permission would help Raz avoid the force of the two-gap argument. Suppose that, in cases in which only first-order reasons are relevant, the first-order reasons favoring an altruistic action of type D would be sufficiently strong to rationally require action in cases in which the only other option was an action of type A or an action of type C. But suppose further that in the actual case there is an exclusionary permission, allowing us to exclude the altruistic reasons favoring D. Given these assumptions, all of our intuitions about the examples in the two-gap argument can be preserved. For although one calculation would favor an action of type D over an action of type A, there would be an alternate calculation available to the agent who had to choose between the two types of action: the calculation based on R_A alone, excluding R_D completely. Clearly this calculation would favor choosing A. Since either calculation is permitted, so is either action. Parallel reasoning explains why, in a choice between actions of type C and actions of type D, either action would again be rationally permissible. And yet, since there is no exclusionary permission that allows the agent to exclude R_A, R_B, or R_C, it can remain true that the agent is rationally required to choose actions of type B over actions of type A, and actions of type C over actions of type B. This accounts for all of the intuitions that were used in the two-gap argument. Similar reasoning could be used to show that a theory that includes exclusionary permissions might also avoid the force of the equal justification argument.

But despite the adequacy of exclusionary permissions in accounting for our intuitions in many cases, they are inferior to the justifying/requiring distinction as a means of saving the phenomena. Their greatest liability is their peculiar ontological status. Where do exclusionary permissions come from? Raz is clear that they cannot be taken for granted, and require some

sort of justification.[26] But it seems very implausible that there will always be a justification for an exclusionary permission whenever we need such a permission to account for our intuitions. This objection, however, is unlikely to impress anyone who is antecedently committed to exclusionary permissions. Such theorists will simply say that either our intuitions are incorrect in those cases, or there is indeed some justification for the exclusionary permission that we have not yet discovered.

A more formal difficulty with exclusionary permissions can be seen by imagining a case in which such a permission allows the exclusion of just one reason from one's deliberation. This allows two possible calculations: one in which the reason is included with all of its strength, and one in which it is excluded completely. The problem here is that we sometimes think that a reason must play a certain *minimal* role in the decision of an agent, even if we agree that it could permissibly play a much greater role. Exclusionary permissions cannot accommodate this. They can only accommodate such a reason's playing *no* role, or its playing a *full* role. For example, suppose that an agent is rushing to catch a bus into the city. She is quite late, and if she misses the bus it will cause her quite a bit of annoyance, although it will not affect anyone else. As she is jogging along the sidewalk, a man with a map and a puzzled face tries to stop her to ask directions. Here the agent has the choice of stopping to help, which will increase her chances of missing the bus, or not stopping, which will leave the man unaided. Let us grant that in this case either action would be permitted, and that this is the result of an exclusionary permission that allows the agent to exclude the reason provided by the puzzled man's interests. Considered in isolation, this choice situation seems to be explained adequately by appeal to this exclusionary permission. But suppose we wish to make the additional conditional claim that, were the inconvenience of stopping to help much less significant (for example, if the buses came with much greater frequency) the woman would be unreasonable not to stop and help. Exclusionary permissions do not seem to allow us this additional claim. For if there is a permission to exclude the altruistic reason in the first case, surely it does not disappear or cease to apply when the strength of the opposing self-interested reason is decreased. One could of

[26] Raz (1999a), p. 90. In Raz (1975), there is an argument for certain exclusionary permissions based on incommensurability. But this cannot be the general form of such an argument, since such a form would never allow the exclusion of only one type of reason – say, altruistic ones. Rather, arguments based on incommensurability will always allow the exclusion of reasons on 'both sides' of the incommensurability. Moreover, the notion of incommensurability itself is problematic.

course say that it does disappear or cease to apply. But at that point the theory of exclusionary permissions seems to lose a good deal of its appeal, especially when compared with the justifying/requiring distinction. One could also argue that exclusionary permissions are sometimes permissions to take certain reasons as having less strength than they actually have, rather than being permissions to exclude certain reasons completely. But again, this move seems in need of a great deal of defense, whereas the justifying/requiring distinction really ought to be taken as the default position. That is, chapter 4 showed that, whatever one's view of rationality, the roles of justifying and requiring are logically distinct. Thus the concepts of requiring strength and justifying strength are also distinct. And it is a much stronger thesis to hold that these strengths are always the same, for any given reason, than that they are not.

The justifying/requiring distinction could explain the above case in the following way. The altruistic reason to help the man with the map has a certain minimal requiring strength, and a relatively greater justifying strength. Its requiring strength is sufficiently great that the agent is not rationally permitted to act against it when the opposing reason is merely that she will have to wait five minutes for the next bus. But the justifying strength of the altruistic reason is still sufficiently great that it can make it rationally permissible for the agent to risk even the extreme annoyance of missing a very infrequent bus. Of course one may disagree with the normative judgments expressed in the discussion of this example. In fact, I myself do disagree with them, since I claim that altruistic reasons have *no* requiring strength, rather than the *minimal* requiring strength to which the explanation appeals. But, as has been mentioned a number of times in previous chapters, nothing of great significance rests on this particular claim. Moreover, and as Raz himself insightfully claims in his presentation of exclusionary permissions, the point here is only to explain how someone who makes the above normative judgments can be interpreted in a coherent way.[27] It is neither here nor there whether such a person is correct in her particular normative judgments.

CONCLUSION

Making a distinction between the requiring and justifying strengths of practical reasons, and holding that altruistic reasons have far less requiring

[27] Raz (1999a), pp. 91, 93.

than justifying strength, allows us to explain why it is rationally permissible (if rather stingy) for middle-class people to donate nothing to charity, but also why it is rationally permissible for them to donate quite a lot. The distinction allows that being selfish and mean, even to the point of violating moral requirements, need not involve any irrationality, but also that it is rationally permissible to devote one's life to the relief of the suffering of others. And yet the distinction, and the associated claim about altruistic reasons, does not make 'everything permitted.' For many reasons have requiring strength: those that involve nontrivial harms to the agent, and, perhaps, those that involve great harms to others.

The distinction between justifying and requiring gives a *sensible* sense to the idea that the reasons in favor of donating two hundred dollars to charity are far stronger than the reasons against doing so: they have far more *justificatory* strength. This is not a special technical sense of 'stronger.' Many ordinary people would agree that saving forty children from malnutrition is, in some relatively straightforward sense, more important than saving two hundred dollars or getting a new winter coat. The distinction between requiring and justifying strength, and the claim that altruistic reasons have more justificatory than requiring strength, gives sense to the claim that despite this, one is not rationally required to donate the money. That is, one is not always rationally required to act on a reason that is, in an important and intuitive sense, clearly the stronger of the two primary reasons bearing on one's choice. While this claim may sound paradoxical in the abstract, the theory that stands behind it provides a better explanation for many of our normative judgments than do the notions of incommensurability or exclusionary permissions.

6

Fitting the view into the contemporary debate

The claim that some reasons have greater justificatory strength than requiring strength entails a number of further claims that are at odds with a good deal of current philosophical dogma. For example, it entails that one need not, rationally, always act on the stronger of two opposed reasons, even in the absence of other relevant considerations. And it holds that this is true whether one takes 'stronger' to mean 'stronger in the requiring role' or 'stronger in the justifying role.'[1] The official view advocated in this book also denies the internalism requirement on practical reasons, for it holds that it is not irrational to be completely unmoved by altruistic reasons. Given these conflicts with contemporary views, and given also what appears to be a significant *structural* difference between the view advocated here and other views – two strength values, as opposed to only one – some readers may have begun to suspect that the notions of practical rationality and reasons for action that form the subject of this book, while interesting and significant, are simply different notions than those of concern to other contemporary philosophers who use the same lexicographical terms. At the very beginning of chapter 1 I explained why this suspicion is unfounded: we are all engaged in the same project of trying to produce an account of the fundamental normative notion relevant to action. Moreover, we all take the relation between this notion and the 'mental functioning' interpretation of rationality to be sufficiently close that a fully informed agent, acting in a subjectively rational way, must also be acting in an objectively rational way – a way that the fundamental normative principle allows.

Nevertheless, it will serve a number of purposes to show how the view developed here is related to certain accounts of the same subject: ideal

[1] The first of these claims is true because the reason with less requiring strength may nevertheless have sufficient justifying strength to make it rationally permissible to act on it. And the second is true because greater justifying strength does nothing to generate a rational requirement.

motive accounts of normative reasons. One of these purposes is to show that these accounts, when one rids them of an assumption that is, independently, quite implausible, yield views that include two values for any given normative reason. These values function in a way that is isomorphic to justifying and requiring strength. All the arguments, therefore, in favor of distinguishing between justifying and requiring strength can be added to the arguments in favor of modifying ideal motive accounts by removing the implausible assumption. And all the arguments against the implausible assumption can be added to the arguments in favor of regarding reasons – the kind of reasons under discussion in this book, and the kind of reasons under discussion in the work of other ethical theorists – as having two independent dimensions of strength.

MOTIVATING IDEAL MOTIVE ACCOUNTS

Normative reasons are reasons of the sort that are involved in claims such as 'One reason for Jones to take a vacation is that it would make him feel more relaxed' or 'Gates ought to donate more to charity for the following reason: it would help a lot of people in a significant way.' Even when such reasons are altruistic, they are not specifically moral. Indeed, one purpose for which philosophers develop accounts of normative reasons is to provide a morally neutral foundation for subsequent arguments that seek to show why one should be moral. Such reasons get their normative significance in virtue of the way they contribute to claims about whether our actions are rationally permissible or irrational. For to call an action 'irrational' in the relevant sense is, among other things, to claim that it ought not be performed. Running parallel to this normative sense of 'reason,' there is another, nonnormative sense, which we might call 'motivational' or 'explanatory.'[2] Motivational or explanatory reasons are involved in claims such as 'Smith spat on Jones for the following reason: he (Smith) hated people of Jones's religion.' As this example suggests, motivational reasons need not provide any *normative* support for the actions they help to explain. For ordinary agents, there is always the possibility that some of the motivational reasons involved in the explanation of a particular action will not be normative reasons. And there is also always the possibility that some

[2] For one clear presentation of the distinction between normative and motivating reasons, see Smith (1994), pp. 94–98. Smith holds that an agent must regard his motivating reasons as normative reasons. The final chapter of this book argues that this is false.

normative reasons applicable to an agent's choice might completely fail to be motivational reasons for that agent. There is, however, an undeniable appeal to the view that when normative and motivational reasons come apart in either of these ways, it must be the result of some failure on the agent's part. Thus, it seems plausible to claim that in an agent who was ideal in relevant respects, all normative reasons would have some motivational pull, and that all motivating reasons would also be normative reasons.

Motivated by the plausibility of this last claim above, one popular and well-known type of account of normative practical reasons associates such reasons with the motives that an agent would have under certain ideal conditions.[3] For example, on such an account the question of whether an agent has a reason to have a cup of coffee is reduced to the question of whether, under certain ideal conditions, that agent would be motivated to some degree to have a cup of coffee.[4] And the question of the strength of a reason is reduced to the question of the strength of the associated hypothetical motivation. It is because of the *ideal* element in the conditions under which it is claimed that the agent would be motivated, that such accounts are sometimes referred to as 'ideal motive accounts.' All such accounts take the following essential form:

> p is a reason for S to do A if, and only if, were S to consider p in the right way he would be given some motivation to do A.[5]

[3] For the canonical versions of this type of account, see Darwall (1983), pp. 41, 81; Brandt (1979), pp. 10–15; Smith (1994), pp. 150–81. Brandt's account is cast in terms of rational desires, rather than reasons. But rational desires function for Brandt in exactly the same way that reasons function for other ideal motive theorists. One might say that Brandt so identifies reasons with rational desires that he has no need for the additional term 'reason.' And indeed the word 'reason' hardly occurs in Brandt (1979). See also Railton (1986).

[4] The phrase 'can be reduced to' is meant to be neutral as between a number of different views about what such accounts are trying to provide: meaning analyses, truth-conditions, or something else. Perhaps least controversially, 'Question *X* can be reduced to question *Y*' may be taken to mean '*X* can be straightforwardly answered by answering *Y*.'

[5] This way of characterizing such views is taken from Darwall (1990), p. 262. Technically, this characterization excludes accounts such as Michael Smith's. For on Smith's account, the hypothetical motivation is not a motivation that the ideal S would himself have, to perform A. Rather, the hypothetical motivation is, for Smith, the desire, on the part of the ideal S, that *the actual* S perform A. Since this distinction is irrelevant to the point of this chapter, I group Smith with Darwall and Brandt. Smith's modification is an attempt to avoid what has been called 'the conditional fallacy.' See Johnson (1999). In J. Gert (2002b) I propose a different solution, based on the idea of an appropriate level of description for action types and desires: descriptions in terms of the *basic ends* of the agent. Nothing of

Among ideal motive accounts, there are both naturalistic versions, and versions that make use of ineliminably normative terminology. The only difference between naturalistic versions and nonnaturalistic ones is that the phrase 'in the right way' in the above general characterization is replaced, in naturalistic accounts, by ideal conditions that can be specified in purely naturalistic terms, while for nonnaturalistic accounts the ideal conditions have an ineliminably normative aspect. This difference, though important in other contexts, will not be relevant to the purposes of this chapter. Nor is it important, for current purposes, whether such accounts are taken as providing an analysis of the meaning of reason-claims, or a statement of their truth-conditions, or something else. Moreover, although the plausibility and usefulness of any ideal motive account will of course depend on exactly how it specifies the relevant ideal conditions, the arguments of this chapter will not depend on a detailed assessment of the merits of any of these specifications.

It is not the primary purpose of this chapter to attack the general strategy of associating practical reasons with idealized motives. Rather, the purpose is to point out an assumption common to existing versions of such accounts, and to examine the consequences of rejecting this assumption. The assumption is that any given consideration would generate a unique degree of motivation in any given agent, if that agent were ideal in relevant respects. We can call this 'the uniqueness assumption.' On an ideal motive account the uniqueness assumption entails that we can always describe the strength of a reason with a single scalar value. If one rejects the uniqueness assumption then one holds that, within a theoretically significant range, a given consideration might generate any degree of motivation, even in an ideal agent. As a result, one will need two values in order to characterize the normative capacities of any given reason. These two values will correspond to the maximum and minimum of the range of acceptable degrees of motivation. The point of this chapter is to show that, with the simple rejection of one unnecessary assumption, ideal motive accounts become much more similar to the account of normative reasons advocated in this book. For the minimum and maximum of the range of acceptable degrees of motivation correspond, respectively, to the requiring and justifying strengths explained in previous chapters.

substance would have to be altered in the present argument in order to accommodate my suggested solution to Johnson's problem. Consequently, that problem, and my solution to it, are not presented here.

THE ASSUMPTION

The uniqueness assumption is sometimes smuggled into ideal motive accounts concealed in the word 'the.' For example, Stephen Darwall makes the uniqueness assumption on behalf of naturalistic ideal motive accounts when he explains that on such accounts:

> the normative force of reasons is fully constituted by *the* motivational pull a consideration exerts when considered in light of knowledge and experience.[6]

Words such as 'would,' 'most,' and 'best' can also help to import the assumption. For example, Michael Smith suggests a way in which facts about what we ought to do, all things considered (in my terminology, what is objectively rational) are fixed by the desires we would have if fully rational. His suggestion is that what we have all things considered reason to do is:

> fixed by the relative strengths of these [fully rational] desires: that is, by facts about what our fully rational selves *would most* want us to do in the relevant circumstances.[7]

That is, Smith equates the normative strengths of reasons with the unique strengths of certain hypothetical desires.

Critics of ideal motive accounts also make this assumption on behalf of the theorists they criticize. Connie Rosati, for example, in her criticism of ideal motive accounts of the good, rightly points out that the different ways in which one might become fully informed are likely to result in one's ending up with different motives. From this she concludes that, when one is constructing a full-information ideal motive account of the good:

> the process of fully informing a person will need in some way to offset the effects of experiential ordering, by requiring, for instance, that a person experience all lives . . . in all possible orders.[8]

[6] Darwall (1990), p. 262, italics mine. Although he does not state it explicitly, it is clear that Darwall makes the uniqueness assumption for nonnaturalistic versions also. See also Smith (1996), p. 167 and Brandt (1979), p. 15.

[7] Smith (1996), p. 167, emphasis added. See also Sidgwick (1981), pp. 111–12; Rawls (1971), pp. 417–18; Brandt (1979), pp. 126–29; Railton (1986), p. 16; Pettit and Smith (1993), pp. 53–79; Rosati (1995), p. 302, esp. n. 15.

[8] Rosati (1995), p. 309. It is true that Rosati is concerned with accounts of what is good for a person, and not with accounts of normative reasons. But the example illustrates the same assumption about the uniqueness of ideal motivation. Moreover, an account of the good for a person can plausibly be regarded as yielding at least a partial account of that person's reasons for action.

But why will these ordering effects need to be offset? The answer seems to be: 'Ideal motive theorists will need to offset these effects, because for their view to be viable, there must be a unique fixed set of motivations that they can claim the ideal agent will possess.' But why is this? This question will be especially hard for Rosati to answer, since she concedes that the effects of experiential ordering need not be regarded as the result of any cognitive *problem*.

It may well be that some ideal motive theorists use definite descriptions such as 'the motivational pull a consideration would generate under ideal circumstances' without committing themselves to the uniqueness assumption. They may be using the phrase 'the motivational pull' in the same way one uses the phrase 'the height of a grown man.' No one thinks that there is a unique height shared by all grown men. Rather, it is clear that this latter phrase indicates a range of heights. Moreover, while the boundaries of the range may be vague, the whole range is not merely the result of vagueness. Both 5′8″ and 6′1″ clearly count as 'the height of a grown man.' Most ideal motive theorists do not appear to be using the phrase 'the motivational pull' in this looser way.[9] But whether or not they are, and whether or not there are other theorists who would immediately accept this chapter's proposed modification, *no ideal motive theorist has explicitly denied the uniqueness assumption, much less explored the theoretical consequences of such a rejection*. This chapter should therefore be of use even to those who have no prior commitment to the uniqueness assumption.

This chapter's objection to the uniqueness assumption is not that motivation can only be measured with a certain roughness, and that more than one estimate of the strength of a reason might therefore be acceptable. Of course this is true, especially for hypothetical motivation. But it is also

[9] For example, Smith (1995), pp. 118–25 believes that the desires of rational agents will converge. Connie Rosati, for the reasons given above, also seems committed to the need for ideal motive theorists to make the assumption, although she is not herself such a theorist. Stephen Darwall almost always writes as if a given consideration either would, or would not, motivate a fully rational agent. See, e.g., Darwall (1983), pp. 134–38. But Darwall is also one of the few theorists who have explicitly recognized the possibility that there might be 'no single correct answer to the question of what reasons there are for a given person to act in a given situation.' See Darwall (1983), p. 241. In fact, however, this claim by Darwall is a consequence of his view that there may be no single correct theory of decision under uncertainty. The absence of such a single correct theory is irrelevant to the question of whether there is a single correct answer to the question of how strong a particular reason is. Rather, it results only in an indeterminacy in what an agent has all-things-considered reason to do.

likely that many ideal motive theorists would tolerate the small range of acceptable strength values that such vagueness implies. Rather, the objection is that there may be a *range* of *clearly* acceptable degrees of motivation that a given consideration could generate, even in an ideal agent. The existence of such a range is different from the existence of vagueness. If two people are told to arrive at a party between 8:00 and 10:00, and one arrives at 8:30 and the other at 9:15, then the reason that both are immune from the charge of being too early or too late is not that there was a certain ineliminable vagueness in the details of the invitation. Rather, both have arrived squarely in the middle of a range of times that are all, in point of being on time, equally 'ideal.' And the specification of this range requires two values.

It is extremely easy to misunderstand the position that results from the rejection of the uniqueness assumption. One might think that the suggested range of acceptable degrees of motivation in an ideal agent will correspond somehow to a range of acceptable degrees of normative strength. But this is a mistaken interpretation. In fact, it is not even clear that it is a coherent interpretation. For what could it mean that a *specific* reason had a *range* of normative strengths? One might take it to mean that the reason could have different normative strengths *on different occasions*. What would this strength depend on? The obvious suggestion is: the actual motivational strength that it provides on each occasion. But this merely confuses normative strength with motivational strength.[10] Nor can one associate the normative strength on a specific occasion with the specific motivational strength that the reason would provide to the agent *under ideal circumstances*. For we are trying to clarify the position that results from the rejection of the uniqueness assumption. And the rejection of the uniqueness assumption is simply the denial that there is any such specific motivational strength.

The correct interpretation of the rejection of the uniqueness assumption is the following. The unique value that was called 'normative strength' is replaced by *two* values. These two values cannot sensibly be called 'minimum normative strength' and 'maximum normative strength.' Rather, they might be called 'minimum rationally permissible degree of motivation' and 'maximum rationally permissible degree of motivation.' When an agent is faced with a decision, the rational options open to him will include all the actions that could result from any combination of rationally

[10] Relatedly, it also confuses the question of the *existence* of a reason, with the question of whether or not an agent *bases* his or her action on that reason.

permissible degrees of motivation – regardless of how likely it is that this particular agent will choose them, given his actual motivational setup.[11] The agent could do *any* of these actions, and be acting rationally. Of course, the agent's action must, in some intuitive way, be *based* on the reasons that make it fall into the set of rational actions, if that *agent* is to be counted as acting rationally, and if the action is to be regarded as *subjectively* rational. And basing an action on a reason may well involve being motivated by that reason. But the question of the existence of reasons, and of their strengths – and therefore the question of the *objective* rationality of an action – is prior to the question of whether or not an agent bases his action on the reasons that make it a rational option. This is also true for accounts that accept the uniqueness assumption, as the following example illustrates. Suppose that the uniquely rational thing for an agent to do in a certain situation is to take some medication in order to save her life. But suppose that the agent takes the medication only because she hates to see pills lying around. This would not make it any the less true that there was a strong reason to take the pills, and that because of this reason the rational thing to do was to take the pills.

By allowing that there might be a range of rationally acceptable options that are independent of (but, barring irrationality, include) what the agent is likely to do, it becomes possible to make the following common sort of claim: it would be perfectly rational for Bill Gates to sell off all his shares of Microsoft, donate the proceeds to relieve the suffering caused by the hurricane in Honduras (and a mass of other suffering), and begin to work on salary for someone else.[12] The fact that someone is very unlikely to perform a certain action does not make it impossible for us to make normative judgments about that action.[13] Perhaps it is true that

[11] The argument of this chapter therefore suggests an internalist position between those which Michael Smith has called 'Humean' and 'Kantian.' For it allows that the normative capacities of reasons can be characterized in an objective way (as the Kantian internalist maintains), and it also allows that different agents might be motivated to differing degrees by the same reason (as the Humean internalist maintains). See Smith (1995), p. 118. Smith's false assumption that Humean and Kantian internalism exhaust the field of possible internalist views is the result of his uncritical acceptance of the uniqueness assumption, and the association of normative strength with unique motivational strength. See Smith (1995), p. 124.

[12] It may be worth noting that this phrase – 'perfectly rational' – as it occurs in real conversation, does not typically indicate anything unique. Rather, it simply indicates that there is nothing *irrational* about a certain action or choice.

[13] Worries that this latitude sanctions radically intransitive preferences, transforming agents into potential 'value pumps,' are addressed towards the end of chapter 7. There I explain and defend the normative significance of a wide variety of formal restrictions.

psychologically impossible actions cannot be *morally* required. But if this particular claim is true, it is plausibly a result of the special link between violations of moral requirements and liability to punishment, and of the practical purposes that punishment serves. On the other hand, and as far as *rationality* is concerned, there is little reason to think that psychologically impossible actions cannot be required. For example, a compulsive gambler, in the middle of a session, may be rationally required to stop gambling, even though we can predict, as well as we can predict any human behavior, that he will not stop.

Perhaps we continue to regard certain actions as rational despite their seeming virtually psychologically impossible because, for practical purposes, it is always at least possible that a given consideration will provide an agent with her motive for acting.[14] But whatever the explanation, it remains true that we can and do make such claims all the time. No matter how surprising someone's action would be, as long as it would fall within a certain range of possibilities, we do not say that it would be irrational for the agent to do it. This is one reason why it is a mistake to try to make the strength of a reason depend upon the strength of the motive it *actually* supplies to an agent on a given occasion. This same mistake also arises when one tries to make reasons relative to something more stable than a passing desire: for example, to a person's *values* or *practical identity*.[15] All such strategies unreasonably narrow the range of actions of which we can correctly say that the agent could perform them (or could have performed them), and be (or have been) acting rationally. Accounts of this kind force us to call actions irrational if they are contrary to the stable values or practical identity of an agent. But we need not – indeed, we would not and should not – count the uncharacteristic altruism of a characteristically egoistic person as irrational.

[14] That is to say, even an account of practical reasons that simply stipulates that certain substantive considerations provide reasons for action will meet what we might call the 'potential explanation' requirement on practical reasons. Of course this does not mean that any given substantive consideration can be reasonably regarded as reason-giving, for there are many other criteria to be met. For a good recent discussion of the explanatory requirement, see Johnson (1999), pp. 58–59.

[15] See, e.g., Copp (1995), pp. 172–85 and Korsgaard (1996a), chs. 3 and 4. Korsgaard may perhaps be able escape this problem by means of her claim (p. 102) that one only has a reason to avoid an action if the performance of that action actually threatens to destroy one's identity. But as she herself admits, this escape comes at a heavy price. It means that one may have no reason of any sort to refrain from even extremely irrational-seeming actions, as long as one's character is sufficiently resilient.

HOW MIGHT THE ASSUMPTION BE FALSE?

It may seem that the very idea of an agent who is ideal or perfect in the relevant respects implies that such an agent would be motivated to some *specific* degree by any given consideration. But this is not true. In order to see how the assumption might be false, we will have to address the naturalistic and nonnaturalistic versions of the ideal motive account separately.

Consider the following simplified ideal motive account of normative reasons

> SIM A fact F is a reason for an agent A to ϕ in circumstances C iff F could motivate A to ϕ in C, iff A were perfectly rational.[16]

SIM is a nonnaturalistic version of the ideal motive account, making use of an ineliminably normative notion of perfect rationality. It may seem impossible that a *perfectly* rational agent would not be motivated to some *specific* degree (the 'most rational' degree) by any given consideration bearing on his choice of action, in any given set of circumstances. And thus, on an ideal motive view, it may seem necessary that the strength of a reason can be given by a single value, perhaps with the admission of some degree of vagueness. But both of these claims are mistaken. One source of these mistakes is the view that perfect rationality is a special sort of psychological state that determines all behavior. But perfect rationality is not a psychological state, especially on nonnaturalistic ideal motive accounts. Rather, it is a normative status, like being a perfect driver. Must we assume that a perfect driver would drive at some *specific* speed on any given stretch of road? Must we assume this, even given specific weather and traffic conditions, etc.? No. Being a perfect driver is a matter of not breaking certain rules, and does not fully determine one's driving style. One of these rules specifies a *range* of acceptable speeds on any given stretch of road. A perfect driver will not exceed the upper limit, or go slower than the lower limit.[17] In the same way, perfect rationality may also be largely a matter

[16] Because the point of the current chapter is not to advance any particular version of SIM, no detailed account of 'perfect rationality' is offered here. Rather, the discussion will rely on particular intuitively plausible claims about the rational status of various actions. It should be fairly clear that the particular judgments that are used in the following discussion are ones that would be assented to by virtually anyone who did not already have a sophisticated philosophical view of practical rationality.

[17] Of course, this is a little simplistic, since there are times when a good driver *would* break the speed limit, and in general the law permits such exceptions.

of not violating certain principles, such as principle P of chapter 3. If so, perfect rationality would not fix the degree of motivation that a given consideration would produce.[18]

Thus we can understand how the uniqueness assumption might be false on a nonnaturalistic version of the ideal motive account. Consider now a naturalistic version of the ideal motive account. Such an account claims that a consideration is a reason for an agent to perform an action in certain circumstances if that consideration could cause an agent to be motivated to perform that action in those circumstances, provided that the agent were also to have some further naturalistic features. According to Brandt, for example, these further naturalistic features would include being fully informed and undergoing a rigorous course of cognitive psychotherapy.[19] Is it a necessary consequence of such an account that any given reason for an agent to perform a certain action in certain specific circumstances would provide a *specific* degree of motivation? Take, for example, the reason one might have to go to France for vacation. The relevant question here is the following: 'How motivated would one be to go to France for vacation, if one were fully informed about what it would be like to go there, and if one underwent a rigorous course of cognitive psychotherapy?' It is very plausible that there is no unique correct answer to this question. This is true even though, if one somehow became fully informed and went through a rigorous course of cognitive psychotherapy, it is at least plausible to hold that one would *then* have some specific degree of motivation to go to France for vacation.[20]

[18] Smith (1994), pp. 166ff has argued that if we hold that there are any normative reasons, then we must also hold that the desires of rational agents will converge. It might seem that if his argument went through, it would show that the suggestion of this chapter must be false. But Smith (2002) has recently clarified his position, and it now includes something he calls 'disjunctive reasons.' The existence of these reasons means that even a fully rational agent, on Smith's view, might desire that his less than fully rational self do any one of a number of mutually exclusive actions, and, importantly, that this desire cannot be understood simply as the conjunction of two (or more) opposing desires of roughly equivalent strength. Thus, even if Smith's convergence thesis is true, it does not seem to exclude, for practical purposes, the view suggested in this chapter. It is a testimony to the attraction of the uniqueness assumption, that Smith seeks to preserve it by introducing desires of such a strange and controversial character. Nor does Smith attempt to defend the postulation of such desires, except by noting that (as they were designed to do) they result in a more acceptable range of rationally permissible actions.

[19] See Brandt (1979), pp. 10–15.

[20] Throughout this chapter I grant the ideal motive theorist the assumption that *actual* people *do* have specific degrees of motivations for such actions. I call this assumption into question elsewhere.

Consider an analogous question, which we can presume to be asked in a certain place at a certain time: 'If I were to drive to Boston starting now, and didn't have any problems on the road, what time would I arrive?' If I *were* to drive to Boston, there is no doubt that the time I would arrive there would be a *specific* time, at least up to the degree of vagueness involved in determining when one has arrived in a city. But this does not mean that there is a specific time that is the correct answer to the question as posed. Rather, a correct answer is more probably something of the following sort: 'Between four-thirty and six.' For the question does not specify the route or the speed at which I would drive, or how long I would spend at rest stops, or a host of other details that might make it more credible that there would be a specific answer. And the question 'How motivated would one be to go to France, if one became fully informed and went through a course of cognitive psychotherapy?' is also missing the sort of details that would make it credible that there was a unique answer.

To make the point even starker, consider another question: 'If I were to go to Boston, what would be the age of the first person I talked to there?' Now, if I were actually to go to Boston and talk to someone there, there is no doubt that the person would have some *specific* age. But this does not mean that there is a unique age that is the correct answer to the question. In fact, plausible correct answers are 'Who knows?' or perhaps 'Between zero and a hundred and fifty.' Similarly, the fact that one would have some specific degree of motivation to go to France if one actually went through cognitive psychotherapy provides no argument at all in favor of the uniqueness assumption. Admittedly, there may well be a connection between going through cognitive psychotherapy and one's degree of motivation to go to France. But this connection need not be sufficiently strong to fix a unique degree of motivation. It may only be sufficiently strong to fix a range of possible degrees of motivation. And the admission that one would have some specific degree of motivation after the psychotherapy provides no argument whatsoever that the connection *is* sufficiently strong to fix a unique right answer. After all, the first person one talked to in Boston would certainly have some *specific* age. What other kinds of ages do people have? And yet it is clear that the answer to the Boston question does not have a unique age as the correct answer, and that there is no connection between being in Boston and the age of the first person one would talk to there.

The general point here can perhaps usefully be made in the language of possible worlds. Consider the question 'How motivated would one be by

the prospect of a trip to France, if one became fully informed about it, and went through cognitive psychotherapy?' In using a framework of possible worlds to answer this question, we first need to consider *all* the accessible worlds in which the antecedent is true: those in which one becomes fully informed and goes through cognitive psychotherapy. Then, looking at the closest of these antecedent worlds, we read off the answer to the question by seeing how motivated one is, in those worlds, to go to France. Because there are a great number of ways in which to become fully informed, and because there are a great number of forms that cognitive psychotherapy can take, it is plausible that there will be a number of closest antecedent worlds. If this is the case, there will be no unique possible world from which we can read off a unique right answer. This is true despite the fact that in *each* of the relevant worlds, one does have *some specific degree* of motivation. It is this last fact that stands behind the misleading truth of the claim that if one were to go through cognitive psychotherapy, one would have a specific (as opposed to a nonspecific) degree of motivation to go to France. But the same sort of fact stands behind the truth of the claim that if one went to Boston, the first person one spoke with would have some specific (as opposed to a nonspecific) age. For, on this sort of analysis of counterfactuals, anything that is true in all of the closest antecedent worlds forms the basis for a true counterfactual claim. The fact remains that the specific degree of motivation in any one of these closest possible worlds may well differ from the specific degree in any other.

It may seem as though the preceding objection to the uniqueness assumption is simply petty quibbling, and that a naturalistic ideal motive account could deal with it by using phrases such as 'other things being equal' or 'in normal circumstances.' But this is false. As Connie Rosati has pointed out, there are many ways in which one might become fully informed.[21] And there are also many forms that cognitive psychotherapy might take. The degree of motivation that a given consideration would generate plausibly depends in a significant way upon these. For example, it has been shown that one's degree of altruistic motivation can depend in a remarkably significant way upon whether or not one has recently found a dime in a payphone.[22] As a result, the naturalistic ideal motive theorist who wants to keep the uniqueness assumption is in the unhappy position of having to claim that there is a specific privileged form of cognitive psychotherapy, and a specific privileged way of becoming fully informed,

[21] Rosati (1995), p. 309. [22] Isen and Levin (1972) referenced in Doris (1998), p. 504.

such that the degree of motivation produced by them, and by them alone, is to be regarded as giving the normative strength of a reason.[23] Moreover, one point of naturalistic ideal motive accounts is to avoid dependence on any prior normative notions. As a result of this, theorists like Brandt cannot specify the acceptable forms of cognitive psychotherapy and full information in terms of acceptable psychological results. In fact, this is the basis of a more destructive objection to naturalistic ideal motive theories. For nothing can guarantee that, after becoming fully informed and after a full course of cognitive psychotherapy – even of some privileged sort – some agents might not persist in having desires that are universally regarded as irrational. These might be the result of brain tumors, chemical depression, stroke, or some other physiological condition. Unfortunately for the naturalistic ideal motive theorist, there does not seem to be much reason to believe that there is a *general naturalistic criterion* that captures all such potential causes of ineradicable irrational desires. Such a general criterion would need to appeal to a normative concept such as 'harm,' or would need to specify certain objects of desire as irrational, which would make reference to cognitive psychotherapy an idle wheel in the account of rational desire.

It seems then that, given the wide range of processes that could count as cognitive psychotherapy, naturalistic ideal motive theorists also should acknowledge the existence of a range of possible degrees of motivation that a given consideration might produce even in an ideal agent.

EVIDENCE AGAINST THE UNIQUENESS ASSUMPTION

So far it has been argued that there is theoretical space for the rejection of the uniqueness assumption in both naturalistic and nonnaturalistic ideal motive accounts. By itself the existence of this theoretical space should be taken to place the burden of proof on those who wish to endorse the uniqueness assumption. For the claim of uniqueness is a much stronger claim than its denial. But at the very least the demonstration of the existence of this theoretical space should clear the path for arguments against the uniqueness assumption. The current section is an attempt to provide one such argument. The strategy is as follows. Ranges have two defining features: a minimum, and a maximum. If we can find a reason for which

[23] For similar claims, see Ripstein (2001), p. 46.

the minimum does not seem to be equal to the maximum, this will provide some evidence against the uniqueness assumption.

Consider the minimum amount of motivation that the prospect of avoiding a certain fairly substantial amount of pain (say, the pain of a badly burnt hand) would cause in a perfectly rational agent.[24] We would expect a rational agent to go to fairly considerable lengths to avoid such an amount of pain: paying a lot of money, for example, or spending all day driving. Indeed, if it were clear that the agent could avoid the pain by suffering a lesser amount of pain that was still quite substantial (say, the pain of a badly burnt finger) we would expect the agent to do so. The minimum level of motivation that the prospect of avoiding the pain of a badly burnt hand would cause in a rational agent is therefore, intuitively speaking, fairly significant. Now consider the *maximum* motivation that would be provided by the prospect of saving a *stranger* from the same pain. It certainly seems rationally permissible, if somewhat saintly, for an agent to be as strongly motivated to prevent someone else's hand from being burnt as he would be to prevent his own hand from being burnt. That is, it would not be irrational to cause one's own hand to be badly burnt, if this were the only means of preventing a similar injury to someone else. Indeed, it would not be irrational to sacrifice one's life to save the life of another, or to risk one's freedom for a chance of gaining the freedom of another. In fact, we can make the following general claim with some confidence.

> M The *maximum* rationally permissible degree of motivation provided by an altruistic consideration is always at least as great as the *minimum* rationally permissible degree of motivation provided by an equivalent self-interested consideration.[25]

M is the weakest claim that guarantees that it would always be rationally *permissible* to suffer some harm oneself in order to prevent the same harm to someone else.

By itself, of course, M is compatible with the uniqueness assumption. After all, the maximum motivation acceptably provided by an

[24] In the following, I will sometimes omit qualifiers such as 'in a perfectly rational agent,' or 'perfectly.' But it is always to be understood that, as SIM specifies, the relevant degrees of motivation are supposed to be those that a *perfectly rational* agent could feel.

[25] 'Equivalent' here means 'involving the same probabilities of the same substantive harms or benefits.'

altruistic consideration might always be as great as the minimum motivation acceptably provided by an equivalent self-interested consideration, simply because these reasons might always cause *the same* unique degree of motivation in an ideally rational agent. Yet, while we would certainly expect a rational agent (let us assume an agent with the financial position of an average university professor) to spend quite a lot of money to avoid a great deal of pain for himself, we would not regard it as irrational for someone to decline to spend this money to spare someone else such pain. We do not, after all, regard it as irrational for people to decline to give money to Oxfam, or to similar famine-relief organizations, when they have no alternate plans for the money, and will only leave it to swell a bank account that they are in no danger of exhausting. If this is right, then the following must be true.

> N The minimum rationally permissible degree of motivation provided by an altruistic consideration is sometimes *less* than the minimum rationally permissible degree provided by an equivalent self-interested consideration.

But it has already been argued that the maximum permissible degree of motivation provided by an altruistic consideration is *at least as great* as the minimum permissible degree provided by the equivalent self-interested consideration. If M and N are both true, then an altruistic consideration must be able to produce a nontrivial range of acceptable degrees of motivation.

One might try to avoid the above conclusion by claiming that the *appearance* of a range of rationally permissible degrees of motivation for a unique altruistic reason is really only the result of conflating two distinct altruistic reasons. True, one might admit, it would not be irrational to save someone's life by sacrificing one's own life. Nor, one might further grant, would it be irrational to decline to give money to famine relief, even though this would also save lives, and even though one had no alternate plans for the money. But this may be true because, when we fill out the descriptions of these actions, we fill them out in such a way that the altruistic reasons are essentially different. For example, in the first case, we may be thinking of sacrificing one's life to save someone in a burning building, while in the second we may be considering the possibility of saving the life of a faceless stranger in a distant land. And it may be that the reason to save a person burning to death right in front of one is different from the reason to save the life of a faceless stranger in a distant land.

Perhaps this is all true. But *even if* the reasons in the two cases are different, this does not argue against the plausibility of ranges of rationally permissible degrees of motivation. For the *maximum* degree of motivation would be, for both of the altruistic reasons in these two cases, essentially the same. This is because any sacrifice that it would be rational to make for the sake of saving someone in a burning building would also be rational to make if it were the only way of saving a faceless stranger in a distant land. Thus, even if one takes the relevance of distance or facelessness seriously, or if one favors some other explanation for why one is not rationally required to donate to Oxfam, it is still hard to defend the idea that there is always a unique rationally mandated degree of motivation for any given reason. On the contrary, suggesting that distance *is* relevant only supports more strongly the view that some reasons can produce a range of rationally permissible degrees of motivation. For the suggestion focuses attention more narrowly on the reason provided by the prospect of saving a distant stranger. And this reason seems clearly to have a low minimum and a high maximum.

In fact, however, distance does *not* seem relevant to the normative strength of a reason. It is true that it seems irrational to fail to take easy means to save someone perishing at arm's length, while it does not seem irrational to fail to send money to Oxfam. But because distance does not seem normatively significant, we should look elsewhere for the source of the apparent irrationality of failing to take easy means help nearby people who are in dire straits. Such a source is not far to seek, and it fits well with an ideal motive account that admits the existence of ranges of acceptable degrees of motivation. Consider what would happen to a person who failed to take easy means to save a nearby person from terrible pain or death. For example, suppose that someone refused to let a gravely injured person use the phone to call an ambulance. If such behavior were to become publicly known, things would almost certainly go very hard for the callous person. This provides an *additional* reason to give the aid. It is this additional *self-interested* reason that is a plausible source of the rational *requirement* to give help in such circumstances, since it appears to be a reason that would produce a reasonably high degree of motivation in an ideally rational agent.[26] Note that when someone's social or legal position is such that they would *not* suffer any significant harm as the result

[26] Of course this does not mean that these self-interested reasons always, or even usually, provide the motivation for such acts. Almost certainly such acts are motivated by concern for the person one is helping. But this does nothing to diminish the force of the argument

of public knowledge of their manifest indifference to (or even their delight in) the suffering of others, then, often, they do not hesitate to show such indifference. Hence the plausibility of the claim that power corrupts.[27] Of course it is grossly immoral for such people to cause suffering, as it is immoral to refuse to allow a gravely injured person to use one's phone. But the moral status of these actions is not the issue here. Rather, the issue is their *rational* status. The conclusion we should draw from these examples is that *by itself*, the prospect of saving someone from pain or death *need* not provide a great deal of motivation, even to a perfectly rational agent. In fact, for all practical purposes it seems that such altruistic reasons might cause virtually *no* motivation in a rational agent. Those who are persuaded by Kantian views will surely protest that this is unreasonable. But given that we regard ourselves and our actions as generally rational, and given that we eat nice dinners, and buy books and music, and go to movies, and so on, while we could spare many people a tremendous amount of needless suffering, this protestation is not very credible.[28] This is, however, consistent with the happier truth that a rational agent might be as strongly motivated to save someone else from pain as she would be to save herself from the same pain. That is, the Godfather was rational, but so was Mother Teresa.

BENEFITS OF REJECTING THE ASSUMPTION

This chapter has so far argued that it is a mistake for ideal motive theorists to assume that a given consideration would provide a unique degree of motivation to an ideally rational agent. This thesis has been defended by example, and also by arguing that the notion of ideal or full rationality

here. The point is that the self-interested reason to give aid in such circumstances plausibly makes it irrational to refuse to do so without a sufficient reason.

[27] It is important here that the sense of corruption is *moral*. We do not have a corresponding saying to the effect that power makes one irrational, for we do not generally regard powerful and untouchable criminals as irrational, at least while they are running their businesses efficiently. Rather, our intuitions are that such people are breaking *moral* rules, and that it is appropriate to *punish* them.

[28] At this point, the notion of imperfect duties sometimes enters the discussion. But whether or not the notion of imperfect duties is entirely coherent or credible, and whether or not it bears on the *rational* (as opposed to the moral) status of actions, it is in any case not very congenial to the view that the rational status of an action is determined by the strengths of the reasons for and against it. And this view, which we might call 'the sufficiency of reasons,' is a natural one for anyone who offers an account of practical reasons that grants unique strength values to normative reasons.

does not imply a determinate psychology, or a determinate set of desires. The view defended so far implies that two different agents might both be rational, and yet one might be, for example, quite altruistic, while the other might not be. It also implies that a single agent might act rationally on two different occasions, even though on the first occasion he acted altruistically, while on the second occasion he acted selfishly. This is in itself a significant and, I think, welcome result, for it comports with our actual judgments of the rationality of actions and agents much better than do views that deny these claims. But the rejection of the uniqueness assumption does more than merely allow ideal motive accounts to better save the phenomena. It also allows them greater precision and variety in their *formal* descriptions of reasons. This is because when they explicitly reject the uniqueness assumption, they will formally characterize reasons with two values rather than with only one. These two values are the minimum and maximum rationally permissible degrees of motivation that the reason can produce. The current section explores some of the benefits of having two values with which to characterize the normative capacities of a reason. The discussion, and the examples used, should make it clear that the benefits stem from the fact that 'minimum rationally acceptable degree of motivation' functions in the same way as requiring strength, and that 'maximim rationally acceptable degree of motivation' functions in the same way as justifying strength.

Consider the following three situations and associated normative claims. Assume in all cases that there are no other significant reasons bearing on the case.

1) An agent finds that she has a certain chance (say, 25 percent) of bringing food and medicine to a group of forty children who need it and would not otherwise get it. But there is also a high chance (say, 75 percent) that the agent will perish painfully in the attempt. In such a case, the agent would be acting in a rationally permissible (if unusually brave) manner, if she decided to try to get the food and medicine to the children. But such an agent would not be irrational to refuse to undertake such a risk.
2) An agent finds two hundred dollars in the street. The agent has recently read in a reputable newspaper that a donation of this amount to Oxfam has a roughly 25 percent chance of being used to provide food and medicine to forty children who need it and would not otherwise get it. In this case, it would be rationally permissible either to donate two

hundred dollars to Oxfam, or to refrain from making the donation. And this is true even if the agent had no alternative plans for the money, and would simply put it into her savings account for indefinite future purposes.

3) An agent with severe allergies finds herself beginning to have a reaction. She is without insurance, but she knows that she can get the required treatment for roughly two hundred dollars at the local emergency room. On the other hand, she also knows that she has a certain chance (say, 25 percent) of riding the reaction out without any permanent ill effects. In this case the agent would certainly be irrational to refuse to pay two hundred dollars to avoid the 75 percent chance of death or permanent ill effects.

If one agrees with the normative claims in the above cases, and if one holds that reasons have a specific strength value and that this value determines, in conflicts of reasons, what is rationally permissible to do, then, as we saw in the previous chapter, it will be hard to deny the following three claims.

1′) The reason provided by a 25 percent chance of saving forty children from the harms of severe malnutrition is of roughly the same strength as the reason provided by the prospect of avoiding a 75 percent chance of premature death.

2′) The reason provided by a 25 percent chance of saving forty children from the harms of severe malnutrition is of roughly the same strength as the reason provided by the prospect of saving two hundred dollars for indefinite future purposes.

3′) The reason provided by the prospect of avoiding a 75 percent chance of premature death is clearly stronger than the reason provided by the prospect of saving two hundred dollars for indefinite future purposes.

As in the example in chapter 5, these three claims are in striking conflict with the expected transitivity of 'of roughly the same strength as' in the domain of practical reasons. But if practical reasons always have one specific strength value, we should expect such transitivity to hold, except perhaps as the result of vagueness. Vagueness, however, does not seem likely to be the explanation of this particular failure of transitivity, since the two reasons involved in (3) are obviously of widely divergent strengths.

On the other hand, if altruistic reasons that have comparatively high maximum rationally acceptable degrees of motivation can also have comparatively low minimum rationally acceptable degrees of motivation, then our intuitions about the three cases are easily preserved. For to say that an altruistic reason has a low minimum rationally acceptable degree of motivation is to say that a rational agent need not be strongly motivated by it. This is why it is rationally acceptable to keep two hundred dollars for indefinite future purposes, even though one knows one might be able to prevent a considerable amount of harm to others with the money. This explains our intuitions about (2). But this is consistent with the claim that such an altruistic reason has quite a high maximum rationally acceptable degree of motivation: sufficiently high that it would be rationally permissible to be more strongly motivated by such a reason than by the prospect of a high risk of painful death. This explains our intuitions about (1). Finally, it is plausible to hold that the maximum rationally acceptable degree of motivation provided by the prospect of saving two hundred dollars will be significantly smaller than the minimum rationally acceptable degree of motivation provided by the prospect of avoiding a high chance of painful death. This explains our intuitions about (3).

Of course there are responses that can be made to defend the uniqueness assumption. But given the arguments of pp. 120–28, which show how and why the assumption might be false, and given the superior explanation of the examples provided by a view that rejects the assumption, the burden of proof should now clearly rest on the shoulders of those who want to defend it. Perhaps the most tempting strategy for someone who wants to preserve the uniqueness assumption is again to try to expand the role of context in determining the content and strength of reasons, as was attempted at p. 126. Following this strategy, one might try to claim that the altruistic reason in (1) is different from and much stronger than the superficially similar altruistic reason in (2), and that this is why the first reason rationally justifies risking one's life, while the second reason cannot even require one to spend two hundred dollars. If this were true, then the three cases would not present a failure in transitivity. But this strategy suffers from the following problem. Unless one independently motivates some rules by which context affects the strengths of reasons, one loses the ability to argue against those who do not share one's intuitions about particular cases. And it is hard to see how one could motivate rules according to which context would affect the strengths of reasons in ways that would preserve our judgments about cases (1) through (3). Why, that is, should

there be a stronger reason to save children from malnutrition when one can only do so by taking a significant risk?[29] The only answer seems to be 'to preserve our intuitions *and* the uniqueness assumption.'

Another strategy for defending the uniqueness assumption from the above argument is to suggest that the normative claims in (1) through (3) *cannot all be true for the same agent*. Since ideal motive accounts can easily make the strengths of reasons agent-relative, there would then be no failure of transitivity. In order to use this strategy, it might be claimed that for the normative claims in (1) to be true of an agent – for it to be genuinely rationally permissible for that agent to risk her life to save a group of strangers – that agent would have to be quite altruistic. And for such an altruistic agent, it might be claimed that the normative claims in (2) would be false. That is, it might be claimed that it would *not* be rationally permissible for such an altruistic agent to keep the two hundred dollars. This sort of objection, which appeals to the fiction of a stable set of motivations that determine the strengths of an agent's reasons, has already been addressed above, at pp. 118–19. To repeat the point made there, it is implausible to suggest that habitually mean and selfish people have no reason to do altruistic actions, or that for such people altruistic actions would be irrational. Perhaps they have no (currently) *motivating* reason to perform such actions. Perhaps, also, it is therefore very unlikely that they will perform such actions. But the human mind is something more complex and unpredictable even than a set of twenty dice. And just as there is a chance that twenty dice, fairly tossed, will all come up six, there is a chance that a selfish person, even without the aid of a stroke or a brain tumor, might see things in an unaccustomed light, and act in an unselfish way.[30] When Sydney Carton does something *far far better* than he has ever done before, this provides not even the hint of an argument that

[29] Or: Why should a significant risk of death provide a weaker reason when one can do a great deal of good by taking it? It is no answer to this question to point out that the reason to avoid the risk is *comparatively* weaker, considered in relation to the reason to provide food and medicine, than considered in relation to the reason to save two hundred dollars. *Comparative* weakness of this sort is admittedly consistent with a unique fixed strength value being assigned to the reason to avoid premature death. But if the strategy being discussed here is to avoid failures of transitivity, then it requires that there be motivated claims about changes in the *absolute* strength values of one or all of the reasons in the examples.

[30] One might be tempted to say that the correct description of what happens in such cases is 'the spontaneous and rationally inexplicable acquisition of an entirely new desire.' Certainly, many cases of sudden and drastic changes in motivation may be the result of irrationality. But in order for this sort of claim to affect the argument here, it must be read as claiming that the *only* way in which any agent could *possibly* be capable of taking

he has therefore done something irrational. And the same remarks can be made about altruistic people acting in uncharacteristically selfish ways.

If one rejects the uniqueness assumption, then cases like (1) through (3) do not force one to abandon the hope of making theoretically useful formal claims about something very much like the strength of reasons. One can even preserve transitivity. All one needs to do is to identify the minimum degree of motivation that a reason could produce in an ideal agent with its *ideal-motive-requiring strength* (or IM-requiring strength, for short). This label is appropriate, since a reason with a greater minimum degree of rationally permissible motivation will of course rationally *require* more of us than a reason with a smaller minimum. Similarly, we can identify the maximum degree of motivation that a reason could produce in an ideal agent with its *IM-justifying strength*. This is an appropriate label, because a reason with a greater maximum degree of rationally permissible motivation will rationally *justify* us in acting against more reasons than would a reason with a lesser maximum. That is, a reason with greater IM-justifying strength will make it rationally permissible to take greater personal risks, and so on. Transitivity of IM-requiring strength and of IM-justifying strength are both consistent with (1) through (3).

An important question is whether some reasons might have a minimum rationally permissible degree of motivation of zero.[31] Given the existence of ranges of rationally permissible motivation, it seems rather arbitrary to assert that the minimum for such ranges must always be nonzero. If there were such reasons, they would provide a counterexample to the internalism requirement on practical reasons. If one uncritically accepts the uniqueness assumption, then this form of reasons internalism may seem trivially true, as indeed it has seemed to many theorists. For if reasons have a unique strength, and if this strength corresponds to the motivation they *would* provide to a rational agent, then any consideration that *could* provide a rational agent with zero motivation *would* provide a rational agent with zero motivation. It would then seem very odd to regard the consideration

both options in all three cases would be for her to acquire a new desire in a rationally inexplicable way at some point. For all that is needed for the argument to go through is the *possibility* that for *one* agent, all the normative claims in (1) through (3) could be true. It may also be worth noting here that the way in which we acquire new desires is typically not by reasoning, or by any intellectual processes at all, and that realistic internalists such as Darwall and Williams allow an important role for imagination, or 'becoming aware of things in a vivid way,' in the rational acquisition of new desires. See Darwall (1983), pp. 39–40, and Williams (1981), pp. 104–5.

[31] The best candidates for such reasons may be *altruistic* reasons that involve the promotion of *benefits* (as opposed to the prevention of *harms*) for others.

as a reason at all. But if one rejects the uniqueness assumption, then zero-minimum reasons could still reasonably be called reasons. For as long as they had nonzero *maxima*, they could account for the rational permissibility of actions that were opposed by other reasons – even by reasons with *nonzero* minima. For example, although it may be rationally acceptable not to lift a finger merely to give a stranger some pleasure, it does seem that one would be rationally justified in going to some pains to do so. That one's action would give pleasure to a stranger may therefore be a reason that violates the internalism requirement.

CONCLUSION

This chapter has not argued in favor of ideal motive accounts of practical reasons. Nor has it even tried to explain what precisely such accounts are best regarded as trying to provide: meaning analyses or truth-conditions, reductions of the normative to the nonnormative, or something else. Rather, the point has been to suggest that *whatever* such accounts are doing, they could do it better if they explicitly rejected the claim that there is a unique degree of motivation that any given consideration would provide to an ideal agent. But if they do reject this assumption, then they must use two values to characterize the motivation a given consideration would provide to the idealized agent: values that specify the minimum and maximum of a range. Discussion of some examples revealed that these two values function in a way that is isomorphic to the way in which requiring and justifying strength function to determine the rational status of actions.

The fact that improved versions of ideal motive accounts give normative reasons features that are isomorphic to requiring and justifying strength supports the view that reasons really do play the two normative roles described in the preceding chapters. The appearance of these features also expands the common ground between the view advocated in this book, and ideal motive accounts of reasons and rationality, and allows ideal motive theorists to understand justifying and requiring strength in their own terms. It is my hope that this common ground will allow for a real philosophical engagement between the two views. However, the same sharing of features that allows for this philosophical engagement may also lead some philosophers to claim that talking about the justifying and requiring roles for practical reasons is just a needlessly complicated way of talking about something that is more intuitively captured by ideal motive accounts. So it may be worthwhile to explain why it is plausible to hold

that the notions of justifying and requiring strength have an explanatory priority over the notions of minimum and maximum rationally permissible degree of motivation.

As was mentioned at the outset of this chapter, ideal motive accounts come in two varieties: naturalistic and nonnaturalistic. The problem with naturalistic versions is that for any naturalistically specified set of conditions, it is possible that an agent might meet them, and yet still have a desire, for example, to scratch the skin off of his legs for no reason. For an agent might have a brain tumor, or a chemical imbalance, or some other physical malady. It does no good here to protest that such accounts are intended to be restricted to rational agents, for we are in search of an account of what rational agents are. Moreover, the prospects are extremely dim for producing a purely naturalistic set of criteria that will sort out chemical and neurophysiological states into those that can be included as part of 'ideal conditions' and those that cannot. Purely statistical criteria will not do. What, for example, would differentiate a statistically rare configuration of the brain that made someone take extra delight in nature from one that made a person self-destructive? One could, of course, use a naturalistic description of the *goals* or *ends* towards which the state inclined the agent: states that inclined agents towards self-desctruction would be disqualified as part of an agent in 'ideal conditions.' But this move gives the game away. For what criteria would we use for selecting the relevant ends? The fact is that we use a normative notion of rationality in sorting psychological states into those that impugn the agent's rationality, and those that do not. Thus, the only plausible form of an ideal motive account will be one that uses an antecedent notion of normative rationality: one that we have been calling a nonnaturalistic ideal motive account. But once we see this, it should become clear that what explains the fact that some considerations yield reasons that have a significant range of rationally acceptable degrees of motivation is the fact that we are using a fundamental principle of rationality with a form similar to those of principles P and Q of chapter 3. That is, what explains the existence of ranges of rationally permissible degrees of motivation is the prior existence of a gap between the justifying and requiring strength of certain reasons.

7

Two concepts of rationality

This book has so far been primarily dedicated to arguing for and explaining a distinction between two normative roles for practical reasons: justifying and requiring. One reason for this is that this distinction is the most controversial aspect of the theory of rationality advocated here. Acceptance of the distinction entails the falsity of a number of extremely widespread assumptions that philosophers make when talking about rationality. But the distinction between these two normative roles cannot be the end of the story. For, as was argued in chapter 4, we should take the notion of wholesale rational status as prior to the notion of a reason for action, and thus as prior to the justifying/requiring distinction as well. The functional role analysis of reasons offered in that chapter took it for granted that we had some way of determining which actions were rational, and which not. So this book would be seriously incomplete without an account of wholesale rational status. Moreover, chapter 4 also claimed that reasons are *directly* relevant only to objective rationality, and not to subjective rationality. Much more remains to be said about these two concepts of rationality. It is the purpose of the current chapter to address these issues, yielding the final account of practical rationality. The two final chapters of the book will then draw out some implications, and explain how the psychology of a rational agent is related to the reasons available to her.

OBJECTIVE AND SUBJECTIVE RATIONALITY

When we point out to someone that she has done something rude or boring, we can expect that at least in some real-life cases she will respond 'So what? I *wanted* to be rude to that person,' or 'Well, *café con leche* is what I like.' But when we point out to someone that she has done something wrong-headed or irrational, this type of response is not appropriate. If someone acknowledges that the kind of action she is contemplating is irrational, then she must also acknowledge that she should not do it – that

the reasons in its favor are insufficient to justify it, even by her own lights. For if the person does offer us a reason for that type of action that she takes to be sufficient to justify it (even: 'it does one good to act like an idiot once in a while'), then she does not regard the action as irrational. Thus, to say that an action is irrational seems to be to say

(1) The action absolutely should not be performed.

On the other hand, there also seems to be a very tight connection between the rational status of an action, and the mental functioning of the agent who performs (or would perform) it. That is, to say that an action is irrational also seems to be to say

(2) If someone performs the action, then something has gone wrong in the practical mental functioning of that person.

Part of the purpose of this chapter is to argue that one notion of irrational action cannot stand immediately behind both (1) and (2). Of course, this claim has already been acknowledged by a significant number of philosophers.[1] But this chapter argues that the relation between actions one should never do, which I have been calling 'objectively irrational,' and actions that indicate a failure in practical mental functioning, which I have been calling 'subjectively irrational,' is of a different nature than is assumed even by philosophers who acknowledge two senses of rationality. In particular, subjective rationality cannot be viewed simply as objective rationality relativized to the beliefs of the agent. Nor can it be viewed as objective rationality relativized to the beliefs that the agent *should* have, given the evidence available to her. Interestingly, it is the distinction between the justifying and requiring roles of practical reasons that is instrumental in demonstrating and explaining the inadequacy of these two attempts to explain subjective rationality in terms of objective rationality. Neither the distinction between subjective and objective rationality, nor the distinction between justifying and requiring, are tremendously complex in themselves. But taken together they yield a view of considerable subtlety and power. It is the purpose of this chapter to explain and defend the view that results from the combination of this pair of distinctions.

[1] See Brandt (1979), pp. 72–73; Rawls (1971), p. 417; Gibbard (1990), pp. 18–19; Harman (1982), p. 127; Raz (1999a), p. 22; Cullity and Gaut (1997), p. 2; Scanlon (1998), p. 30.

WHAT ACTIONS SHOULD ABSOLUTELY NOT BE DONE?

What is it we are trying to express when we call an action 'irrational' in the sense expressed by (1)? If actions that are irrational in this sense should, as a matter of conceptual necessity, never be done, then the question 'Why not be irrational?' should be pointless. For this question to be pointless, it should be analytic that there can never be a sufficient reason or a compelling argument to perform an action that is understood to be irrational in this sense, and that there is in fact always a reason *not* to do it. This gives some clue as to what we may mean when we say that an action is irrational in this sense. We may mean that no one could ever sincerely offer anything as a sufficient reason for such an action. Of course, we can *insincerely* offer reasons in favor of actions that we regard as irrational, and we can even say to someone 'I think you should do this action,' when we regard the action as irrational. For we may not care if the person does something irrational. We may in fact want that person to behave in such a way. This may be because we dislike the person, or because the action will, in some way, benefit us. But if we are being sincere, we cannot say 'You should do the action, for such-and-such reasons' if we regard the action as irrational, in the sense given by (1). Now, we are most often sincere in our recommendations when we are speaking with our friends. Therefore, a good heuristic in thinking about what it is to regard something as irrational in the sense given by (1) is provided by keeping in mind that it should in general not be possible for anyone to recommend an action to a friend if one regards it as irrational, in this sense.[2] But one should not take this heuristic too literally. The real question of relevance is whether anyone could sincerely offer considerations in support of the claim that the agent should perform a particular action. Whether the person is a friend or not is really neither here nor there: it is simply a mental device to help us see more clearly whether an argument in favor of performing the action could actually be offered sincerely. Another way of putting the question is therefore the following: are there features of the action that someone could cite to the agent in support of the claim that the agent ought to do

[2] This heuristic strongly suggests that formal accounts of objective rationality as the maximization of the satisfaction of one's considered and fully-informed preferences are inadequate. For one's friend may be in a state such that her considered and fully-informed preferences would be for extremely self-destructive things. Nothing in the notion of 'full consideration' or 'full information' rules out this possibility. At least this possibility remains unless these notions illicitly import some substantive criteria: for example, that one cannot count as 'fully-informed' if one's preferences continue to be self-destructive. See Sen and Williams (1982), pp. 9–12; Parfit (1986), p. 500.

the action, such that the agent would not be puzzled to understand that the action was being recommended on *those* grounds?

Note that the general *impossibility* of arguing sincerely that someone should do something that one regards as irrational does not imply the general *possibility* of offering an argument or reason in favor of any action one takes to be rational. For not only is it analytic that one cannot offer a reason that one takes to be sufficient to justify an action that one regards as irrational in this sense, but it must also be true that one takes there to be a reason *against* such an action: a reason that places the action in need of justification in the first place. For if there were no such reason *against* an action *A*, then it would be perfectly reasonable to ask 'Why not do *A*?' even if there were no reason (and thus no argument) *for* doing *A*. This is why twiddling one's thumbs, doodling, or deciding to walk clockwise around the block (as opposed to counter-clockwise) are all perfectly rational: it makes sense to ask 'Why not do it?' Now, it may not be possible to *recommend* such trivial actions, since sincere recommendation may always involve having reasons. So if the heuristic offered above – asking 'Could someone sincerely recommend the action to a friend?' – were regarded as providing a definition, such actions would be classified as irrational. But someone could certainly *allow* a friend to do such actions. There is no reason against them, and as a result, they are perfectly allowable. Therefore we should widen the heuristic so that such *allowable* actions are not considered irrational either.

THE OFFICIAL ACCOUNT

The strategy for the official account of rationality is first to describe an attitude that one might have towards a possible action. This attitude, which featured prominently in the preceding section, will be called 'regarding an action as irrational,' and it will correspond to the fundamental sense of 'irrational' given by (1) above. The claim is then made that *objective* irrationality is a property of those actions that are of a type that would prompt this attitude in the overwhelming majority of people. That is, an action type is objectively irrational if, and only if, it prompts this attitude in the overwhelming majority of people, and an action token is objectively irrational if, and only if, it is a token of an irrational action type. The first and most important of these two biconditionals preserves an appropriate vagueness in 'objectively irrational,' since it is a vague matter what 'overwhelming majority' means. That biconditional, however, should not be

taken as giving the meaning of 'irrational action.' Rather, it will be argued that the meaning of 'irrational action' is better understood as given by a substantive reference-fixing definition – that is, by a description of those substantive characteristics in virtue of which an action would prompt the relevant attitude in the overwhelming majority of people. It is, of course, an open question whether the substantive definition is sufficiently accurate – that is, whether it manages to capture, with an allowance for vagueness, those actions that people would call 'irrational' if it were explained to them what this term was meant to capture: a type of action that no one could sincerely recommend. But the potential for error here does not matter a great deal, unless the divergence is significant. This said, the account of regarding an action as irrational is as follows:

> A1 A person *regards an action as irrational* iff that person cannot see, and does not believe that there are, any consequences of the action that could allow someone sincerely to advise someone else to do it.[3]

With this account, it is possible to provide an account of objective irrationality in the following way.

> A2 An action is objectively irrational iff virtually everyone would regard the action as irrational, if they were fully informed about all nontrivial consequences of the action.

I do not mean to beg questions by using the word 'consequences' rather than 'features' in A1 and A2. I use 'consequences' for purposes of clarity, because all the features that appear in A3 below *are* consequences. Should this turn out to be a mistake, I do not believe anything of importance in the rest of the account is affected. Moreover, I grant that there is some plausibility to the idea that there are moral rules – for example, 'Keep your promise' – that can sometimes provide rational justification without the prospect of any of the right sort of consequences for anyone. But the applicability of such a moral rule would, in my view, only provide a rationally *justifying* reason, even if it provided a moral *requirement*.[4]

[3] This attitude is not to be taken in an overly intellectualist way, so that a person would have to *think to himself* that he cannot see any consequences that would allow for a sincere recommendation. It is enough that the person would be puzzled if someone were to try sincerely to recommend it – or even allow it – based on the likelihood of the various possible consequences.

[4] I do not believe that any problems of circularity would result if my account of rationality were to be used in the relevant account of morality. One way of avoiding such a

As was mentioned above, A2 should not be taken as suggesting that to call an action 'irrational' is to assert that there is a nearly unanimous agreement in judgments of a certain sort, any more than to call a banana 'yellow' is to make such a claim. But it is not irrelevant to the meaning of the word 'irrational' that there is such near unanimity. For this near unanimity, because it allows for ostensive teaching of the word, also allows it to share the grammar of objective descriptive words, and to have an objective referent. Making use of this fact, and substituting a plausible extensional equivalent for the right-hand side of A2, we get the following:

> A3 An action is objectively irrational iff it involves a nontrivial risk, to the agent, of nontrivial pain, disability, loss of pleasure, or loss of freedom, or premature death without a sufficient chance that someone (not necessarily the agent) will avoid one of these same consequences, or will get pleasure, ability, or freedom, to a compensating degree.[5,6]

One significant misunderstanding that may arise at this point comes from reading the word 'irrational' in A2 independently of the technical and syncategorematic phrase 'regard as irrational,' as it is defined in A1. If one reads A2 in this mistaken way then A3 is unlikely to seem as plausible

problem would be by explicitly restricting the notion of rationality to the consequentialist version when developing a moral theory. Another more interesting and perhaps more plausible way would be by solving a sort of 'normative differential equation,' the result of which would be the limit of the following series. The first term is the consequentialist account of rationality, the second term is an account modified by moral reasons yielded by the moral rules that emerge from a moral theory that takes rationality to be as given by the first term, the third term is an account modified by the moral reasons yielded by the moral rules that emerge from a moral theory that takes rationality to be as given by the second term, and so on. There is no reason to doubt that such an account would fail to yield determinate accounts of both rationality and morality. This is because the nonconsequentialist reasons will always be of relatively small importance compared to the consequentialist ones, as long as morality, as a system, is rightly regarded as primarily concerned with the welfare of sentient beings.

[5] My debt to Bernard Gert is very great here. See B. Gert (1998), chs. 2 and 3. At the time I incurred this debt, Gert did not make the distinction between objective and subjective rationality, or formally distinguish the justifying and requiring roles of reasons. He has since modified his view in response to the arguments presented here.

[6] The phrases 'a *sufficient* chance' and 'to a compensating degree' of course retain a normative aspect. These normative phrases could be eliminated by specifying, descriptively, the types of trade-offs that the overwhelming majority of people regard as rational. I have not eliminated them because to do so would make A3 too cumbersome to serve its illustrative purpose. For this same reason, A3 does not reflect the possibility that egregious harms to others may sometimes make an action irrational.

as it is here claimed to be. For there is a significant group of philosophers who, if fully informed, would nevertheless regard it as perfectly rational – in a sense *different* from that given in A1 – for an agent to chop off her little finger, given that the agent had a sincere desire to do so, and no opposing desires of sufficient strength. Humeans such as Bernard Williams hold such a view.[7] Now, there are a whole range of arguments against views that invest desires with this kind of normative significance. For example, Joseph Raz, Jonathan Dancy, Thomas Scanlon, and Warren Quinn have all recently argued that it is not desires that have normative relevance, but the reasons behind them, and that if there are no reasons *behind* one's desires then those desires are whims at best and pathological at worst.[8] These arguments are very persuasive, and they certainly support the general view offered here. But in fact they are not relevant at this point. Rather, what is important is to recall that we arrived at A1 by analysis of what we mean when we use terms like 'irrational' in a certain fundamental way: the way indicated by (1). The present account relies on the supposition, defended below, that there is sufficient agreement in what people regard as irrational (in the relevant sense) to yield an objective matter of fact as to what kinds of actions really are irrational – in the sense of 'objectively irrational' – and what kinds are not. When 'regard as irrational' is understood in this way, A3 is extremely plausible. Additional strong support for A3 will come from discussion of subjective irrationality: the kind of irrationality that has a much closer connection to practical mental functioning, moral responsibility, free will, and so on. For there will be a simple account, in terms of A3, of subjective irrationality. The plausibility of this account will provide additional support for A3.

As we saw in chapter 1, any account of subjective rationality should be consistent with the idea that a fully informed agent, performing an objectively rational action, will typically be performing a subjectively rational one also.[9] Call this the 'objective-subjective implication.' Against A3, and making use of this implication, many theorists have claimed that all actions that harm others for morally insufficient reasons are not only immoral, but are also irrational; in other words, that moral requirements are also

[7] See Williams (1981), p. 105.

[8] Quinn (1995), p. 195; Dancy (2000), pp. 35–38; Raz (1999b), pp. 50–64; Scanlon (1998), pp. 35–42.

[9] More accurately, such an agent must at least have the possibility of thereby performing a subjectively rational action: such an action should only be subjectively irrational if it is done 'for the wrong reasons.' This qualification will not be important in what follows.

rational requirements. A3, taken together with the objective-subjective implication, certainly is inconsistent with this claim, for it allows it to be subjectively rational to hurt other people for profit. Ultimately the best argument against this view will be the greater adequacy of the view offered here, taken as a whole.[10] But more pointedly, it can also be urged that this objection simply goes against the way in which the notion of 'irrationality' is understood by competent language speakers. Of course I do not mean to appeal to intuitions about the use of the very word 'irrational,' much less to the phrase 'subjectively irrational.' The first of these is rarely used by normal people, and the second is a technical term. Rather, I mean that it is very plausible that there is overwhelming agreement if it is understood that we are using the term 'subjectively irrational' to categorize those actions that involve a kind of failure in practical mental functioning that is relevant to questions of moral responsibility and so on, whether or not those failures are sufficiently extreme to have much importance in particular cases. That is, 'subjectively irrational' is meant to collect the spectrum of actions that range from 'silly' and 'stupid,' through 'boneheaded' and 'a bad idea,' all the way up to 'crazy,' 'insane,' and worse. We generally would not say, for example, that it is *irrational*, in this sense, to embezzle money, or to cheat on one's partner, or even directly to hurt others by acting on vengeful impulses. Or, if we do say that such actions are irrational, the reason we offer is almost always 'because you might get caught,' and almost never 'because it might hurt someone.' In general we only say that an immoral action is irrational *in virtue of the harm it does to others*, if that harm is very significant. For example, we might well say this if we discovered that our friend had someone tied up in the basement, and was preparing to cut off that person's toes merely as an experiment. But though these types of actions receive a disproportionate amount of attention in philosophical and fictional literature, they are far from the typical cases of immoral action. In typical cases it is the applicability of a reason like the following – that one's friend may get caught and punished – that explains why one feels the need to cite some justifying reason in recommending an immoral action to anyone. But such a reason need not be of the right kind to *morally* justify the action. Of course, the rational permissibility of a great deal of petty immoral action does nothing to argue against the view that morally required action is also, and always, rationally permissible. A3,

[10] See also Svavarsdóttir (1999), pp. 161–219. Svavarsdóttir makes use of many of the same intuitions about the rationality of action that this book has relied upon.

and the associated account of subjective rationality, does not speak in any way against the rationality of acting morally; it only speaks against the necessary *ir*rationality of acting *im*morally.

A more plausible but equally flawed argument that moral duties provide altruistic rational requirements comes from considering possible harms to one's family or friends as a result of one's actions. Suppose, for example, that someone with a fairly high-paying job and a good deal of savings decides to give away virtually all of his money, to quit his job, and to get a new job at a very low salary, working for a nonprofit organization that promotes the availability of contraception and prenatal care in Third World countries. Though this behavior sounds extreme, it does not sound irrational until we add the fact that the man has a wife and children towards whom he has specific duties, and who will suffer a great deal as a result of his 'altruism.' This might seem to suggest that specific duties to others can provide altruistic reasons of substantial (rationally) requiring power. But this example is not as simple as it looks. First, it may well be that in some cases the best explanation for this sort of behavior is that the agent is suffering from a mental illness: that may be the best explanation for many drastic changes in behavior. In such cases the action may be irrational, *in the sense of proceeding from a failure of practical mental functioning*. But not all such cases need be explained in this way, and in any case we are not considering this sense of 'irrational' at the moment. Moreover, even the suspicion that such action is irrational in this latter sense depends covertly on the assumption that the agent cares a great deal about his wife and children. If we explicitly reject this assumption, and are really convinced that a father has grown so estranged from his wife and children that his duties to them, though real, are felt by him to be rather a burden, then his extreme altruistic behavior towards strangers, described above, no longer seems irrational even in the 'mental functioning' sense. It only seems to be tinted with an uncommon degree of coldness and cruelty. If we take care to separate questions of proper mental functioning from questions of what actions can be sincerely recommended, then 'cruel-altruistic' action no longer seems irrational in the fundamental sense at issue here. For the question of relevance is: could someone concerned for the agent sincerely recommend to him that he allow his own family to suffer for the sake of nameless strangers? And the answer is 'Yes.' Peter Singer, for example, has recommended exactly this.[11] And even the people

[11] Singer (1972), pp. 229–42.

who do not act as he suggests are not puzzled when they consider the basis on which Singer makes his recommendation. That is, people do not regard such action as irrational, in the sense of regard-as-irrational given in A1.

It is important to note here that no *argument* is being offered here for the correctness of the substantive description of those actions that one can sincerely recommend. The above discussion is only attempting to *describe* the wide range of actions which, on consideration, we think people can and cannot recommend in this way. The current aim is to get the *extensions* of the concepts 'objectively rational action' and 'objectively irrational action' right. One of the main features of the description is that the existence of compensating (or even more than compensating) benefits for others does not generally prevent us from making comprehensible recommendations of action in the face of the reasons those benefits provide. No argument is being offered for this description, because if it is correct, there will not be any argument possible. To what normative principle could such an argument possibly appeal? Any principle that was not exactly equivalent to the correct description would be wrong. But any principle that *was* exactly equivalent would be exactly equivalent, and therefore could not play a justifying role in an argument for the correctness of the description: it would just be a repetition of the description. When we are making accounts of rationality as the fundamental normative term, a correct description of the fundamental way we should always behave is the end of the normative road. It should not surprise us that this is the correct method for *fundamental* normative theory, even if it is not the correct method for, say, moral theory. For there cannot be any *more fundamental* normative principles that could ever provide a normative argument in favor of *fundamental* normative principles. On the other hand, moral theory generally does appeal to more fundamental notions. At bottom, 'objectively rational action' must simply be described, based primarily upon the way normal language speakers learn to use the relevant normative words like 'reason,' 'makes sense,' and 'recommend.' If someone acknowledges that, given the meanings of these terms, there is a reason against performing an action – a reason of the sort that it makes no sense to act against without some countervailing reason – and that there is no countervailing reason that anyone could sincerely recommend acting on (so that, at least in that person's estimation, the action is objectively irrational), then she cannot ask 'But is it really true that there is a reason against it, and that it would make no sense to do it?'

The above remarks might be summarized as two claims:

(a) there are no untaught-but-intuitively-accessible meanings for normative words like 'reason,' 'justify,' or 'ought' that make it possible to ask whether it is *really* true that the things that everyone takes to provide reasons *really* provide them, or whether the reasons that everyone takes as providing sufficient justification for a comprehensible recommendation *really* do provide such justification;
(b) nor, since we are trying to provide an account of the *fundamental* normative principle applicable to action, is there any *more fundamental* normative principle to which one might appeal.

These two points justify the heavy use of examples in this book, since those examples might really be used by parents when teaching their children normative concepts.[12] Of course it is possible that certain isolated parts of the actual use of terms like 'justified' and 'reason' could reasonably be repudiated by normal language speakers. For example, we might see that the way these terms have been taught has been influenced by the social or racial climate, and that we therefore have a good explanation for the inclusion of some considerations as 'reasons' that really do not fit the broader pattern of use. But these sorts of local corrections to the accepted meanings of such terms cannot undermine the general structure and content of the concepts as revealed in actual considered usage.[13]

It should not be thought that appeal is being made here to some principle of the following sort: if the overwhelming majority of people regard an action as irrational, that makes it irrational. That would be quite a dubious normative principle. Moreover, it would leave the account open to a standard objection that is raised against overly simple response-dependent accounts of notions of all sorts. In this case the objection would be that if we all suddenly went crazy, perhaps as the result of a psychoactive gas from outer space, this would be completely irrelevant to the rational status of actions. One standard response to this objection is to rigidify the account by the inclusion of the word 'actually.' In this case, the word 'actually' could be inserted, for example, before the word 'would' in the right-hand side of A2, yielding the following:

[12] Again, it does not matter if the examples go against popular theoretical views that imply, for example, that immoral behavior is really irrational. Such views cannot be right if the method advocated here is correct.

[13] See Raz (1999b), pp. 161–81 for arguments in a similar spirit, that allow only limited improvements in our understandings of basic normative notions.

AA An action is objectively irrational iff virtually everyone actually would regard the action as irrational, if they were fully informed about all nontrivial consequences of the action.

With this modification, 'irrational' would continue to refer to its *actual* referent even if the responses of the overwhelming majority of people should change from what they *actually* are at the moment.[14] This move, however, though popular, and though it does save response-dependent accounts from the objection, seems motivated entirely by the need to avoid the force of the objection. Otherwise it seems ad hoc. Why, that is, should the concept of 'irrational action' (or other response-dependent concepts, such as 'yellow') privilege the *actual* responses of human beings over those of the future, especially since we will be using those same words in the future, when our then-actual responses may have changed? The current account does not provide an answer to precisely this question, since it denies that any reference to the responses of human beings, actual or otherwise, is part of the meaning of 'irrational action' (or of 'yellow'). Nevertheless, the current account does explain why there is a connection between the overwhelming responses of *actual* human beings, and the referent of 'irrational action' (and of 'yellow'): it is actual human beings who have participated in the processes whereby people have come to learn the meanings of these words.

Thus, on the current account A2 is not a meaning claim, and is not asserted to be analytic. If it were, then we should have to suspend our judgment as to the rational status of certain actions in all cases in which we were unsure whether our views were shared by the overwhelming majority of other people.[15] But the overwhelming agreement of other people, though relevant to the rational status of actions, is not relevant in this way. Rather, the fact that there happens to be overwhelming agreement in what people regard as irrational (again, in the sense given by A1) is what allows for the ostensive teaching of the concept of irrational action, and the related concept of a normative practical reason. And once the meaning is taught, people can rely on their spontaneous judgments of rationality just as reliably as they can rely on their spontaneous judgments of color. The meaning of 'irrational' contains no more reference to the existence

[14] For the canonical explanation of this use of 'actually,' see Davies and Humberstone (1980), pp. 22–25.

[15] This seems to be the sort of view David Lewis has in mind with respect to the notion of value. See Lewis (1989).

of overwhelming agreement than does the meaning of 'yellow.' Rather, the meaning of this word can be explained by reference to the class of actions that members of the linguistic community would use in teaching the concept. A3 specifies this class. That means that the meaning of 'irrational' is more adequately given by A3 than by A2. Rather than giving the meaning of 'irrational,' A2 is an important part of what, following Philip Pettit, we might call the 'genealogy' of the concept of rational action. But it is a substantive description, such as A3, that comes much closer to giving the meaning.

'But I don't want to know what comes *close* to giving the meaning, I want to know *the meaning*.' This seems to me a misguided request. One explains a concept when one explains how to use a word in a certain way. This is why verbal definitions often serve to explain concepts and give the meanings of words. But often such purely verbal definitions are not sufficient – as in the case of 'yellow,' for example, and other concepts taught primarily by ostension – and then other explanations are required. I am claiming that A3 is much more useful in such an explanation than A2 would be. But by itself A3 might well be insufficient. It will then be useful, in explaining the concept of objectively irrational action, to mention that such actions are the actions to which one typically will have the attitude given in A1. The possible usefulness, in explaining the concept of an objectively irrational action, of mentioning the attitude of 'regarding as irrational,' may be what explains the attraction of expressivist views such as those of Allan Gibbard and Simon Blackburn. These philosophers seem content with claims such as A1: claims that limit themselves to describing the attitude one typically (they might use a stronger word) expresses when one uses a normative word. But such an account, unsupplemented by claims like A2 and A3, cannot differentiate between response-dependent words such as 'funny,' for which the criterion of correct use involves the possession of the right response by the speaker, and response-dependent words such as 'yellow,' for which the criterion involves picking out the correct items. There is a fact of the matter with regard to yellowness, in virtue of the overwhelming agreement of people in their visual responses, so that anyone who does not share this response is called 'color-blind.'[16]

[16] This fact of the matter is consistent with there being no fact of the matter, with regard to certain yellows, as to whether or not they are slightly greenish or slightly reddish. For an opposed view, see Hardin (1993), p. 91. It seems to me that Hardin makes too much of the fact that very fine-grained distinctions can be made differently by people who are equally 'normal.' It is possible that there is no fact of the matter as to whether the color

A similar claim is *not* true for 'funny,' since you and I can find different things funny, without either of us having to regard the other as wrong. But the overwhelming agreement in response to irrational actions (and as will be seen presently) to harms and benefits means that people who fail to have the typical response can be identified, and their responses labeled defective. When the degree of agreement allows for this to happen, this permits the development of an objective notion for which an expressivist analysis will be wrong.[17]

To return to the description: when we call an action 'irrational' in the fundamental sense described above, we take it that the action involves a risk of harm to the agent without a compensating benefit to anyone else.[18] Put this way, it should seem plausible that we have indeed reached the end of the normative road. For there does not appear to be any answer to the question 'Why shouldn't I act irrationally, in that sense?' One could of course answer 'Because such action will harm you without benefiting anyone else.' But that is no *further reason*. It is simply a repetition of the definition. It only sounds like an answer because it would have been an answer if the question had been 'Why shouldn't I do *this particular* action?' in a case in which the particular action turned out to be irrational.

As has already been mentioned, the above response-dependent account of objectively irrational action is similar to response-dependent accounts that might be given of other notions that are objective despite being relative to human nature, such as accounts of what is yellow, poisonous, comfortable, sweet, and so on. There will always be a small minority of people who are not harmed by certain substances, who dislike a certain room-temperature, who think oranges are the same color as lemons, or who cannot taste the difference between sugar and salt. The existence of such people does nothing to falsify the claims that arsenic is a poison,

of a fire engine perfectly matches a particular tristimulus value, while there is a fact of the matter as to whether it is *red*.

[17] I explain this in more detail in J. Gert (2002a), in which I use the 'possession of the attitude by the language learner' criterion to argue against an expressivist analysis of 'morally wrong.' Michael Ridge has presented interesting counterarguments to this particular illustration of the criterion, but his arguments do nothing to diminish the force of the points here, as applied to the concepts of irrational action, reason, harm, or benefit. Compare B. Gert (1998), pp. 90–91, in which the objectivity of yellowness is explained in terms of 'standard conditions' and so on. This common method of ensuring the objectivity of response-dependent concepts waits on an analysis of 'standard' in a way that my statistical method does not.

[18] I use the normative words 'harm' and 'benefit' here as a shorthand for the sort of consequences mentioned in A3. This use is justified below.

that seventy degrees Fahrenheit is a comfortable room temperature, that oranges are not the same color as lemons, and that sugar is sweet. This is true despite the fact that the tiny minority may sometimes be the only group to hold some true belief. Let us suppose, for example, that there was a time when the overwhelming majority of people believed that the Earth was flat. This did nothing to the shape of the Earth, or to the falsity of the claim 'The Earth is flat.' But when there is an overwhelming agreement in judgments that have no independent method of verification, or about which the only arguments will be those initiated by the skeptic, then the judgments or reactions of the tiny minority are not rightly regarded as legitimate alternate views, but are, rather, to be regarded as wrong. It cannot be, for example, that the vast majority of people are wrong in thinking that 'hello' is a greeting, that grass is green, that sugar is sweet, or that it is objectively irrational to risk the loss of one's arm if no one is going to benefit.[19]

A2 and A3 can provide an account of harm and benefit in the following way. The class of actions that are objectively irrational, according to A2 and A3, is pretty well fixed. Therefore, even if we did not have the normative concepts of harm and benefit, we could notice, as A3 claims, that objectively irrational actions are those that involve risk to the agent of pain, premature death, disability, and so on, without a certain chance of bringing someone pleasure, freedom, and so on, or of helping someone to avoid pain, premature death, disability, and so on. If we did notice this descriptive structure, we could *invent* the notions of harm and benefit, and of a benefit compensating for a harm, in the following ways.

> *Harms* are the *types* of consequences of an action, to an agent, that can make the action objectively irrational: i.e., death, pain, etc.[20]
>
> *Benefits* are the *types* of consequences of an action, to an agent, that can prevent the action from being objectively irrational,

[19] It is no objection to this sort of response-dependent analysis of rationality that, whereas blueness and sweetness are descriptive properties, rationality is a normative one. For the point here is that rationality, in both senses at issue in this chapter, *is* a descriptive property, *as well as* a normative one. The normativity comes from the nature of the subjective response mentioned in A1, which is linked to human motivation in a way that having a subjective experience of blueness or sweetness is not.

[20] The reference to the agent here does not prevent us from talking about harms that *my* actions might do to *you*. For this definition of harm serves to pick out certain substantive consequence *types*. The same is true for definitions of benefit and of compensating.

> in those cases where the action would be objectively irrational without those consequences: i.e., pleasure, ability, reduction of pain, etc.
>
> *Compensating* benefits, relative to certain harms, are benefits that actually *would* make it so that it was not objectively irrational for an agent to suffer the harms: i.e., the unpleasantness of running five miles is compensated for by the pleasure one feels afterwards, and by the contribution the running makes towards increasing various of one's abilities.

Unsurprisingly, given the extreme usefulness of the concepts of harm, benefit, and compensating benefit, we did not have to wait to invent them. Defining 'harm' and 'benefit' in terms of objective irrationality allows us to understand why avoiding a harm always counts as a benefit, but why not getting a benefit does not always count as a harm. It also explains why it may be a harm to *lose* a certain benefit, when it would not be a harm to be prevented from *getting* it.

It may seem implausible to some readers that notions such as rationality, harm, and benefit are as objective as the notions of sweetness or yellowness.[21] For it may seem that there simply will not be any actions of which it is true that '*virtually everyone* would regard the action as irrational, if they were fully informed,' as A2 claims. But the strength of this objection should be significantly reduced by attention to the following three points.

First, there is vagueness in the notions of sweetness and yellowness, just as there is vagueness in the notions of objective and subjective rationality, and harm. Everyone admits that there are irresolvable disputes as to whether a certain substance is sweet, or a certain object yellow. Therefore, similarly irresolvable disputes as to whether a certain type of action is objectively irrational, or whether a certain consequence counts as a harm, are not in themselves any argument against the objectivity of those notions. It is in fact a virtue of the above account that it does not contentiously impose an unrealistically rigid precision on the relevant notions. As long as the disputes are sufficiently marginal, they can be attributed to the vagueness inherent in almost all objective notions.

Second, the disputes *are* marginal, *especially* with regard to the question of what counts as a harm or a benefit. There is overwhelming agreement

[21] Precisely how objective these notions are is a matter of philosophical interest, but is not relevant for present purposes. I will be satisfied to convince readers that it is just as much a matter of fact that pain is a harm as that sugar is sweet.

that pain and death, for example, are harms, even though there may not be such overwhelming agreement on the question of precisely how much pain it is worth risking in order to avoid a certain chance of death. One need not agree that a very painful treatment is worth a 5 percent chance of avoiding a premature death, in order to agree that the choice involves a balance of harms. That is, virtually everyone agrees that one absolutely ought not take the treatment if it has no chance of increasing one's life-span (or providing some other benefit). And virtually everyone agrees that one absolutely ought not refuse the treatment if it were completely painless and cost-free (unless living the extra time would bring with it some other cost).

Third, the disputes are marginal even with regard to the question of the relative strengths of reasons, and thus with regard to the question of the objective rationality of particular types of actions. To see this, it is important to keep in mind that such agreements are not the same as, and do not entail, agreement in what one actually does, recommends, or even allows. It is possible to recommend one option while recognizing that many other options – even options one would recommend that a friend *not* do – are not objectively irrational. You and I might agree, for example, that there is nothing objectively irrational in giving up a high-paying job in order to do less pleasant but more socially beneficial work. And yet you might choose to make this sacrifice and urge me to do the same, while I might continue to draw my six-figure income as a computer consultant, and press you to act similarly. That there is overwhelming agreement in what kinds of actions are objectively rational is shown by the fact that almost all of us regard as rational (in the sense given by A1) almost all of the actions that almost all people actually perform, whether we approve of them or not. That is, almost all of the different career and lifestyle choices, almost all of the choices that medical patients make, either to undergo or to refrain from treatment, almost all of the mundane choices of what to eat, wear, watch, say, and so on, are ones for which we can see the reasons, and which do not mystify us or call for the special explanations that actions typically require when we regard them as irrational.[22]

[22] Thomas Scanlon describes something very similar to the attitude of seeing the reasons that other people have, without being moved by them oneself. See the Appendix to Scanlon (1998). But because Scanlon takes the notion of a reason as primitive (p. 17), he is forced to take the idea of 'counting in favor of an action' as univocal, and he cannot distinguish justifying power from requiring power. As a result, his univocal interpretation is essentially 'requiring' (p. 61). Thus he cannot explicitly say that, without some sort of

RATIONALITY AS RELATED TO PROPER MENTAL FUNCTIONING

The above account of objectively irrational action makes it objectively rational to make great sacrifices for others, and it also makes it objectively rational to ignore the interests of others to a very great degree. But according to A3, an action that is likely to hurt the agent a great deal, but which is also likely to save someone else's life, would be classified as objectively rational, *even if the agent were completely unaware* of the potential benefits of the action. This seems wrong in an account of rational action, if such an account is meant to capture something about proper practical mental functioning. But this is fine. A1 through A3 do not provide an account of rationality in this sense at all. Rather, the class of actions A3 describes includes mistaken action that is harmful to the agent. Since A2 and A3 emerged at the end of an argument that started with the idea that there is a class of actions of which it should make no sense to ask the question 'Why shouldn't I ever perform actions that belong to that class?' it should be clear that *objective irrationality* is the fundamental normative notion applying to action.

There is, however, a notion that is more properly termed 'rationality,' and which has an intimate connection with proper practical mental functioning. Of course the use of the phrase 'proper mental functioning' here has nothing to do with the contingent 'purposes' served by human reason: purposes that might figure in an evolutionary or theological explanation of why our reasoning abilities have their current shape. Rather, the current account is an analysis of the concept of 'irrational action' as it is used by philosophers, doctors, lawyers, and others, in a way that is intimately related to our assessment of the agents who perform them, and that is therefore related to the notions of moral responsibility, freedom of the will, competence to give consent, and so on. The link with these other important notions should always be kept in mind when rival accounts of rationality are offered, especially Kantian accounts according to which it is irrational to be immoral. Perhaps there is *some* sense of 'irrational' that these accounts capture. But if there is, it is a sense that has very little to do with moral responsibility, competence to give consent, self-control, and so on. One main point of this chapter is to distinguish the notion of objective

deficiency, one can see that someone else has reasons, and that one is in the same relevant circumstances, and yet remain unmoved. Instead, the most he can say is that one can see that one has, oneself, reason "not to scorn" the ideals that someone else takes to provide him with reasons, and reason "not to mock those who take it seriously" (p. 370).

rationality from the notion of rationality in this latter sense, and to provide an account of the relation between the two. In general, philosophers have used the term 'rationality' to refer to both of these notions. When I use the term without qualification, I should always be understood to be talking about the subjective notion.

One might think that there is a simple definition of subjective rationality in terms of objective rationality, which runs as follows:

> R1 An action is subjectively irrational if, relative to the beliefs of the agent, it is objectively irrational.[23]

Such an account would correctly make it subjectively irrational to do something that one knew would be harmful to oneself, even if, unbeknownst to one, it were very likely to save someone else's life, and therefore even if the action were objectively rational. Thus R1 accounts for the type of case that shows that A3 does not capture the notion of rationality that is related to proper mental functioning. But such a simple definition will not work. One problem is that an agent may conspicuously *lack* a belief: for example, the belief that her action will be extremely harmful to her. If the agent *should* believe that an action will be harmful in this way, but does not, then that agent's action may well be subjectively irrational.[24] But R1 will not classify it as subjectively irrational, since R1 relativizes to *actual beliefs*. For example, suppose that I believe that I can fly, and therefore that I will not fall to my death when I jump off of the roof of my apartment building. Despite the fact that I do not believe I will be harmed by jumping, it is still subjectively irrational to jump.[25] Why? Because I *should* believe that I will be harmed. In response to such cases one might try to patch up the definition in the following way:

[23] See Brandt (1979), pp. 72–73; Gibbard (1990), pp. 18–19; Harman (1982), p. 127; Raz (1999a), p. 22.

[24] For present purposes, 'should believe' can be taken to mean something like 'could be faulted for not believing.' It does not mean 'would be delusional not to believe' – although of course delusional agents will also often fail to believe what they should, in the relevant sense, believe. Worries about the sense of 'should' at issue here will not carry over into the account of rationality offered below, since that account does not refer to what the agent should believe. In fact, the account can be used in a straightforward way to give a relatively clear content to this important sense of 'should.'

[25] This example was chosen for vividness, but one might also be acting irrationally, although less drastically so, in eating another oyster. Eating the oyster would be irrational if one should (but does not) believe that it will cause the same distress that it typically does.

R2 An action is subjectively irrational if, relative to the beliefs that the agent *should* have, it is objectively irrational.[26]

But suppose now that an agent *should* believe that her action will save someone's life, but that the agent does *not actually* believe this. If the action is one that she *should and does* believe will be very painful, then it would clearly be irrational for her to do it, for she does not believe it will have any compensating benefit. Yet definition R2 will not classify it as irrational. What is going on here? Our strategy, in definitions R1 and R2, for determining the subjective rationality of an action has been first to relativize the action to a set of beliefs, and then to determine its objective rationality. But it seems as though we cannot relativize to the *actual* beliefs of the agent, because of the relevance of beliefs that the agent should have. But we cannot relativize to the beliefs that the agent *should* have either, because of the relevance of the beliefs that the agent actually does have. And of course one cannot simply relativize to both, since there may then be no univocal verdict in cases in which an agent has one belief about the consequences, but should have a different one.

At this point one might make the following suggestion, which, since it seems to follow from some remarks of Hume's, we can call 'the Humean suggestion.' This suggestion admits that, speaking in a rough and inaccurate way, we often call an action irrational if it is based on an irrational belief. But *strictly* speaking, the suggestion continues, irrational action is captured perfectly well by R1. That is, this suggestion holds that someone's jumping out of a window, irrationally believing that he can fly, may nevertheless be a perfectly rational *action*. It is just the *belief* that it is based upon that is irrational. There are a number of responses to the Humean suggestion. The first response is that one should not lose sight of the ultimate goal of the account of subjective irrationality here. We are trying to capture a sense of 'irrational' that already exists, and that bears a close relation to freedom, moral responsibility, disabilities of the will, competence to give consent, and so on. It is easy to become diverted from this purpose by the attractions of conceptual simplicity. One possible result of such a diversion

[26] This definition follows a pattern used by Rawls in defining what he calls 'subjective rationality' in relation to what he calls 'objective rationality.' See Rawls (1971), p. 417. See also Cullity and Gaut (1997), p. 2. Cullity and Gaut relativize practical rationality not to what we actually believe, but to *what we are rationally justified in believing*. This last phrase is ambiguous as between 'what, given the evidence, we would be rationally justified in believing (whether or not we actually believe it)' and 'those of our actual beliefs that are also rational.' Either choice will encounter problems.

is that one ends up defining two 'cleaner' notions of irrational action, the first of which is R1, and the second of which is 'action based on irrational belief or culpable ignorance.' Call the first 'acting on flawed desires' and the second 'acting on flawed beliefs.' Neither of these notions does what we want. For example, when considering questions of moral responsibility, it is often a matter of indifference whether someone's irrational behavior is the result of an affective disorder, or of delusions.[27] Moreover, not all actions that are based on irrational beliefs are viewed in the same way, with reference to moral responsibility, free will, etc. If I irrationally believe that wearing green brings me small pieces of good luck, and for that reason I wear green, this action would not typically be regarded as irrational at all. And if it were, it would not be regarded as being nearly as irrational as my jumping out of the window, thinking I could fly. This difference in the rational status of the action has nothing to do with the degree of irrationality of the belief.

The second response to the Humean suggestion is directed against one of the worries that inspires it. The worry is that unless we take the suggestion, our account of irrational action is going to be *drastically* inelegant. That is, one might worry that unless we divide the actions we call 'irrational' into the two suggested classes, the resulting account will include reference to an unwholesome mix of actual beliefs and beliefs that the agent should have. Readers who are sympathetic to the Humean suggestion for this reason are advised to wait until the official account is offered below. That account captures the differential relevance of actual beliefs and beliefs that the agent should have. But it does so in a way that is neither ad hoc nor disjunctive, and it also does so in a way that finds a parallel in theoretical rationality.

Let us return therefore to the failures of R1 and R2. The problem, again, is that we can relativize neither to the *actual* beliefs of the agent, nor to the beliefs that the agent *should* have. In seeing why this is so, it is useful to note that it is a matter of importance *who* is going to be benefited or harmed as a result of one's action (or failure to act). Failing to note the more or less obvious evidence that one's action will result in *one's own* significant harm can, by itself, be sufficient to convict one of acting irrationally. Failing to note evidence of the same sort that one's action will result in the same harm for someone *else* is not, by itself, sufficient.

[27] Consider a case in which one's agoraphobia provides a valid moral excuse. In such a case, it does not really matter if the phobia is described as an irrational fear of the outdoors, or the irrational belief that something terribly bad will happen if one does go outdoors.

Of course such a failure still counts as failure of theoretical rationality. Depending on the circumstances it may also count as a moral failure. And if the person is, for example, one's beloved child, spouse, or friend, it may also count as a failure of practical rationality. But the claim here is about what counts, *by itself*, as a sufficient condition for acting irrationally. The loss of a beloved child, spouse, or friend normally entails substantial harms for oneself, and it is these harms that are most plausibly taken to ground the charge of irrationality in such cases. Of course this does not mean that when one acts to save one's beloved child, one is motivated by the prospect of avoiding harms to oneself. In fact, love is as much a matter of acting on the reasons that stem from the interests of one's beloved as it is a matter of suffering when one's beloved is harmed.

The relevance of *who* is going to be harmed by one's action, when one should be but is not aware of the likelihood of such harm, not only supports the preceding arguments against the Humean suggestion, but also points towards the following relativization:

> R3 An action is irrational if, relative to the beliefs that the agent should have about harms the agent may suffer, and to the beliefs that the agent *actually* has about benefits to anyone at all, it is objectively irrational.

This relativization is simply a response to the counterexamples to R1 and R2. It is therefore, and admittedly, ad hoc. But it comes very close to working. Nevertheless, it does not work. Consider the following case. Suppose someone decides to kill herself in a painful way in order to produce terrible guilt feelings in her parents. In the circumstances she can see that a side-effect will be that her death will save the lives of a handful of other people. But let us suppose these others are people whom the agent hates and wishes dead. Thus the side-effect is completely undesired. But, let us further suppose, the agent's wish for the death of these other people is not very great. Though she feels mild regret that her own death will save them, she still believes that, overall, it is worth killing herself to punish her parents. Now, if she does kill herself, her action is not irrational according to definition R3. For that definition makes no reference to the motives of the agent. It mentions only the agent's beliefs, and the beliefs the agent should have. Because it involves only beliefs, it cannot distinguish cases in which a heroic person dies *in order to* save other people, from cases in which a spiteful person dies *with the mere distasteful knowledge that* she will thereby save other people. What seems important in this latter case is that

the reason in favor of suicide, namely that suicide will save several other people from death, does not motivate the agent at all. In fact, she would kill herself even more readily in the absence of this consideration.

Note, however, that though the motivations of the agent are relevant to the question of whether saving other people's lives rationally *justifies* suicide, the motivations of the agent are not relevant to the question of whether saving the agent's own life is (in the absence of strong countervailing reasons) rationally *required*. That is, when we are considering the subjective rationality of an action, the motivations of the agent generally impact the relevance of possible benefits to *others* in a logically distinct and more significant way than such motivations impact the relevance of harm to the *agent*. For if the agent is not motivated by benefits to others, those benefits generally play little or no role in determining the rationality of the action; roughly speaking, it is as if those benefits were not possible consequences. But even if the agent is not motivated by possible harm to herself, those harms continue to provide requiring reasons against the action.[28]

Derek Parfit's notions of *what we have most reason to do* and *what is most rational for us to do* parallel my notions of objective and subjective rationality. Trying to clarify the relation between the former and the latter, Parfit notes that "[w]hile reasons are provided by the facts, the rationality of our desires and acts depends instead on what we believe, or given the evidence, ought rationally to believe."[29] It is to Parfit's credit that he notices the relevance, to subjective rationality, both of *actual* beliefs, and of beliefs that we *ought* to have. But he does not note that what we 'ought rationally to believe' can only make actions rationally *required*, and cannot ever provide a rational *justification* unless we *also* believe it. That is, he does not consider the sort of cases that provide counterexamples to R2. Even if Parfit did make this distinction, his account of the relation between objective and subjective rationality would only be as adequate as R3, for it does not mention

[28] Philippa Foot has noticed something like this, in recognizing the different logical roles played by the objective *interests* of the agent, and the objects of the subjective *desires* of the agent. See Foot (1978b), p. 156. And David Copp makes a similar distinction between the objective needs and the subjective values of agents in determining the rationality of their actions. See Copp (1995), pp. 172–85. My view *explains* these asymmetries by reference to an agent/other asymmetry in the notion of objective rationality, which has itself been independently motivated. It should also be noted that neither Foot nor Copp make any explicit distinction between objective and subjective rationality.

[29] Parfit (1997), p. 99.

the differential relevance of the agent's motivation to the justifying and requiring roles of reasons.

The above considerations lead to a final and most extravagantly ad hoc definition:

> R4 An action is irrational if, relative to the beliefs that the agent should have about harms the agent may suffer, and to the beliefs about benefits to others that the agent both (a) actually has and (b) is moved by, it is objectively irrational.[30]

One obvious problem with this definition is that beliefs motivate people to different degrees. This problem, though it provides a strong objection to the definition, could be overcome with a definition still more complicated. But that does not matter. The point to note here is that definition R4 is so ad hoc that, even if it were correct, it would fail to supply almost any insight into the nature of irrational action. A modified definition would fail even more spectacularly. The definition does nothing to suggest a unifying principle behind each of the counterexamples. The lesson we should learn from the failure of definitions R1 through R4 is that the relation between objective and subjective rationality is not fruitfully conceived as these definitions have conceived it. That is, subjective rationality is not fruitfully conceived as objective rationality relative to some special class of beliefs.

One of the reasons for the failure of the above definitions should be obvious from the nature of the counterexamples to R1 and R2. The counterexample to R1 took advantage of the fact that reasons with considerable requiring strength can be relevant to the rationality of an agent's action even though the agent does not see those reasons: it is sufficient that the agent *should* see them. The counterexample to R2 took advantage of the fact that reasons that primarily play a justifying role are not relevant to the rationality of an agent's action unless they are *actually* seen by the agent. R3 implicitly took account of the justifying/requiring distinction by treating primarily justifying reasons (i.e., those involving benefits to others) differently from reasons that also require (i.e., those involving harm to the agent). This is why it represented such an increase in adequacy. But it also failed, for the justifying power of a reason is not relevant to the subjective rationality of an action unless it provides the agent with a motive. Unless

[30] This definition is not meant to involve any controversial claims about beliefs being motivating on their own. As far as R4 is concerned, a belief can move a person in virtue of an antecedent desire, or by some other means.

one has the justifying/requiring distinction in place in one's account of objective rationality, it will be almost impossible to understand these failures to define subjective rationality in terms of objective rationality. For there does not appear to be any other explanation for the fact that some considerations continue to be relevant to the rationality of action regardless of whether they are actually believed by the agent, while other considerations do not appear to be relevant unless they are actually believed. The difference here is emphatically not the difference between culpable and nonculpable ignorance.[31] For even if we eliminate all nonculpable ignorance by *fiat*, stipulating that the agent should be aware of all the relevant information, none of the counterexamples to R1 through R4 are altered in any way. For none of those counterexamples made any reference to nonculpable ignorance.

Given the above remarks, the reader may now be expecting a definition of irrationality that explicitly mentions the requiring and justifying roles of reasons.[32] But what we would really like is an account that *explains* why belief and desire have a differential impact on the requiring and justifying roles of reasons, when the move is made from objective to subjective rationality. This would be superior to an account that merely took account of this differential impact. Here then is the official definition of subjectively irrational action:

> R5 An action is subjectively irrational iff it proceeds from a state of the agent that (a) normally puts an agent at increased risk of performing objectively irrational actions, and (b) has its adverse effect by influencing the formation of intentions in the light of sensory evidence and beliefs.[33]

R5 explains the following facts. It explains why it is relevant to the rationality of an action that an agent *should* have a belief that his action will cause him harm, and not exclusively relevant that he actually *does* have such a belief. For if an agent fails to have such a belief when, in the light of sensory evidence and other beliefs, he should have had it, then he has

[31] Although this distinction is, of course, relevant to the rationality of action in other ways.

[32] Indeed, Bernard Gert has recently been pushed by the preceding criticisms into just such a view.

[33] 'Normally,' because science-fiction worlds in which a guardian angel secretly guarantees that all my actions are objectively rational are nevertheless worlds in which I can act irrationally. A useful gloss of 'normally' is 'in the circumstances responsible for the development of the concept of subjectively irrational action.' In this connection, see Millikan (2000), pp. 61–68.

failed to see something that he should have seen, and, given that what he failed to see is a *requiring* reason, his failure shows that he is at increased risk of doing things that are objectively irrational. This explains the counterexample to R1. On the other hand, R5 also explains why it is relevant to the rationality of an action that an agent *actually* have a belief about the compensating benefits of his action and other primarily justifying reasons, rather than merely that he *should* have such a belief. For if the agent merely *should* have such a belief, but does not have it, and acts in a way that he knows will bring him harm, this shows that he is insufficiently averse to the perceived harm – a requiring reason – and thus that he is at increased risk of doing objectively irrational things. This explains the counterexample to R2. A similar explanation accounts for the fact that beliefs about the benefits to others must supply a *motive* for the agent if they are to rationally justify her action. That is, suppose one knows two things about one's action: that it will cause one a lot of pain, and that it will prevent someone else from suffering a comparable amount of pain. If one does not care that the action will prevent someone else from suffering pain, and goes ahead and does the action, then one's action involves a failure to be appropriately averse to one's own pain. This explains the counterexample to R3.

Clause (b) links subjective irrationality explicitly to the will. It thus eliminates factors such as blindness and clumsiness as relevant sources of increased risk of acting in an objectively irrational way. If someone is blind or clumsy, and consequently at increased risk of harming himself, we do not call his actions 'irrational,' but 'unfortunate.'[34] Clause (b) also provides some explanation for the fact that 'irrational' is a term of practical criticism. For while criticism of actions that stem from blindness or clumsiness would typically have no point, actions that meet clause (b) are often ones that could be avoided by an effort of will, and are consequently ones for which criticism could have a point. Indeed, part of the point of having the 'mental functioning' concept of irrational action (less metaphorically: part of what explains the presence in the language of words that allow one to have this concept) is that it helps us to provide the very criticism to which such actions are liable. In some cases, of course, the state of the agent from which the action proceeds is a disability of the will – say, an addiction – so that criticism may be doomed to failure. Thus, while 'rationally ought'

[34] This remark indicates the sense of 'increased' in 'increased risk.' It means 'greater risk than the population at large.'

generally implies 'can,' it does not always do so. Addictions, phobias, and compulsions certainly cause people to do things that, rationally speaking, they ought not do. And yet it is unclear whether, in any important sense of 'can' that goes beyond mere physical ability, they can always avoid doing those things. Clause (b) also suggests that some actions stemming from extremely low intelligence should be classified as irrational. Although this may go against common usage to some degree, it is worth noting that extremely low intelligence is linked to diminished moral responsibility, diminished competence to give consent, and even diminished freedom of the will, in the same way that more 'intelligent' irrationality is. Moreover, the sorts of everyday irrational actions with which this book is concerned are often called 'stupid' or 'dumb' even when they do not proceed from low intelligence at all. That these words are nevertheless used suggests that actions that stem from actual stupidity share important features with more paradigmatically irrational actions.[35]

On this account of practical rationality, objective rationality may be seen as playing a role similar to that played by truth in accounts of theoretical rationality.[36] That is, one might define 'irrational belief' in the following way:

> B1 A belief is irrational iff it proceeds from a state of the agent of a kind that (a) normally puts an agent at increased risk of believing false things, and (b) has its adverse effect by influencing the formation of beliefs in the light of sensory evidence and other beliefs.

Although it is beyond the scope of this book argue for it, B1 does seem to capture much of the content of 'irrational belief,' where irrational beliefs are ones that we take to reflect negatively on the cognitive capacities of the believer. The plausibility of B1 as an account of irrational belief would reinforce this book's account of the relation between objective and subjective rationality. But of course there are so many differences between beliefs and actions that we should not expect B1 to be perfectly adequate, even if R5 is correct.

R5, taken together with A2 and A3, accounts for the plausibility of a great diversity of alternate accounts of rationality, and for our intuitions in some puzzling cases. For example, R5 accounts for the plausibility of

[35] I thank Maria Victoria Costa for these points.
[36] In this connection, see Anscombe (1995), pp. 32–33.

almost all purely formal accounts of rationality, such as 'considered preference' accounts. For if one is acting against one's considered preferences, this indicates a sort of malfunctioning in the will that places one at increased risk of doing objectively irrational actions. But R5 also accounts for the possibility of irrationally acting *in accord with* one's considered preferences, in those cases in which those preferences are themselves objectively irrational. Purely formal accounts of practical rationality notoriously fail to do this, so that it becomes perfectly in accord with rationality to prefer the destruction of the entire world to the scratching of one's little finger, or to opt, for no reason, against taking the medicine that will return one to perfect health and a pleasant life.[37] R5 also accounts for the normativity of instrumental reason, since failures in instrumental rationality are certainly characteristic of a state of increased risk of doing objectively irrational actions. This is true even if the particular failure of instrumental rationality turns out to be 'for the best.' For suppose that it was blind luck that the failure was for the best. Then similar failures in the future are likely to be harmful. On the other hand, suppose that the failure in instrumental rationality was for the best because the end to which one failed to take appropriate means was a harmful end. In this case, even if one is 'cured' of having the harmful end, one still has a *problem*, and it is from this *problem* that the original failure proceeded. Similarly, R5 explains why it counts as irrational to act against one's own normative judgments – at least if the capacity for coming to such judgments can plausibly be seen as a tool for increasing one's chances of acting in an objectively rational way. Thus Thomas Scanlon's restricted understanding of 'irrational' can be understood as corresponding to one specific form of a much more general phenomenon, and its negative normative significance can be explained. R5 also accounts for the normative significance of formal restrictions on preferences, such as transitivity, without making it irrational to change one's preferences, and without falling prey to an objection that affects 'dutch book' arguments for the same formal features: the objection that one of one's strong preferences may be to fall victim to the dutch book. In all these cases – failure to act on one's considered preferences, failure to have objectively rational ends, failure to take the proper means to one's ends, failure to act on one's judgment regarding how one ought to act, failure to be sufficiently consistent in one's preference-ranking – there is a different *type of state* that explains why one is at increased risk of doing objectively

[37] Hume (1978), p. 415; Williams (1981), p. 105.

irrational actions. This shows how R5 might be used to create a sort of *taxonomy* of practical irrationality. In connection with the idea of such a taxonomy, it is worth remarking that there is a sort of prejudice against substantive accounts of rationality that offer a list of reason-providing considerations, such as pain, premature death, knowledge, ability, and so on.[38] Such lists are accused of being arbitrary. But it is amazing how many *formal* accounts of rationality offer similar lists without incurring the same sort of criticism.[39] True, these lists are lists of formal conditions, which are more appealing to philosophers. But all that the members of these lists of formal conditions typically have to recommend themselves is a very high degree of surface plausibility – just a bit less surface plausibility, in my estimation, than the claim that premature death is bad, and pleasure good. R5 offers a way of justifying these lists of formal conditions.

R5 also accounts for the subjective irrationality of behaviors that result from mental disorders such as schizophrenia, which other accounts will have a surprisingly difficult time explaining. For example, suppose that Jim hears voices in his head telling him that wearing tin foil under his clothes is the only way to protect himself from some dangerous rays. Let us grant, for the sake of argument, that wearing tin foil in this way is sufficiently uncomfortable that one requires a justification for doing so. Now, if Jim knows nothing of psychology, it is not completely unreasonable – in a certain sense – for him to take the voices in his head that identify themselves as powerful aliens to *be* powerful aliens. And it would not be unreasonable therefore for Jim to believe that the tin foil they recommend really will help him avoid the effects of the dangerous rays. But despite all this 'reasonableness,' someone like Jim, who hears voices telling him to wear tin foil under his clothes and who therefore wears it, is acting irrationally. The Noah of the Bible, on the other hand, was not acting irrationally in suffering the ridicule of his neighbors by building the ark. Epistemically, Jim and Noah are in pretty much the same situation. How then, can we call the one's actions 'irrational' and the other's 'rational'? By the appeal that R5 makes to the risk-increasing nature of the state that causes the relevant behavior.

Finally, R5 allows a satisfactory treatment of 'Schelling cases.' For example, consider the case discussed in Parfit.[40] Some thieves are threatening

[38] It is probably worth making it clear that the account of objective rationality offered in this chapter does not do so: it only results in the *extension being specifiable* in terms of a list.
[39] See especially Nozick (1993) and Brandt (1979).
[40] Parfit (1986), pp. 12–13.

one's family with death unless one opens a safe containing valuables within the next fifteen minutes. Under the circumstances, if one takes an 'irrationality pill,' this will prevent the thieves from being able to press one to open the safe, since one will not respond rationally to threats to oneself or the things one cares about. *Taking* the pill is both objectively and subjectively rational, according to R5 and A3. What about the subsequent 'crazy' actions? Since, by stipulation, the pill puts one in a state in which one does not try to avoid harms to oneself or to the things one cares about, the actions that are the result of having taken the pill are irrational.[41] And yet the actions are also, in a sense, objectively rational: like the original action of *taking* the pill, these actions do involve, albeit somewhat indirectly, an adequate compensating benefit. Given that the actions are objectively rational in this sense, it might seem reasonable to suggest that they must also be rational. After all, they are the result of a state that is, *in the circumstances*, producing objectively rational actions. But the relevant state of the agent here is something like the following: not being disposed to avoid harms to oneself, or to the things one cares deeply about. It is just false that the state the pill induces is the following: not being disposed to avoid harms to oneself, or to the things one cares deeply about, *when this disposition is useful*.[42] After all, if the thieves suddenly run away, one will continue to act in crazy ways. Thus the actions that result from having taken the pill count as objectively rational in a sense, but as paradigmatically subjectively irrational.

CONCLUSION

The purely justificatory role of altruistic reasons is of great importance to an understanding of rationality in the 'proper mental functioning' sense. Because of their exclusively justificatory role, altruistic reasons need not motivate even rational agents to any degree. That is, even fully informed agents can in general be completely unmotivated by such reasons, and yet always act in an objectively rational way.[43] For if they always avoid harms

[41] The pill does not simply make one no longer care about these things.

[42] For related points, see Parfit (2001), pp. 85–86.

[43] A slightly moderated claim would be that altruistic reasons are *primarily* justifying, so that fully informed rational agents could be *almost* completely unmotivated by them. This moderation of the view does not have any substantial effect on the position for which this book is arguing, or on its criticisms of other views of rationality. For it still requires a sharp logical distinction between the justifying and requiring strengths of reasons. See pp. 89–92.

to themselves, they are at no increased risk of acting in an objectively irrational way. Now consider the case in which an agent performs an action that he knows will harm him considerably, but where he also knows that the action will benefit someone else to a comparable extent. In such a case there are two possible explanations for the action. One is that the agent was motivated by the altruistic reason. The other is that the agent was insufficiently averse to his own harm. In the former case, the action does not involve any failure of mental functioning. That is, the agent who acts from such motives is not at increased risk of performing objectively irrational actions. But in the latter case, in which the agent is insufficiently averse to his own harm, the action does involve such a failure. That is, the agent *is* at increased risk of performing objectively irrational actions. It is primarily in distinguishing these two cases that the motivations of the agent become relevant to the rational status of an action. And the relevant question is: 'Is the agent motivated by the prospect of benefiting *someone else*?'

Thus there is an important role for the contingent desires of the agent in determining which reasons are relevant to the (subjective) rational status of her actions, and how important those reasons are. But this role is limited to reasons that play an exclusively or primarily justificatory role: reasons involving the interests of others, and reasons that involve benefits to the agent, as opposed to harms. This accounts for much of what is plausible in Humean accounts of rationality. But the account offered in this chapter also allows for the Kantian intuition that there are some categorical requirements of reason – that there are some actions that are irrational regardless of the contingent desires of the agent. Of course, these requirements are not, as the Kantian typically argues, moral requirements.[44] However, much of the appeal of Kantian views stems from dissatisfaction with accounts that invest brute desire with too much normative significance, rather than from a prior sense that morality is rationally required. Kantians deny that our desires provide rational justification. Instead, they hold that justification is provided by the reasons that stand behind our desires. The view offered in this book is in wholehearted agreement with this intuition.

[44] Philippa Foot (1978b), pp. 148–56 offers a view that also responds to this combination of Humean and Kantian intuitions, suggesting that reasons stem from the interests *and* desires of the agent, but that only the former provide categorical rational requirements. David Copp (1995) offers a similar 'needs and values' theory of rationality. The account offered in this chapter explains the plausibility of these other views.

8

Internalism and different kinds of reasons

The purpose of the present chapter is to bring the requiring/justifying distinction to bear on a central controversy in contemporary ethical theory: the internalism/externalism debate.[1] This debate concerns the nature of the relation between practical reasons and the desires of the agents who have those reasons. Crudely put, internalists hold that there is a very strong relation between the desires of a (rational) agent, and the reasons that such an agent has, while externalists hold that the reasons that an agent has are given by features of her situation in the world, and are independent of her attitudes towards those features. The nature of the relation between practical reasons and desire is of obvious relevance to a large number of central philosophical and practical questions, including the rational status of morally required behavior, and the reasonableness of punishing people who act in significantly immoral ways.

Parties to the internalism/externalism debate have typically assumed that, with regard to practical reasons, either internalism or externalism is correct. And they have assumed that if, for example, internalism should turn out to be the correct account, then there will be a single correct interpretation of internalism that holds for all practical reasons. In bringing the requiring/justifying distinction to bear on the internalism/externalism debate, one major point of this chapter is that these assumptions are almost certainly false, and that a failure to see this has hamstrung the discussion

[1] The terms 'internalism' and 'externalism' are of course used to describe a wide range of views. As understood in this chapter, internalism and externalism are views about whether or not it is a necessary condition on the existence of a practical reason, that the agent have a related motivation. Other versions of internalism concern a putative necessary condition on judgments or beliefs about reasons, rather than on the existence of reasons. For the distinction between judgment and existence internalism, see Darwall (1983), pp. 54–56. There are also debates between internalists and externalists about moral reasons, moral obligations, and moral judgments, although such debates often presuppose some view about practical reasons internalism more generally.

from the beginning.[2] Practical reasons have, as we have seen, two logically distinct normative roles in determining the rational status of action: requiring and justifying. The relation of desire to the requiring role of a practical reason will almost certainly be very different from the relation of desire to the justifying role of the very same reason.

INTERNALISM AND EXTERNALISM

All internalists could endorse the following general claim about the link between practical reasons and the motivations of the agents who have those reasons: any practical reason must find a corresponding motivation in the agent. But there are two significantly different ways of reading the 'must' in 'must find a corresponding motivation.' Some internalists take it as making the existence of reasons depend on the existence of some corresponding contingent antecedent motivation in the agent. These internalists hold that if an agent has a reason, then it follows as a matter of conceptual necessity that the agent actually has a relevant antecedent motivation. For example, an agent would be held to have a reason to take a walk, or to help a stranger, only if that agent had some relevant desire, or adhered to some relevant principle, or was (in the case of helping the stranger) benevolently disposed, or if some other relevant motivational claim were antecedently true of the agent. Of course it need not be the case that the agent actually has a motivation to do the very action for which he has a reason: internalists of the sort under discussion here do not deny the possibility of irrational action, or of ignorance or mistake about the reasons one has. But they do hold that if an agent has a reason to perform some action, then the agent does indeed actually have some kind of motivation from which rational processes could produce a motivation to perform the action. Moreover, according to this form of internalism, the relevant rational processes can only produce such motivation from antecedent motivation, so that the actual reasons an agent has are always contingent on his antecedent motivational setup. I will call philosophers who hold this form of internalism 'Humeans,'[3] and I will call their interpretation of internalism 'the Humean interpretation.'

[2] The most recent instantiation of the debate might be dated as beginning in 1981, with the republication of Williams (1981).

[3] Bernard Williams is such a philosopher. See Williams (1981).

Other internalists take the 'must' in 'must find a corresponding motivation' as a normative requirement on agents, rather than as a necessary condition for the existence of reasons. These internalists hold that practical reasons are given by the situation in which the agent finds herself, and these reasons must find a corresponding motivation if the agent is to avoid the charge of irrationality. What is distinctive about the internalists of this second type is that they hold that reasons for action are conceptually independent of any contingencies in the motivational setup of the agent who has those reasons.[4] I will call philosophers who hold an internalism of this sort 'Kantians,' and I will call their interpretation of internalism 'the Kantian interpretation.'[5] There are two ways of being a Kantian internalist. The first way is to hold that there are certain rational processes that would produce specific motivations in any agent, completely independently of that agent's contingent antecedent motivation. This way of being a Kantian internalist involves denying the Humean internalist's claim that rational processes cannot produce motivation except from antecedent motivation.[6] A second way of being a Kantian internalist is to hold that there are simply some motivations that one is rationally required to have, even though there may be no rational processes that would necessarily bring one to have them.[7] This sort of Kantian might hold that the desires to avoid death and pain, for example, are simply rationally required, in virtue of death and pain being bad things. For such a Kantian, failure to have one of these desires would by itself be enough to convict one of irrationality. Sometimes this sort of Kantian internalist is called an externalist, because the status of a consideration as a reason seems so far removed from the nature of the agent whose reason it is.[8] But this chapter will reserve

[4] Of course our desires are likely to have a causal impact on substantive consequences such as our likelihood of feeling frustration or satisfaction, or our likelihood of success in the relevant action. And these consequences may constitute reasons.

[5] The labels 'Humean' and 'Kantian' are not intended to smuggle in covert interpretations of Hume and Kant. In fact, neither Hume nor Kant talk explicitly about normative reasons in the typical senses in which such reasons are understood in current debates about internalism. But there are modern Humeans and modern Kantians for whom the labels are appropriate.

[6] Christine Korsgaard is such a philosopher. See Korsgaard (1996b). She does not actually deny the Humean's claim in this article, but only makes room for the denial, which she makes explicit in, for example, Korsgaard (1996a), ch. 4.

[7] This type of Kantian internalist would be classified as a "non-constitutive existence internalist" by Darwall. The former Kantian internalist, and the Humean, would both be classified as "constitutive existence internalists." See Darwall (1992), pp. 158–59, 165.

[8] See, for example, Parfit (1997), p. 101 and Brink (1986), p. 36.

the term 'externalist' for an easily distinguishable and much more extreme view.

Externalists about practical reasons, as here understood, deny the claims of both Humean and Kantian internalists. They deny the Humean claim by holding that our reasons, in a given situation, do not depend in any conceptual way upon our contingent desires. Thus, a characteristic externalist claim would be that the prospect of restored health always provides an agent with a reason, whether or not that agent is interested in being restored to health. Of course such an externalist would also grant that if restored health will bring a host of significant problems, then there might be countervailing reasons that make it rationally permissible not to want, overall, to be restored to health. But the reasons at issue in the internalism/externalism debate are *pro tanto* ones: the sort that can be opposed or augmented by other *pro tanto* reasons.

The externalist also denies the Kantian internalist's claim, and holds, against the Kantian, that some reasons need not motivate us even if we are fully rational. Thus, the externalist might claim that while it would not be irrational in any way to be completely unmotivated to do anything to entertain a small child left in one's charge for the afternoon, the prospect of providing pleasure to that child still provides a reason that would justify taking some trouble to do so. Such a position is not very popular, since it is almost impossible to see what such a reason claim would amount to without first recognizing the requiring/justifying distinction.[9] This near impossibility is perhaps the result of the philosopher's tendency to simplify matters in order to see more clearly. In this case, such a tendency leads to the consideration of cases in which only one reason is relevant. But when one considers a reason in isolation in this way, only the requiring role is likely to be apparent. In order for justifying strength to be relevant, there must be something to justify: that is, there must already be opposing reasons with some requiring strength. If one simplifies by removing countervailing reasons, and thus only sees the requiring role, then the externalist position is going to appear conceptually confused. This explanation also helps explain why Joseph Raz, who holds a position very similar to the externalist in most cases, nevertheless seems to hold that if there is only one reason relevant to a choice, then the agent is rationally required to be moved by it.[10]

[9] See, e.g., Kagan (1989), pp. 378–80.

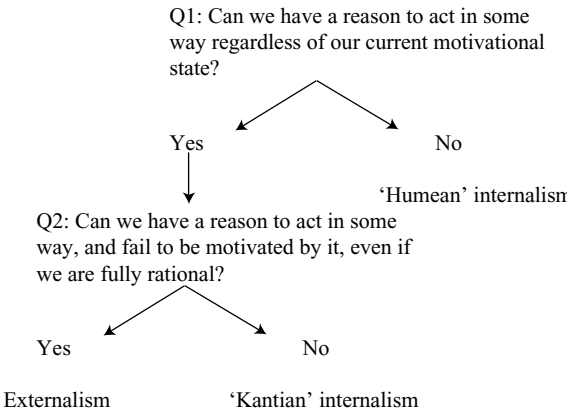
[10] See Raz (1999b), pp. 90–117.

Of course not all parties to the internalism/externalism debate are committed to the exclusive truth of only one of the views just outlined. Some, for example, are Kantian internalists about reasons involving the needs or objective interests of the agent, and yet hold that there are other reasons that depend upon the agent's contingent motivations.[11] Such views are more likely to approach the truth, and seem partly to have been shaped by an implicit recognition of the justifying/requiring distinction. For simplicity's sake, I do not discuss these hybrid accounts. But it should be clear that despite a more delicate feeling for the normative phenomena, they are unlikely to be adequate unless they make their recognition explicit.

Many recent discussions of internalism and externalism are extremely subtle, and it may appear that the present discussion is overlooking important distinctions. But the distinctions between Humean internalism, Kantian internalism, and externalism, as they have been drawn here, are not intended to be subtle. The points that will be made here do not depend in any way on the sorts of fine distinctions that are made elsewhere. Indeed, the reader will already have noticed that there is an important dispute within the ranks of those whom I have described as 'Kantian internalists,' between those who hold that reasons stem from the existence of objective goods or evils, and those who hold that reasons stem from some imperative-producing feature of human rational processes.[12] But in arguing that Kantian internalism is the correct view of reasons in their requiring role, there is no need to resolve this dispute. Significant work will have been done if it is shown that neither Humean internalism, nor externalism, is a viable view of requiring reasons. Here, then, is a summary diagram explaining the relation between Humean internalism, Kantian internalism, and externalism, as those notions are used in the remainder of this chapter:

[11] For instance, David Copp's 'needs and values' view of rationality places a similar division in the center of his account of rationality. See Copp (1995), pp. 172–85. See also Foot (1978b), pp. 148–56.

[12] Garrett Cullity and Berys Gaut call the following two forms of Kantian internalism "the recognitional view" and "legislative universalism," respectively, attributing the former to Aristotle and the latter to Kant. These two sorts of Kantian internalists share a commitment to what Cullity and Gaut call "categorical reasons." See Cullity and Gaut (1997), pp. 3–5.



In what follows, I will first present cases that seem to favor each version of internalism over the other. Then I will argue that the plausibility of each case depends very significantly upon whether we are considering the justifying role of reasons, or the requiring role.

An example favoring the Humean interpretation over the Kantian

On January 13, 1982, Air Florida's Flight 90 out of Washington's National Airport crashed into a bridge over the Potomac and landed in the freezing water. An anonymous man, to whom a life-line was repeatedly given, repeatedly passed it to others in the freezing water, until he himself finally succumbed to the cold and drowned. What are we to say about the rational status of this man's actions, and about the contributions made to their rational status by the reasons that favor and oppose them? Certainly there was a powerful reason *against* passing the life-line to another person: that the man risked death in doing so. Were it not for the possibility of saving other people, it would have been irrational to have acted against this reason. That is, it would have been irrational to have passed the life-preserver to someone else if no one else really needed it. But in the actual case, there were also very powerful reasons in *favor* of the heroic actions: that several lives could be saved. These reasons made it rationally permissible to act against the other reason.

Let us now alter our description of the case. Suppose the heroic man was in fact not motivated in his heroic actions by the thought of saving the

freezing passengers. Rather, a woman with whom he had had one date, and who worked at the local television station, had decided not to go out with him again because, by his own admission, he never did anything heroic. The crash gave him the opportunity to display some heroism, and he was confident that this woman, seeing his behavior in the editing room, would reconsider her decision and give him another chance. Let us suppose that he was reasonable in believing this about the woman. Let us also stipulate that if he had not believed that she would see the footage of his actions, and if he had not believed she would then go out on a second date with him, he wouldn't even have considered trying to save the passengers. This is what is meant by saying that he was not motivated by the thought of saving the passengers, but only by the prospect of a second date. And, let us say, it also made no difference to him whether he thought the passengers would really be saved by his actions – he only cared that it looked as if he cared. Let us even suppose that he almost 'saved' a passenger who was already dead. What are we now to say about the rationality of this man's actions? The reasons that justified his risking his own life in the first, real-life, case do not seem, in this fictional case, to do so. That is, the fact that the man knows that he is saving the lives of the passengers does nothing to mitigate the irrationality of his actions. Despite the fact that he knows that he is saving the lives of these people, his actions really demonstrate only a pathological obsession with a woman whom he hardly knows.

This type of case suggests that the antecedent motivations of the agent play a role in determining whether a given consideration is a normative reason. For in a case in which an agent is strongly motivated to save the lives of others at high personal risk, the fact that the action will probably save those lives seems obviously to count as a strong reason. But in the case where the agent merely knows that his actions will save those lives, but does not care about that fact, then the fact that the action will probably save those lives does not seem to count as a reason at all. This apparent dependence upon antecedent motivation argues in favor of the Humean interpretation of internalism. Of course it does not provide a deductive proof that the Kantian internalist is wrong, even in this case. A committed Kantian internalist could still maintain that the heroically strong motivation to save the freezing passengers would have been produced in any fully rational agent. This would probably imply that to save one's own life when one could have saved five others is in fact irrational. Such a view, and the corresponding implication, seem false, but I will not argue here that they are. At the end of this chapter the Kantian should

simply have less philosophical motivation to make such claims. For such theorists should recognize, by that point, that they have been presenting their arguments without realizing that reasons can play two distinct normative roles in determining the rational status of actions. In general their arguments involve only the requiring role of reasons. And so they should not assume that their arguments apply equally well when the primary role of a reason is to justify.

An example favoring the Kantian interpretation over the Humean

Paradoxically, the following example comes from Bernard Williams, who is himself a Humean internalist. The reader may or may not feel that Williams has in fact provided a *reductio* of his own view, but the example should at least show that any Humean as honest as Williams is will be compelled to accept some counterintuitive consequences.

To set the stage for the example, Williams first concedes that "insofar as there are determinately recognizable needs, there can be an agent who lacks any interest in getting what he in fact needs."[13] The example itself is that of a sick person who has no desire to take the medicine that will restore his health. What should we say about the rationality of this sick person's refusal of medicine? Recall that the case is not one in which the agent wants to die, perhaps as a release from suffering. Rather, this sick person simply lacks a desire to be healthy. Williams's point would remain the same even if we stipulated that this agent knew that, if restored to health, he would live a life of uninterrupted virtue and felicity. Most nonphilosophers, including trained psychologists, would almost certainly conclude that such a person was suffering from a mental illness (perhaps depression), and that part of the very illness lay in a failure to be moved by reasons that would move a more rational person.[14] But Williams, and other adherents to Humean internalism, cannot say this. Rather, Williams is very explicit in endorsing the view that if such an agent:

> really is uninterested in pursuing what he needs; and this is not the product of false belief; and he could not reach any such motive from motives he has by the kind of deliberative processes we have discussed; then I think we do have to say that in the internal sense he indeed has no reason to pursue these things.[15]

[13] Williams (1981), p. 105.

[14] See Deigh (1996), pp. 133–59.

[15] Williams (1981), p. 105. It is worth noting the phrase "I think we do have to say" here. This is not the kind of phrase that precedes a conclusion one regards as independently plausible.

In this case, the failure to have a desire for health seems to count against the rationality of the agent. In general, if an agent is completely unconcerned with things like the prospect of disease, death, injury, loss of freedom, and so on, this does not lead us to conclude that the agent has no reason to avoid these things. Rather, such failures in motivation are all the data we need in order to conclude that the agent is not completely rational. Thus the example of the irrational sick person is consistent with the Kantian interpretation of internalism, but argues against the Humean interpretation of internalism. Of course it does not provide a deductive proof that the Humean is wrong. When we confront a stubborn Humean with a sick person who refuses medicine, that Humean can always dogmatically assert that the sick person really does have some desire – however weak – for health, or that the sick person really has no reason to take the medicine. But for the Humean who takes the first of these routes, the question of the strength of the reason will remain a problem, since it will be hard to get a strong reason out of a weak motivation. And the second route is not very attractive. In any case, at the end of this chapter the Humean should simply have less philosophical motivation to make such dogmatic claims. For such theorists should recognize, by that point, that they have been presenting their arguments without realizing that reasons can play two different normative roles in determining the rational status of actions. In general their arguments involve only the justifying role of reasons. And so they should not assume that their arguments apply equally well when the primary role of a reason is to require.

THE PROPOSAL

The overarching critical point of this chapter is that the ongoing debate about practical reasons internalism has been hopelessly confused on account of a failure to recognize that justifying and requiring are two separate roles of normative reasons. The remainder of the chapter offers arguments in favor of the following two more positive claims:

(1) Kantian internalism is true of requiring reasons.
(2) Externalism is true of justifying reasons.

The irrational sick person revisited

In the case of the sick person who refuses medicine the reason at issue – that the medicine will restore his health – is relevant as a requirement on action,

and not as a justification for acting against any other reasons. This can be seen by noting that in discussion of this example, what is at issue is whether or not it would be irrational to ignore this reason, and to continue to refuse the medicine, unless one had a powerful opposing reason that could justify doing so. The Kantian internalist's plausible position is that the agent would be irrational, and that this is true regardless of the antecedent motives of the agent. If the agent did not have the required corresponding antecedent motivation, this would only show that the agent was to some degree irrational. It would not show that the reason was not a reason. Here is the argument for the Kantian view. Suppose that we have some reason, and that it rationally requires an action. If this is so, then what is being required by the reason is not only action, but also motivation itself. The point can be put in the form of a rhetorical question: How could we be rationally required to do anything in particular – to avoid illness or pain, for example – unless we were also rationally required to have the motives that would lead us to do that particular thing? Suppose, for example, that one of our rational requirements is the requirement to preserve our ability to reason. And suppose that on a certain occasion it is necessary to take some medication to comply with this requirement. If an agent takes this medication with the false belief that it will destroy his ability to reason, is he nevertheless immune to rational criticism? The above rhetorical question is, I think, what stands behind the Kantian internalist's intuition about the possibility of rational assessment not only of actions, but of motives. Where requirements on action are at issue, there are implicit requirements on motives. And this means that the motives cannot be antecedently necessary conditions on the requirements, and, a fortiori, that requiring reasons do not depend upon antecedent motivations. That is, the Humean interpretation of internalism does not apply to requiring reasons. As a consequence, the Kantian interpretation does apply, although this is not merely the result of a simple argument by elimination. Rather, it is because the reasons at issue are rational requirements that the Kantian view is correct. This is because to say that these reasons have some requiring strength is simply to say that failing to be motivated by these reasons is sufficient to convict the agent of some degree of irrationality. And that is just what the Kantian asserts.

Derek Parfit offers the same sort of argument in favor of Kantian internalism, and in favor of the idea that there are rational requirements on motives. He argues as follows:

We have reason to try to achieve some aim when, and because, it is relevantly worth achieving. Since these are reasons for *being* motivated, we would have these reasons even if, when we were aware of them, that awareness did not motivate us. But, if we are rational, it will.[16]

Parfit's argument here is good, insofar as we are considering reasons in their requiring role. But Parfit's short final sentence here exposes his assumption that all reasons are prima facie rational requirements.

It might seem that a parallel argument would also support the claim that Humean internalism must be false of justifying reasons, and therefore that Kantian internalism must be a correct account of reasons in both roles. One might try to argue in the following way: suppose that one has a reason that would justify acting in a certain way, but not require it. Then this reason would also justify being motivated to act that way, although it would not require it. This shows that we would have the reason for being motivated in that way, even if we were not so motivated. So far so good. But although this conclusion does argue against Humean internalism for justifying reasons, it does not support Kantian internalism for such reasons. For the conclusion does not establish that a fully rational agent would be motivated by the justifying reasons. It only establishes that those reasons would exist independently of the motivations of the agent. If Kantian internalism were true of such a reason, the agent *would* be rationally required to be motivated by it. But by stipulation, the reason is not a requiring one. So it seems in fact that the strong form of externalism presented above will be true of justifying reasons. But before we endorse this conclusion we should re-examine the case that seemed to provide such strong support for Humean internalism: the Air Florida case.

The Air Florida case revisited

In the Air Florida case the reason at issue – that the man's action saved the lives of other people – is relevant as a justification of action, not as something that requires action. This can be seen by noting that in discussion of the case, the question is emphatically not whether the heroic man's actions were rationally required. Rather, the discussion centers on the question of whether the action of the fictional passenger *is made rationally permissible* – i.e., is rationally justified – by the reason. The action, involving such grave

[16] Parfit (1997), p. 130.

personal risk, is obviously one that stands in need of justification. That is, without a powerful justification, it would have been irrational to have done what the heroic man did, refusing the life-line time after time, until he finally succumbed to the cold. But it was not irrational; the altruistic reason provided the necessary justification. In the discussion of this case, the contingent motivations of the agent seemed to make a great deal of difference to the question of whether or not the agent's action was indeed rationally justified (the question of moral justification is neither here nor there). When he was motivated by the fact that he could save lives, then his action was rationally permissible, though not required. But when he was motivated to act in such a risky way only by the prospect of a second date, then his action did not seem rationally permissible. This seemed to support the Humean internalist. But now we can see that, at best, it supports Humean internalism as it applies to reasons in their justifying role.

In fact, I write 'at best' above because there are differing intuitions about what one should say about the second, fictional case: the case in which the man saved the lives of the other passengers only because of the prospect of a second date. On the one hand, it is very tempting to say that this man's actions were irrational, since it is simply not worth it to risk one's life to get a second date (especially with someone who refuses a second date because one is not heroic enough). On the other hand, it is also tempting to say that *whenever* one can save the lives of a number of other people, it is rationally permissible to risk one's life to do so. The difference in intuitions here may seem to stem from a difference in the kind of thing one is evaluating: action tokens, or action types.[17] If one takes this route in explaining the different intuitions, then one can say the following: the first intuition ('irrational') is strong when one is evaluating the token action of the pathological date-seeker, while the second intuition ('rational') is strong when one is evaluating the action type of this fictional man – an

[17] This is the line I took in the paper from which this chapter was derived, although I now regard it as an inferior solution. Robert Audi also addresses this issue. See Audi (1985). Joseph Heath makes a similar distinction, in terms of *ex ante* and *ex post* questions about actions. See Heath (1997). Heath points out that Humeans like Williams conflate these two forms of justification. But Heath, like Audi, does not distinguish requiring from justifying reasons. This explains, in part, why Heath (p. 471) ends up accepting the Kantian intuition that "acting morally is, in one sense, just acting rationally." Here the qualifier "in one sense" does not indicate, benignly, the reasonable claim that moral action is rationally permissible. Rather, it expresses the much stronger view that ideal rational justification tells us to abide by moral requirements.

action type of which the action of the actual hero of Air Florida's Flight 90 was also a token. The question then becomes: are justifying reasons relevant to action types, or to action tokens, or perhaps to both?

It seems to me that it is misguided to attempt to assess the rational status of action tokens by appeal to reasons, unless this is simply understood as assessing the rational status of the relevant action type that the action token instantiates. Here is one reason for such a doubt. If a consideration can fail to be a justifying reason for an action token because that token was not based on the consideration (as one is tempted to say in the fictional Air Florida case), then it seems we could have action tokens that were required but not justified. For example, suppose that an agent is rationally required to take a significantly painful two-day treatment that will cure him of a life-threatening disease and restore him to perfect health and the prospect of a long and happy life. We can stipulate that the painfulness of the treatment, though significant, is not so great that it would be rational to refuse the treatment and die, just in order to avoid the pain. Suppose now that the agent does undergo the treatment, but only because he has been promised a bowl of cherry ice cream; he had refused the treatment and all other inducements until the prospect of cherry ice cream made its appearance. If the potential justifying reason here – that the agent will be restored to health – fails to be an actual justifying reason because it does not find a corresponding motivation in the agent, then we seem forced to say that taking the treatment was rationally required, but that there was nevertheless no reason that would justify the patient in undergoing it. This simply sounds too strange. We should favor an alternative description, if a plausible one is available. And one is: the action was both required and justified in terms of objective rationality, which is the sense of rationality to which – as we have already seen in chapter 4 – reasons are directly relevant. That is, there is a reason for the treatment that justifies undergoing it. But despite its favorable objective rational status, and despite the presence of the justifying reason, it remains the case that the agent performed the action for the wrong reasons. Because of this, and because (in this case) such a performance was the result of his being insufficiently averse to the painfulness of the treatment (or of his being much too concerned with cherry ice cream) his action was subjectively irrational. That is, we can say the following three things in this case: that undergoing the treatment was objectively rationally required and justified; that there was a justifying reason to do so; that the actual agent's undergoing of the treatment was subjectively irrational. One important consequence of this discussion is

therefore that it highlights one context in which a failure to distinguish subjective from objective rationality is likely to result in a wrong conception of practical reasons. In the case of 'acting for the wrong reasons,' the simple stipulation that the agent is fully informed is not guaranteed to bring the subjective and objective rationality of an action into agreement. If, in such a context, one tries to gain insight into practical reasons by stipulating full information and then using the intuition that a certain action is *subjectively* irrational, one will be led into error. For it is to *objective* rationality that practical reasons are directly relevant, and in this context one will have come to a wrong conclusion about the objective rational status of the action.[18] No argument in this book has made use of an intuition that a particular action was subjectively irrational in support of any claim about that action's objective rational status, unless, in addition to stipulating full information, it was also made clear that the relevant action would have been subjectively irrational *no matter what its etiology*. For example, if a fully informed agent performs an action, *the only relevant consequences of which* will be a moment's pleasure and a lifetime's misery, then the subjective irrationality of this action implies its objective irrationality because there is no way such an action could be subjectively rational.

A second reason for regarding the justifying role of reasons as relative to relevant action types, and not to action tokens, is the following. There is reason to wonder what exactly the justification of an action token could possibly be, if it is not simply the justification of the relevant action type. For it is very plausible that arguments seeking to show that an action is required or justified must proceed in general terms if they are to make the status of those actions intelligible.[19] That is, if the available reasons show a certain action token to be, say, rationally justified, then it seems fair to say that any other action that is similar in the relevant respects must also be rationally justified by the same reasons. If this is right, then what seems to be the justification of an action token is really only the justification of

[18] See pp. 8–9 for a discussion of Gibbard's equation of subjective and objective rationality under conditions of full information. See also pp. 69–72 for other cases in which a focus on subjective rationality misleadingly suggests that, for example, normative judgments provide reasons because acting against one's normative judgment is a species of subjective rationality. At pp. 163–64 many distinct forms of subjective rationality are described, and each of these can also form the basis for mistaken conclusions regarding what sorts of considerations provide normative reasons.

[19] See Raz (1999b), p. 220.

the relevant action type. Thus, I conclude that the justification of actions is always the justification of action types. Both the heroic rescue, and the rescue motivated by a pathological desire for a second date, may be equally rational. For despite differences in the motives that produced them, they may still be tokens of the same relevant action type. This is especially plausible when one recalls that the relevant type, for purposes of assessing objective rationality, does not include the motives from which the action springs. Moreover, any difference between the two rescues does not seem likely to be of much consequence to their objective rational status, for they are alike in their most important reason-giving features: both involve the risk of drowning, and the chance to save several people from a similar fate.

Because the two rescues are instances of virtually the same relevant type, either both rescues are objectively rational or both are objectively irrational. It is objectively rational to save several people at the risk of one's own life, though it would also be objectively rational to decline to do so, in favor of one's own safety. This means that the reason 'that the action will save the lives of several strangers' is quite a strong justifying reason, for it can justify actions that involve a significant risk of death for the agent. So this altruistic reason is a justifying reason independent of the motivations of the agent. Humean internalism is therefore false of such altruistic reasons. Is Kantian internalism therefore true of these reasons? Not necessarily. For if it is subjectively rationally permissible, but not required, to act on this reason at the risk of one's own life, then it seems possible that even a fully rational agent might be unmotivated to act on the reason. This would mean that externalism was true of justifying reasons: that even a fully rational agent might have such a reason and be unmotivated by it. And in fact, even if all rational agents must be motivated to some minimal degree by such a reason, in which case Kantian internalism would still be *technically* true of such justifying reasons, it would also have lost much of its interest. For the strength of such a reason, as a justification, would not correspond to this minimal rationally required degree of motivation. Rather, this minimum would correspond to the strength of the reason *as a requirement*. When one is determining the justifying strength of the altruistic reason, it doesn't matter whether all rational agents would be motivated by it to some minimal degree. The technical truth or falsity of Kantian internalism is irrelevant when one is considering reasons in their justifying role.

My conclusion therefore is that justifying reasons do not depend in any important way on the antecedent motivations of the agent, so that Humean internalism is false of them. And Kantian internalism is also false of such reasons, since it seems that even a fully rational agent might not be moved by them. To the claim that a fully rational agent would be motivated to some minimal degree by primarily justifying reasons, the response is that this minimal motivation is completely unconnected to the justifying role of those reasons. Instead, it gives a measure of the minimal requiring strength that those reasons possess. Therefore, when considering reasons in their justificatory role, externalism is the most illuminating view. Humean internalism only seems plausible when one wrongly looks to subjective rationality for direct insight into the existence of justifying reasons. That is, there are cases in which a justifying reason fails to motivate an agent, and in which that very reason provides the justification without which the action (which is objectively rational) would have been objectively irrational. In such cases the reason does not make the agent's action *subjectively* rational. And this can lead to the wrong conclusion that the reason is in fact not a reason.

EVIDENCE

Is a failure to appreciate the justifying/requiring distinction really hindering progress in the current internalism/externalism debate? It should seem very plausible that it is. Requiring and justifying simply are two conceptually distinct but equally important roles for normative reasons. If an argument for some particular form of internalism were to rely at some point on the assumption that reasons provide *pro tanto* rational requirements, this would render the argument invalid for reasons that do not provide such requirements. And if an argument relied on some feature of justification, this might render the argument irrelevant to the capacity of reasons to require. Despite these *a priori* claims about the destructive effect of a failure to recognize the justifying/requiring distinction, it may be worthwhile to point to at least a few places in which this failure has a noticeable effect.

In his seminal "Internal and External Reasons," Bernard Williams argues against externalism and against Kantian internalism by arguing against the existence of reasons that are completely independent of the contingent motivational setup of the agent. One now famous example is

that of Owen Wingrave, whose father (on Williams's rephrasing) insists that Owen has a reason to join the army. Williams explains that Owen hates the thought of the army, and has no other existing motives from which rational processes could generate a desire to join the army. Now, it is admitted by internalists and externalists alike that Owen might well not have any reason to join the army, so the example may perhaps not have been the fairest that Williams could have chosen. But, in fairness to Williams, he is here concerned not so much with the truth of Owen's father's claim, as with what such a reason claim could possibly mean. Williams thinks that the only thing the externalist could possibly mean by a reason claim is that the agent would be irrational if he failed to act on the reason: that the reason provides a rational requirement.[20] Under this assumption, Owen's father's claim certainly looks like bluff, and that is what Williams regards it as. But once one recognizes the possibility that some reasons might serve only to justify, one can interpret the externalist merely as saying that the reason would make it rationally permissible to perform the action, despite the reasons against it. Perhaps even this is false in the Owen Wingrave example. But there are other examples the externalist might use, and Williams simply never addresses this possibility.

Here is an additional example. John Tilley, in "Motivation and Practical Reasons," provides perhaps the clearest illustration of the acceptance of internalism based on the undefended assumption that all reasons are *pro tanto* rational requirements. It is indeed a great virtue of Tilley that he explicitly articulates the premise that does so much covert work in other defenses of internalism. He puts the claim in the following form:

(1) Reasons are facts we are rationally required to act upon. That is, if F is a reason for A to do D, and A is both aware of F and without any reasons that compete with F, then A is rationally required to act on F and do D, assuming she is not hindered from so acting.[21]

[20] Williams (1981), pp. 110–11. Williams is clear about what he himself means when he claims that an agent has a reason to do a particular action: he means that there are rational processes that could produce a motivation to do the action for which it is claimed that there is a reason. Interestingly, on a charitable reading this does not mean that the agent, if rational, will, as a matter of necessity, go through the relevant process, and will therefore, with equal necessity, be motivated to do the action. So Williams's own reason claim may not need to be taken as expressing a rational requirement.

[21] Tilley (1997), p. 113.

Internalism then follows.[22] For if an agent is such that she would not be moved to action by a given reason when opposing reasons were removed, then she is disposed to violate a rational requirement, and is not fully rational. But no defense is offered for (1), even though Tilley acknowledges that it plays a crucial role in arguments for moral subjectivism, which he rightly sees as a troublesome conclusion. He simply does not see (1) as controversial at all, explicitly claiming that even externalists agree with it. But of course he has in mind the sort of externalists, like Parfit and Brink, whom I have placed in the Kantian internalist camp. Had Tilley appreciated the difference between justifying and requiring, it would have been far more difficult for him to have regarded (1) as uncontroversial and axiomatic.[23]

CONCLUSION

In current normative theory there are at least two versions of internalism about practical reasons. We may call them 'Humean internalism' and 'Kantian internalism.' And there is a strong version of externalism that denies both internalist views. This chapter has made a suggestion about what the correct view is: that there is no unique correct view that applies to all types of reasons. Rather, Kantian internalism is true of reasons insofar as they play a requiring role, while externalism is true of reasons insofar as they play a justifying role. Regardless of whether each of the arguments for these particular conclusions goes through, the distinction between requirement and justification remains. As was pointed out at pp. 68–69, this is a logical distinction that does not depend at all upon any controversial substantive normative claims. Despite its logical nature, this distinction has not been acknowledged or even noted by anyone who advocates any form of internalism. But because justification is not the same as requirement, it is unlikely that any one form of internalism could possibly be the whole story. Philosophers who advocate a unified internalism will have to show, at least, that requiring and justifying strength necessarily co-vary, or

[22] Tilley's rendering of internalism, on the same page, is the following: "If F is a reason for A to do D, and A is aware of F, then barring all impediments and practical reasons that compete with F, agent A will be moved to D by her awareness of F – assuming she is rational."

[23] Christine Korsgaard, in Korsgaard (1996b), makes the same assumption as Tilley, and infers the same conclusion. She is not as explicit as Tilley, however, since her primary goal in that paper is not to establish internalism, but rather (correctly) to expose a significant undefended assumption behind Humean internalism.

that (more surprisingly) their view applies equally well to both. Perhaps in logic or mathematics justification and requirement are indeed the same. That is, perhaps any theorem that is justified by the rules of logic is also, in a sense, required by them. But it is a characteristic mistake of philosophers to take the simplest and most formal models as the purest and most ideal. Any attention to real practical arguments that are offered about actual matters of importance will show that practical rationality is completely different from logic or mathematics. It is something vastly more complex, about which no simple claims are likely simply to be true.

9

Brute rationality

One significant implication of the view of rationality offered in chapter 7 is that as long as an action does not stem from the kind of mental malfunction that would put the agent at increased risk of suffering harms without compensating benefits for anyone else, that person's action is subjectively rational. However, many contemporary philosophers hold that for an action to be rational in this sense, or even intelligible, it must somehow involve the judgment, by the agent, that the ends of the action are good. For example, Jonathan Dancy, Warren Quinn, Joseph Raz, and Thomas Scanlon have recently and independently presented theories according to which intentional action is action undertaken for a reason, and undertaking an action for a reason requires that one see something in the action as being of value, or as being a reason-giving feature. Not surprisingly these philosophers also hold that we have desires for reasons, at least when these desires are not simply urges that seize us. Having a desire for a reason involves, for Scanlon and for Raz, the judgment that the object of the desire is good in some way, while Quinn holds that the same sort of judgment is required in order for an action to be rational.[1] And Dancy holds that the reasons for which an agent acts, whether good or bad, must at least be regarded by the agent as favoring the action.[2] Despite many points of disagreement, all of these philosophers would be able to agree to the following general claim: rational action involves the making of normative judgments.[3] The concern

[1] Raz (1999b), pp. 8, 23, 62, 291; Scanlon (1998), pp. 18, 23–24, 33–35, 56–57; Quinn (1995), pp. 195, 200, 203, 205.

[2] Dancy (2000), pp. 129, 136.

[3] In what follows I will make the simplifying assumption that the agent is not mistaken or ignorant about any relevant nonnormative matters of fact, such as whether the liquid in his glass is gin or gasoline. Even with this assumption, Scanlon would not agree with the claim to which this note is attached, if 'rational' is interpreted in his idiolect. Unfortunately Scanlon has no term that captures the sense of 'rational' as intended here, although it is a common sense. In Scanlon's terminology, what is meant by 'rational' in the present context would be expressed by the phrase 'not open to rational criticism, except with regard to

in this final chapter is primarily to deny this view of rational action. In denying that rational action involves the making of normative judgments, I will advance the same claim about intentional action more generally. This is because rational action is a kind of intentional action. And indeed all of the philosophers discussed here hold their view of rational action *because* they hold the same view of intentional action: that it also requires the making of normative judgments. Rational action, for these philosophers, is then simply intentional action that *actually is* justified by the reasons for which the agent acted. Thus, the locus of dispute in what follows will often be the nature of intentional action, even though the ultimate goal is an account of rational action that does not involve the making of normative judgments. It will not be possible to specify the notion of rational action at issue here with great precision, since the four philosophers being discussed would themselves be unable to agree completely on any proposal, and would surely disagree with many aspects of the theory of rationality offered in this book. But we could perhaps all agree that rational action is intentional action that is free from failures of instrumental rationality, and is a response to the reasons of which the agent is aware, in a way that is appropriate, given the normative significance of those reasons.[4]

An additional theoretical commitment common to the four philosophers mentioned above is a denial of the basic normative significance of desire. Dancy, Quinn, Raz, and Scanlon would all find something substantially correct in the following: it is not desires that rationally justify or require actions; rather, it is the reasons that stand behind those desires. Let us call this view 'the objective reasons thesis.' The objective reasons thesis is set up in opposition to neo-Humean views of rationality that regard desires as the source of all of our practical reasons. The view offered in this book also strongly opposes the neo-Humean view, and I have argued against it in a number of places in previous chapters. But because the arguments of the present chapter are aimed at philosophers who, like Dancy, Quinn, Raz, and Scanlon, have also already rejected such views, I will not repeat these arguments. Rather, the purpose of this chapter is to separate the objective reasons thesis, which is a view with much merit, from the

nonculpable ignorance or mistake about nonnormative matters of fact.' Understanding 'rational' in this sense, and given the assumption of full and accurate information, Scanlon could agree with the above claim. See below for more remarks on problems with Scanlon's terminology.

[4] Again, for terminological reasons Scanlon would dissent from this as a characterization of *rational* action. But he would, I hope, concede that the notion described here is an important one.

view that rational action requires agents to make normative judgments, however implicit, unconscious, or inchoate. This latter view, which we might call 'the judgment thesis,' misrepresents and overintellectualizes the vast majority of our everyday choices, desires, and actions.

WHAT ARE NORMATIVE JUDGMENTS?

The plausibility of the judgment thesis, and the plausibility of attributing it to Dancy, Quinn, Raz, and Scanlon, will depend a great deal on what we take normative judgments to be. For purposes of present discussion, normative judgments are judgments that something is good or bad, or of value, or that something provides a reason.[5] But is important to make clear at the outset that these judgments need not be explicit or conscious. None of the philosophers mentioned above hold the extreme view that explicit normative judgments are required for desire or for action. And yet the judgments at issue cannot be mere dispositions to agree to certain normative propositions. They actually play a role in generating motivation, and so they must be regarded as psychologically present in some fairly strong sense.[6]

Perhaps the weakest interpretation of the claim that an agent is making a normative judgment is that something 'appears of value' to the agent, or 'appears to be a reason.' By way of comparison, it is certainly possible for something to appear blue to a perceiver without that perceiver having the concept 'blue' in any sense other than that in which babies or apes could be said to have that concept. If the normative judgments that Dancy, Quinn, Raz, and Scanlon have in mind are like this, then they could certainly avoid the charge of overintellectualizing our desires and choices.[7] The question of importance for this weak understanding of normative

[5] There are important questions about the relation between value and reasons. But since the arguments of this chapter will deny that any sort of normative or evaluative judgment is typically involved in desire or intention, the distinction between value and reasons will not affect any of the arguments that follow.

[6] The word 'generating' is here meant to be neutral as between causal and noncausal accounts of the relation between reasons (or beliefs about reasons) and motivation. For Dancy, for example, the above claim is meant to be compatible with the idea that an appropriate normative judgment is one of the 'enabling conditions' for a reason to be an agent's reason.

[7] On the other hand, and as we will see, such a characterization runs the risk of blurring the line between cognitivist and noncognitivist accounts of motivation. For example, Philippa Foot and Simon Blackburn both place 'recognition of reason for action' among the attitudes and feelings that are characteristic of noncognitivist views. See Blackburn (1995), p. 36.

judgments is whether it makes sense to think of something 'appearing good' or 'appearing to be a reason' in the same way that it makes sense to think of something appearing blue.

One argument in favor of this view comes from Scanlon.[8] We all understand what it would be like for someone to seem, for example, trustworthy, when we actually judge that she is not. This suggests that there can be normative appearances that are distinct from explicit normative judgments, and that can even conflict with them. Scanlon takes this case of seeming trustworthy to involve a normative appearance that parallels the "vague appeal to a normative category" that he asserts is involved in typical desires. But what is going on in such cases? When something appears blue although we explicitly judge that it is not really blue, we can explain what is going on in the following way: the agent's visual experience is as if the object perceived were blue. But we cannot explain Scanlon's example in the same way unless we posit a special faculty that perceives trustworthiness, or, at the very least, a distinctive phenomenology of trustworthiness. Then we could explain normative appearances by saying that the appearances produced by this faculty were as if it were mediating the perception of someone who was really trustworthy, or that the phenomenological experience of the perceiver was of the distinctive sort typically produced by trustworthy people. But such a special faculty represents quite a hard bullet to bite: the arguments of this chapter will have accomplished quite a bit if they compel anyone to bite it. And the idea that there is a distinctive phenomenology of trustworthiness is hardly more appealing.[9] Moreover, there is a straightforward alternative. We can say that because of the unconscious influence of features of the seemingly trustworthy woman's behavior, facial expression, and so on, we are strongly disposed to believe what she says, or to follow her advice, or otherwise to respond to her in ways that are appropriate for genuinely trustworthy people. If we also have the explicit belief that she is *not* trustworthy, then if we notice that we often (for example) form beliefs based on what she

[8] It is unclear how this solution would be applied to account for the "standing normative judgments" to which Scanlon (1998), p. 24 appeals in explaining our unreflective desires and choices.

[9] This point is very similar to the Wittgensteinian point that there is no distinctive or essential phenomenology of 'grasping' or 'understanding' a rule. Rather, what is important is that one can go on in the right way. And since one may *think* that one can go on in the right way, or one may confidently *try* to go on in the right way, when one cannot, there is a place in our language for phrases like 'it seemed to me that I understood, although I did not.' The fact that there may sometimes be a certain felt experience at the moment of comprehension is no argument against Wittgenstein.

says, this is an important piece of information, worth being able to report and discuss. In such situations, we say that although she appears trustworthy, she is not. But this does not mean that anything is going on here that is like an object's appearing blue. Rather, it is simply a way of noting that the untrustworthy woman elicits behavior or beliefs or feelings in a way that is appropriate only when dealing with a genuinely trustworthy person.[10]

On an analysis of the above sort, what would it be for an end achievable by action to seem good? It would simply be for it to elicit responses that are appropriate when dealing with an end that genuinely is good. What are such responses? They are simply desires, or states of motivatedness, or affective states of that sort. Thus desire is not to be *explained* by normative appearances in the strong sense that the judgment thesis maintains. Rather, the notion of 'appearing good' gives us a way of talking about desires (and other affective states) for things that we judge not to be desirable. This is consistent with the claim that the notion of 'appearing good' also gives us a way of talking about things that we *do* (or would) judge to be desirable, based on their appearance: 'That peach appears particularly good.' But there is a strong case to be made that seems-talk is parasitic on is-talk, even in cases where 'seems' can legitimately be taken to suggest 'is.' That is, it is unlikely that we would have seems-talk in a domain if there were no possibility of misleading perception, wrong judgment, or inappropriate reaction in that domain. And the claim here is that for blueness, the relevant possibility is misperception, while for goodness and other normative terms, the relevant possibility is inappropriate affective response.

None of this requires us to deny that desires are quite often to be explained by reference to one's awareness of something that is *in fact* good: for example, by reference to one's awareness that one's action will give pleasure to a friend.[11] It is only to deny that desires are to be explained by reference to the fact that something *appears* to be good: not even in those

[10] I do not mean to suggest that there is a uniquely appropriate pattern of response to trustworthy people. Rather, there are a wide variety of responses to particular people that would only have been appropriate had the person been trustworthy, and when we note that we tend to have a significant number of these responses, despite our knowledge of the untrustworthiness of the person, we can usefully say that the person appears trustworthy. This same point goes for other normative properties and the responses they characteristically elicit.

[11] Awareness of something that is in fact good is different from awareness of the fact that something is good. One of the targets of the current argument is a tendency to conflate these two things.

cases in which the thing that motivates is in fact *not* good. In both cases, the desire can be explained by awareness of the fact: whether it is the fact that the action will help a friend, which does give a reason, or the fact that the action will hurt someone of whom one is envious, which (arguably) does not. In the latter case, if we say that the fact that the action would hurt the person 'appeared to the agent to provide a reason' we need only be taken as indicating that the agent desired to do the action because of it.[12] We need not be committing ourselves to the idea that the prospect of hurting the other person appeared *as having any normative property*. One may wish to object here that desires can clearly be explained by reference to something's having *appeared* to be the case to the agent; perhaps it appeared to the agent that a certain object was a real peach, when in fact it was only a convincing fake. But there is no need to deny this, for even in such cases the relevant judgment would not be a normative one.

One should not be fooled by the use of the same word in 'appears blue,' 'appears trustworthy,' and 'appears good.' This mere linguistic similarity is very weak evidence that anything formally the same is going on in each case. There are, admittedly, similarities between appearing blue and appearing good: the linguistic similarity does suggest this much. But it is very plausible that perceptual appearance is the paradigmatic case, and that the other cases are derivative, or that some other relation holds between them. In cases in which an object appears blue, an agent will tend to behave as if the object really were blue, until the agent comes to believe that it is not actually blue. And this behavior is to be *explained* by a phenomenological

[12] It seems to me that this is the strongest claim that Dennis Stampe can defend in Stampe (1987), although it may be sufficient for his purpose in that paper, which is to show that the fact that one desires something gives one a (possibly quite bad) reason to pursue it. He writes that "desires are reliable indicators of what would be good, and that their authority involves nothing more, and nothing less" (p. 374). But this 'nothing more' means that their authority does not depend on their involving a *presentation of their object as good*. And indeed, I think that Stampe's "seems$_d$" can simply be taken as a misleading technical replacement for 'is desired.' I don't deny that desire shares many points of similarity with perception – including being a fairly reliable indicator of something – but taking the point as literally as Stampe does forces him to look for an *existing object* of perception, which he finds in one's own physiological state. This does not sit well with his claim that the desire that *p* is a state ideally caused by the fact that it would be good that *p*. For suppose I desire my daughter to marry well. This desire, ideally, would be caused by the fact that it would be good for her to marry well. But the desire is caused by some physiological state of *mine*. In order to avoid the obviously false claim that the goodness of my daughter's marrying well actually *is* a state of my own body, Stampe instead makes the claim that the relevant state of my own body is *such that* it would be good if my daughter married well. See Stampe (1987), pp. 355, 372–75.

experience that the agent is having, by way of his faculty of sight. Similarly, in cases in which a person appears trustworthy, or an action appears good, the agent tends to behave as if the person were really trustworthy, or the action really good, until the agent comes to believe that the person is not really trustworthy, or the action is not really good. This similarity is sufficient to explain the use of the word 'appears' in both types of case. But in the latter case the antecedent behavior – believing what the seemingly trustworthy person says, taking her advice, etc. – is not to be explained by any phenomenological experience the agent is having, distinct from the experiences of sight, sound, and so on. Rather, it is to be explained by the following fact: that we humans are set up to respond more or less automatically to certain perceptible features of the world, including other people's tone of voice and facial expression.

A defender of Scanlon may wish to point out here that objects appear to us as cups or trees all the time, and that this fact does nothing to imply that we have a special faculty that perceives cups or trees. Nor does it suggest that there is any distinctive phenomenology of cup or tree perception. When we say that something appears as a cup, this can simply mean, for example, that the object looks like a cup would be expected to look under normal conditions, or that the way it looks would lead one to believe that it is a cup.[13] Why cannot 'appearing to be a reason' be modeled on this sort of appearance? Why need it be compared to appearing to have some basic phenomenal quality like blueness? The answer is: let 'appearing to be a reason' be modeled on as complex a sensory appearance as one likes, it will not help Scanlon. For, as William Alston points out, when we understand 'appears as' in these ways, "a phenomenal concept is, so to say, always in the background, even when not explicitly employed."[14] Suppose then that we interpret 'appears as a reason' as 'appears as a reason would appear under normal conditions.' What is the phenomenal concept in the background? It does not seem that there always is one. Sometimes, perhaps, a reason elicits a felt desire, or some other felt affective response. Perhaps sometimes there is even a distinct phenomenology of something appearing to be a

[13] These interpretations of 'appears as' are what William Alston calls, respectively, the *comparative* and the *doxastic*. He also describes an *epistemic* concept, and presumably there are a good number of other ways of interpreting our 'appears as' claims. See Alston (1991), p. 45. It may be worth noting that the doxastic interpretation would not be an attractive way for Scanlon to understand the normative appearances that he claims stand behind desires, since children have desires before they have sufficient conceptual apparatus to form a belief that something is a reason.

[14] Alston (1991), p. 45.

reason: a felt experience that one would wish to describe in exactly this way. But it is just implausible to assert that this sort of experience precedes or underlies all desire, or even all rational desire.[15]

Now, in attempting to understand what the normative judgments might be that Dancy, Quinn, Raz, and Scanlon have in mind, it is not extremely helpful merely to be told that they are not to be understood by strict analogy with perceptual appearances. But surely the burden here is on these philosophers to explain what they mean when they speak of an agent's taking a fact to show an action to be good, or an agent's conceiving a consideration as favoring an action.[16] Perhaps Scanlon has done so, in offering his analogy with perception. But that analogy cannot be taken strictly. So it remains for Scanlon to explain how the analogy is actually to be taken. In what follows it will be assumed, for the sake of argument, that there is some plausible understanding of 'normative judgment' that makes such judgments less explicit than occurrent thoughts, but also gives them a distinctively normative content. And it will be argued that no such normative judgments are required for rational action, or for intentional action more generally.

THE BASIC PICTURE

The view offered in this book does not merely dispute the claim that rational action requires agents to make *correct* normative judgments. It holds that rational action, and intentional action more generally, does not require the making of *any* normative judgments. While agreeing that virtually all of our desires, even at the most basic level, are held for reasons, and that virtually all of our actions are undertaken for reasons, I also hold that it is quite common for human action and desire to involve no normative judgment, even of an unconscious, inexplicit, inchoate sort. Rather, it is sufficient for one to act for a reason that there be a reason to act, and that one act because of it – at least if the 'because' here is interpreted as indicating the presence of the right sort of causal or psychological mechanism.[17]

[15] Stampe (1987), p. 359 seems to appreciate the truth of this point, but not its significance.
[16] These phrases come from Raz (1999b), p. 24 and Dancy (2000), p. 129.
[17] This qualification is necessary to deal with the problem of wayward causal chains. To take a well-known example, I may become so unnerved by my sudden awareness of a reason to kill my mountain-climbing partner that my hands begin to shake and I end up letting go of the rope, killing him. This should not count as an action done for a reason. The problem of characterizing the 'right way' in which the reason must be a cause of my

Similarly, it is sufficient for one to desire something for a reason that there be a reason to desire it, and that one desire it because of that reason. In neither case need one judge that the reason is a reason. One need not even have the concept of 'a reason.' And as long as the reasons for one's actions or desires are sufficient to justify them, that is all it takes for those actions or desires to be rational.

Here is the picture of reasons and rational action that will be set up against the judgment thesis. Although the first two claims, which concern normative reasons, are part of the developed theory worked out in previous chapters, the five points taken together are by no means intended as a full-blown theory of rational agency. Rather, they are offered as a framework within which one might develop such a theory.[18]

(1) There are facts of the matter about whether or not there are reasons that speak in favor of particular actions, and facts about what those reasons are. For example, it is a fact that the prospect of having one's fingers burnt provides a reason to avoid touching a certain very hot object, though it can be rationally permissible to act against such a reason, if there are countervailing reasons of sufficient justifying strength.

(2) Reasons for actions are, at least generally, completely independent of the desires or other noncognitive attitudes of the agents who have them. This claim and the previous one express views that are shared by Dancy, Quinn, Raz, Scanlon, and myself. Again, because this chapter is directed primarily at these philosophers and at those who hold similar views, no additional defense of these first two claims will be offered here. This chapter does not attempt to persuade Humeans to abandon their desire-based views, but only to convince those who already reject such views that they should take a further step. Nevertheless, in order to clarify the view being offered, it may be worth noting that claim (2) goes significantly beyond claim (1). Claim (1) is consistent with the view that the reasons that speak in favor of a particular action are somehow determined by the desires of the agent – either as she is, or in some ideal state. Claim (2) denies such a view.

action is beyond the scope of this book. See Davidson (1980), p. 79 and Mele (2000). It may be worth noting that simply adding an appropriate normative judgment to the causal or psychological mechanism does nothing to eliminate the possibility of these sorts of wayward chains.

[18] For readers worried that no plausible theory will fit into the basic picture, it may be worth mentioning that one of the most influential causal accounts of intentional action, Alfred Mele's, is fully consistent with the view advocated here.

(3) A desire may be understood in an intuitive and very broad way as a disposition to act so as to bring about certain states of affairs. We can call this state of affairs the *end* of the desire. 'End' should be as broadly understood as 'desire' here, so that it can include, for example, the state of affairs of my avoiding some pain, or the state of affairs of my acting as God commands. This particular picture of desire, however, is not at all crucial to anything that follows. Those who favor some alternate account of desire should be able to substitute their own favored picture, modified in relevant ways, into the basic view I am here presenting. In particular, those who wish to regard desire as a state of motivatedness, where the motivation comes from the antecedent apprehension of some fact, should feel free to do so. Such states of motivatedness will still have ends, as they are understood here. What is important is that ends are states of affairs that would be realized or promoted by acting successfully on a desire. Whether one has the desire because of one's logically antecedent perception that such ends could be achieved by one's action, or whether desires are somehow brute, is neither here nor there for current purposes. The only accounts of desire that are incompatible with these purposes are accounts that tie desires conceptually to normative judgments.

(4) If an agent acts on a certain desire, then the end of that desire – something quite distinct from the desire itself – forms part of the *reason why* the agent acted as he did.[19] One and the same thing can be both a *reason for* action, and a *reason why* an agent does an action. In fact, it is typically true that the reasons why an agent acts are also reasons for acting in that way. But this does not mean that there is only one notion of reason in play. For one thing, it is *not* typically true that all the reasons for an action are reasons why an agent acts. One important reason for this is that there can be reasons of which the agent is completely unaware, or which the agent has insufficient conceptual apparatus to appreciate. And even when we restrict attention to the reasons of which the agent is aware, generally there will be too many for an agent to take them all into account. Moreover, it seems possible that a language could have the means to express the notion of a reason why, without having the means to express the notion of a reason for. It is only the notion of a *reason for* that is normative.[20] In giving the reason why an agent

[19] 'Part of,' only because there may be more than one desire involved. In cases in which an agent clearly acts because of one and only one dominant desire, then the end of that desire simply is the reason why the agent acted as he did. The reasons why an agent acts as he does are always to be given in terms of the ends of the agent's relevant desires.

[20] That the notion of 'reason for' is a normative *notion* is consistent with the claim that particular reasons for action are themselves nonnormative matters of fact. For example, it

acted as he did, one is neither making a normative claim, nor imputing one to the agent. One is only making a claim about what motivated the agent. Quite commonly the things that motivate agents are nonnormative matters of fact, as that an action is the only means of avoiding getting one's feet wet.[21] In the following discussion the bare term 'reason' should always be understood in the normative sense, as meaning 'reason for': what Dancy calls 'a good reason.'[22] In order to avoid confusion, the terms 'motive,' 'end,' 'motivating reason,' or 'explanatory reason' will generally be used instead of 'reason why.' Also, when the phrases 'reason why' or 'explanatory reason' are used, they should be interpreted in a stronger sense than merely 'causally explanatory.' Rather, they indicate reasons that play the right sort of causal or psychological role in the production of the action.

(5) By the time a normal human being has grown to adulthood, she will have acquired the concept of a reason. Normal adults can, unsurprisingly, make wrong judgments about reasons, just as they can make wrong judgments about the colors of objects, but typically their judgments are correct. Moreover, although there are disagreements about whether something is a reason, or about how much it can justify or require, these disagreements are as inevitable, marginal, and conceptually unimportant as disagreements as to whether something is green or blue. This is shown by the fact that almost all of us can see the reasons for almost all of the actions that almost all people actually perform, whether we approve of them or not.[23]

is a reason for taking some medicine *that it will stop the pain*. That this fact has normative significance does not mean that it itself is a normative fact. Normative facts are of the following sort: that pain is bad, that providing pleasure to others provides a reason, etc.

[21] Or again, if one likes, what typically motivates an agent is her *awareness* of, or *beliefs* about, such nonnormative matters of fact. For current purposes there is no need to enter this debate.

[22] Dancy (2000), p. 4. Dancy would perhaps deny that there are two senses of 'reason,' holding instead that there are two sorts of questions that reasons are intended to answer. Since I do not dispute Dancy's claim that normative and explanatory reasons are the same *kind* of entity – states of affairs – this disagreement is relatively unimportant.

[23] In fact, the general correctness of our reasons-beliefs is not essential to the argument, since it only bears on how many of our actions are rational, and not whether our desires are produced by these beliefs. Nevertheless, see Scanlon (1998), p. 71 for similar skepticism about widespread disagreement as to what counts as a reason. In order to appreciate this point it is important to realize that agreement about what counts as a reason, and agreement about how much such reasons can justify or require, is perfectly consistent with widely divergent tendencies to act on those reasons, even among rational people. This is because normative reasons are plausibly regarded not as uniquely fixing a rational choice (or small range of optimal choices), but rather as setting limits on a relatively wide range of rationally permissible options. See Raz (1999b), p. 100.

Here is how the basic picture describes some typical rational actions. One reason for desiring to eat ice cream, and for eating it, is that one will get some pleasant sensations by eating it. If a three-year-old desires to eat ice cream, and eats it, because he knows he will get those pleasant sensations, then that child acts for a (motivating) reason, and also has a (motivating) reason for its desire. Since that motivating reason is also a normative reason, we can say that the child acts for a *good* reason. If this reason is unopposed by any other reasons of which the child is or should be aware, and which would make it irrational to eat the ice cream, then the action is rational. If a fifty-three-year-old desires to eat ice cream, the same things are also true. In this latter case the ice cream eater, being older and wiser, does indeed have normative concepts, and would probably immediately assent to the claim that pleasure was good, and that the pleasure of eating ice cream provides a reason to eat it. But these facts about what the older agent would immediately assent to are the result of his having, unsurprisingly, certain concepts and beliefs. Such an agent would also probably immediately assent to the claim that ice cream is a dairy product, or, for that matter, that chickens are animals. But none of these beliefs has anything to do with the genesis of his desire to eat ice cream, or with his eating it, or with the fact that his eating it is a perfectly rational action. For philosophers who are disposed to believe that desires are somehow the *result* of normative judgments, the immediate assent that most adult agents would give to the relevant normative judgments will seem to confirm their opinion.[24] But this immediate assent is also explicable as the result of motivationally inert basic knowledge of what is good and what is bad.

EXPLAINING THE JUDGMENT THESIS

Why are Dancy, Quinn, Raz, and Scanlon all drawn to the judgment thesis? There is no unified answer to this question. Each philosopher says slightly different things in its defense. But very often one of them can be seen as filling a gap in another's argument, or as answering a further objection. The following four sections attempt to do three things: to present the relevant arguments, with some indication of how they reinforce each other, to explain how one might come to hold them, and to criticize them, lending support to the basic picture outlined above.

[24] See Raz (1999b), p. 233.

Quinn

Quinn endorses the judgment thesis at the end of an argument against the view that desires, understood as bare functional states, could possibly justify action.[25] His argument makes use of the example of a person who is disposed to turn radios on whenever he sees them off. This person sees nothing good in doing this, but is quite strongly disposed to do it anyway. Quinn reasonably expects his readers to agree that such a person has no reason for turning on radios simply in virtue of being in this functional state. Perhaps the person might feel uncomfortable until he turned on a visible radio. And in such a case he might have some reason to do so. But then it would be the prospect of relieving the discomfort that provided the reason, and not the stipulated end of the desire, which is simply to have the radio switched on. Quinn then goes on to point out that if the bare disposition to act in a certain way does not provide a reason in the bizarre cases, then it does not provide a reason in the normal cases either. Rather, in the normal case, in which an agent is disposed, for example, to promote his own health, it is not the disposition that provides the reason to do so, but something else: something to do with the goodness of the prospect of increased health. It is at this point in the argument, up to which I wholeheartedly agree, that Quinn makes an additional unnecessary move and commits himself to the judgment thesis. Here is the relevant passage.

> That I am psychologically set up to head in a certain way cannot by itself rationalize my will's going along with the set-up. For that I need the *thought* that the direction in which I am psychologically pointed leads to something good (either in act or result), or takes me away from something bad.[26]

The first sentence here is correct. Simply being set up to switch radios on does not give one a reason to do so. Nor does being set up to pursue pleasure and health give one a reason to pursue these things. Something else is needed. But Quinn adds the wrong thing. It is not the *thought* that the end of my desire is good that makes action rational; it is the *fact* that

[25] Quinn (1995), pp. 189–95.

[26] Quinn (1995), p. 195, italics Quinn's. It is possible to give a *de re* reading to the content of the thought referred to in the second sentence. Then the thought would be the following: 'the direction in which the agent is psychologically pointed leads to something which is, in fact (but perhaps not in the agent's thought) good.' But this reading is inconsistent with Quinn's repeated assertions elsewhere that "some kind of evaluation" is "typically present in basic desire" (p. 200) and that "desires and preferences rationalize only because of the value judgments they involve" (p. 201).

the end of my desire is good.[27] That fact alone is enough to explain the difference in rational status between actions based on a desire to switch on radios, and actions based on a desire to maintain one's health. The claim here is not that merely having a genuinely good end is sufficient to make one's action rational. Surely it is not, since one may well be irrational in pursuing even a good end. Rather, the claim here is only that what accounts for the difference in rational status between switching radios on and pursuing health need not be the presence of a normative judgment. Rather, one can give a sufficient explanation for the difference in status in this case simply by reference to the pointlessness of the one end, and the goodness of the other.

Quinn's own strategy for arguing against the normative significance of desire works just as well against his own view that normative judgments have normative significance. Consider a person who not only desires to switch radios on, but who also makes the normative judgment that it is good to do so. Suppose that he is totally sincere in making this judgment, and that all the evidence supports attributing it to him. Suppose also that this judgment is basic: i.e., that it is not based on false beliefs about the usefulness, for some further ends, of switching on these radios, but that it is a judgment that switching radios on is simply, in itself, good. This agent and his actions do not seem any more rational than the agent who merely has a bare desire to switch radios on. So the basic normative judgment does nothing to rationalize action in this case. Thus, following Quinn's reasoning, even when basic normative judgments are *correct* (as, for example, that getting pleasure is good), it is not the normative judgment that rationalizes action in accord with those judgments. To paraphrase Quinn:

> No normative judgment can, by itself, make the contribution to rationalizing action that adherents to the judgment thesis suppose it to have. This is true even if the judgment is correct: i.e., that pleasure or health are good. For pleasure or health provide a point to their pursuit that does not consist in the fact that they are *judged* to have a point.[28]

[27] One should not be misled by the language of this sentence into thinking that the current issue is whether one should regard beliefs or facts as the kind of thing that motivates us. That is emphatically not the issue. For even if one holds that it is the *belief* that jogging will promote my health that motivates me, it is still the *fact* about the goodness of health that makes the motivation rational. And I need make no corresponding normative judgment for it to do so.

[28] This is a modification of a passage from Quinn (1995), p. 195.

In line with this paraphrase, Quinn does in fact seem to deny that a bare normative judgment is itself sufficient to rationalize desire or action. That is, he considers cases in which one makes a false normative judgment – assessing some end as choice-worthy when it is not – and acts on the corresponding desire. And what he says of the resultant choice is that it is intelligible, but not rational.[29] But why should we postulate the existence of a normative judgment that mediates between the perception of an end, whether it be choice-worthy or not, and an action, if the difference between merely intelligible and actually rational action depends not on the judgment, but on the choice-worthiness of the goal? One answer to this question might be that the normative judgment is required to make the action intelligible, and that this is a necessary condition on its being rational. We will assess this suggestion when we discuss Joseph Raz, but since Raz also picks up a thread of argument started by Jonathan Dancy, let us turn to Dancy first.

Dancy

Dancy holds that "[w]e can normally explain an agent's doing what he did by specifying the reasons in the light of which he acted."[30] Understood in a certain way, there is nothing objectionable in such a claim. But Dancy further holds that "[i]t is required for this sort of explanation that those features be present to the agent's consciousness – indeed, that they somehow be conceived as favoring the action."[31] Similarly, Dancy writes that "the explanation of action, at least that of intentional action, can always be achieved by laying out the considerations in the light of which the agent saw the action as desirable, sensible, required."[32] It is these further claims about what is involved in being motivated by a reason that this chapter is arguing against.

Why does Dancy endorse the judgment thesis? It may initially be smuggled in with the technical phrase 'in the light of'. Dancy defines this notion early in *Practical Reality* as the relation between an agent and the reasons for

[29] Quinn (1995), p. 201. [30] Dancy (2000), p. 5.

[31] Dancy (2000), p. 129. It is only the second part of this claim that will be disputed.

[32] Dancy (2000), p. 136. There is a way of reading the "always" in this claim so that Dancy is specifying only a sufficient condition for motivational explanation. In light of other claims, however, it seems most consistent to read the claim as meaning 'in principle, it is always possible to explain intentional action by laying out the considerations in the light of which the agent saw the action as desirable.'

which the agent acted.[33] Since, for Dancy, reasons are not psychological states, this relation is between the agent and (typically) an external state of affairs, as that it is very likely to rain. The relation holds whenever the state of affairs motivated the agent. So to say that an agent acted in the light of *x* is merely to say that *x* motivated the agent to act. Defined in this way, the phrase 'in the light of' is unproblematic. Moreover, it serves a useful purpose, since acting in the light of some consideration is different from acting (partly) because one believes that some state of affairs obtains. For example, and as Dancy mentions, it turns out that people are much more likely to give to charity if they believe that others are doing the same. Yet it does not generally seem appropriate to say that the fact that other people are giving to charity is an agent's reason for donating, or that an agent is motivated by it. That other people are giving does have an influence, and this influence involves something psychological, but it is insufficient to ground the claim that the agent acted for the reason that other people were donating – that the agent acted *in the light of* that consideration.

What is required to ground a claim that an agent acted in the light of some consideration? Dancy seems to assume that part of what is required is the normative judgment, by the agent, that the consideration favors the action. At one of the points where Dancy makes the claim that the purpose of psychological explanation is to reveal the light in which the agent came to do what he did, he adds that this light *cannot but be regarded as an evaluative light*.[34] But there is no argument in support of this claim. There is, however, an explanation for why it should have misleadingly appeared to Dancy that psychological explanations of actions must take for granted that the agent saw the things that motivated her in a positive evaluative light.[35] For it is indeed bizarre for an agent to be motivated by some consideration, and yet not to regard that consideration in a positive evaluative light. Since it is bizarre, we do not feel that we have received a satisfying explanation when this happens. But the reason why it is bizarre is that we humans are typically motivated by considerations that are in

[33] Dancy (2000), p. 6.

[34] Dancy (2000), p. 97. It is unclear whether Dancy is speaking for Nagel or for himself at this point. But even if he is only representing Nagel with this claim, in many other places he says essentially the same thing. See, e.g., the passages above, taken from Dancy (2000), pp. 129, 136.

[35] I write 'take for granted' here so that these evaluative judgments can function as what Dancy calls 'enabling conditions,' rather than as a proper part of the explanation of action. The current argument will speak equally strongly against this sort of role for normative judgments.

fact good reasons, and, equally typically, we (adults) have correct beliefs about those considerations being good reasons. When we are motivated by something that we do not regard as a good reason, there are then two important possibilities as to what might be going on. The first is that we are right to regard the consideration as not being a reason. Then our desire, and the related action, may well be bizarre.[36] But this need not be because we lack a necessary normative judgment that would have made the action intelligible. It may be because we are motivated by something that is not a good reason, and that is in fact a strange source of motivation. The second possibility is that we are motivated by a good reason that we actively judge *not* to be a good reason. This is bizarre also. But in this case what is left unexplained is the fact that we regard a consideration as not being a good reason, when it is one. This is especially bizarre if we are actually motivated by the reason.

But surely, one might say, if an agent regards a certain end as of great value, then he has a reason to pursue it. Doesn't this show that normative judgments provide reasons for action?[37] No. Dancy himself provides the form of argument with which to deal with this sort of claim. His presentation of the argument comes in response to the parallel claim that having a desire must give one a reason, since we sometimes give advice to people, based on knowledge of their desires, even though we think they have no reason for those desires. Dancy's example involves a couple who have decided to insulate their house so well that no sound can penetrate from the outside.[38] Suppose one holds that there is no reason to have this desire, or to undertake this project. One might still advise these people to buy a certain brand of insulation, rather than another, given their end. This seems to show that a bare desire can ground a reason.

[36] 'May well be bizarre,' rather than 'will be bizarre,' because of common motives like revenge, envy, and hatred, which can make actions intelligible without seeming, even to the agent, to provide a normative reason in their favor.

[37] The view suggested here, which we might call 'the constitutive judgment thesis,' seems to add the following claim to the judgment thesis: the normative judgments that are required for rational action are not merely a necessary condition for such action, but somehow contribute to the reasons for which the agent acts. This stronger view should not be attributed to Dancy. It is presented here because it may appeal to other philosophers, and because Dancy himself provides the model for an argument against it. Moreover, even if normative judgments do *sometimes contribute* to an agent's reasons for action, this does nothing to support the claim that normative judgments are *always present* in every case of intentional or rational action.

[38] Dancy (2000), p. 34.

Dancy's response to the above argument is to disallow the relevant detachment in the following form of argument:

You ought, if *e* is your end, to pursue *e* in way *w*.
e is your end.
So you ought to pursue *e* in way *w*.[39]

The 'ought' in the first premise governs the whole of the conditional, and not merely the consequent. As a result, *modus ponens* certainly does not warrant the conclusion. And Dancy's claim is that no other inference rule does either. Rather, the point of the first premise is to assert that there is a certain kind of irrationality involved in having *e* as an end and failing to pursue it in way *w*: a kind of irrationality that goes beyond merely having *e* as an end in the first place. One does not have an extra reason, in virtue of one's desire for *e*. Rather, one merely has an additional way in which to be irrational: taking some other means to *e* than *w*.[40]

It should be clear how to modify Dancy's argument to deal with the above suggestion that normative judgments can give reasons. Suppose that someone regards switching radios on as a very good thing. If this is the case, then there is something irrational going on when this person tries to turn radios on by blowing on them. Because of this irrationality, one might be tempted to make the conditional claim that if a person regards switching radios on as good, then he has a reason to use his fingers to switch them on. In a particular case, in which someone actually does make the odd judgment that switching radios on is very good, one might be tempted to use this conditional claim to infer that the person has a reason to use his fingers to switch radios on. But this latter conclusion is as unwarranted as the claim that the couple in Dancy's example have a reason to buy a certain type of insulation. Regarding a nonreason as a reason does not convert it into a reason. Nor does it give one a reason to pursue the things 'favored' by the nonreason. Rather, there is a form of irrationality that consists in regarding something as a reason (rightly or wrongly), and then failing to act in a manner that is consistent with this judgment. This does not mean that one is acting rationally if one *does* act in a manner that is

[39] Dancy (2000), p. 43.

[40] For additional arguments against the same sort of detachment, see Greenspan (1975), pp. 271–74; Hare (1971), pp. 85–89; Darwall (1983), pp. 15–16 and 46–48; Broome (1999), pp. 409–11 and 415–17. Darwall points out that the detachment is at least more plausible when the relevant ends are intended, and not merely desired, but both Darwall and Broome hold that even understood in this way, the detachment is not valid.

consistent with one's wrong judgment. It only means one is avoiding a *further* irrationality.[41]

Raz

Joseph Raz tries to provide some argument for a crucial assumption that Dancy makes but does not defend: that a normative judgment is part of the 'in the light of' relation.[42] It will be easiest to discuss Raz if we first make more explicit one of the main views that he is attacking, and that this chapter is defending: the view that what makes a desire or choice intelligible (but not necessarily rational) is a matter of brute regularity. Because this will be the main view discussed in the present section, the focus will be on intelligible action, rather than on rational action. But Raz does not endorse the judgment thesis because he thinks that rational action, as a distinct subclass of intelligible action, requires the making of normative judgments. Rather, he endorses the judgment thesis because it holds that *all* intelligible action requires such judgments. Because of this, the arguments that follow bear directly on the judgment thesis.

In order to engage in argument against Raz's claim, the relevant notion of intelligibility needs to be identified. After all, even the most psychologically opaque behavior may (at some future point) be perfectly intelligible neurophysiologically. And even within a psychological theory of action, once we stipulate that the agent has certain completely bizarre basic desires, we may still regard the resulting action as intelligible, in a sense, if we can see that it comes about through the normal operations of the mechanisms that the theory of action postulates. What, then, is the sort of intelligibility Raz has in mind? Raz cannot simply stipulate that action is intelligible – in the relevant sense – only if we can see what the agent took to be of value in the action, or only if we can see what the agent took to be reasons in its favor. Characterizing intelligibility in this way would beg the question in Raz's favor, since we cannot be expected to see what the agent took to be of value in the action unless the agent did in fact take something in the action to be of value. And it is one purpose of this book to defend the claim that intelligible actions need not involve the agent's making any such judgment. I think the relevant sense of intelligibility is the following: an action is intelligible if there is a story about what motivated the action – a

[41] In this connection, see Lawrence (1995), pp. 137–38.
[42] See Raz (1999b), pp. 22–45. Raz even uses the phrase 'in the light of' (p. 24).

story that would answer the question 'What was the motivation behind that action?' in a way that would not leave us puzzled. This is extremely close to Raz's own characterization of 'typical intentional actions,' which he characterizes as:

> actions about which their agents have a story to tell (i.e., actions manifesting an internal viewpoint about what one is doing, or is about to do), a story which explains why one acted as one did . . . a story which shows what about the situation or the action made it, the action, an intelligible object of choice for the agent, given who he is and how he saw things at the time.[43]

This characterization does not mention normative judgments at all, and thus can be taken as common ground between Raz's view and my own.

In making the claim that brute regularity can make desire or choice intelligible, the word 'regularity' is not meant to suggest anything that could, even in principle, be formulated in terms of an exceptionless law, or indeed in terms of a law of any sort. Rather, it is only meant to describe events (or connections between events) that are sufficiently common that we are completely unsurprised when they happen. For example, it is a brute regularity that humans are motivated by the prospect of food, sex, rest, intellectual stimulation, the novel, and so on. It is because we are so familiar with these common kinds of motivations that we are unsurprised by them. And because we are unsurprised when people act from these sorts of motives, we are not puzzled by stories that explain their actions in terms of these sorts of motives. Raz opposes this view. He thinks that such brute regularities do nothing to make desire, or action, intelligible. Once this view is rejected, Raz needs to find something else that can make an action intelligible. What he offers is the fact that the agent sees certain considerations as providing reasons to perform it. In his own words, he holds:

> that the central type of human action is intentional action; that intentional action is action for a reason; and that reasons are facts in virtue of which those actions are good in some respect and to some degree.[44]

One might uncharitably read these claims as suggesting that intentional action is always action undertaken because of some fact in virtue of which the action *actually is* good in some respect. But this is too strong a reading.

[43] Raz (1999b), p. 24. That this is a characterization of *intentional* rather than *intelligible* action does not matter, for it is clear from the passage, and surrounding remarks, that Raz is not making any significant distinction between the two at that point in the book.

[44] Raz (1999b), p. 23.

Raz certainly allows that intentional action can be undertaken based on the mistaken judgment that something is a reason. Such actions are not counterexamples to his view, Raz writes, "for in the eyes of their agents they are good."[45] This implies that on his view intentional action only requires that the agent make an appropriate normative judgment, whether correct or incorrect. But such a view is too strong: intentional action does not require any normative judgments at all.

Raz himself supplies three examples of psychological intelligibility that is the result of nothing but contingent regularity, although, unsurprisingly, he does not offer the examples under that description. The first example is that it is intelligible that hunger makes concentration difficult.[46] One might object to the use of this example against Raz, claiming that he only intends it as an instance of an intelligible neuropsychological explanation. But in fact the example is chosen to elucidate the meaning of 'intelligible' in the claim that *morality* is intelligible, so it cannot be interpreted in this benign way. Now, what is it that makes it intelligible that hunger makes concentration difficult? As Raz points out, we are all sufficiently familiar with hunger and its effects when we are trying to concentrate to understand what someone means when they explain that the reason they could not concentrate was that they had not eaten in many hours. But this is simply appeal to a brute regularity, and to our familiarity with it. Had evolution taken another twist, it might have been that hunger actually sharpened one's abilities to concentrate. Had that been the case, then we would understand why someone would be pleased that they were hungry before an exam, rather than unhappy about it. One might try to deny that brute regularity is what is doing the work in this case. For example, one could claim that the intelligibility here depends on nothing more than the connection between concentration and distraction, and between distraction and unbidden sensations, i.e. ones whose occurrence is not, or is not entirely, under our volitional control. But why should the occurrence of unbidden sensations constitute a distraction, and not a spur to greater concentration? Again, the answer is simply that this is a brute regularity in human beings. One should not be fooled by the word

[45] Raz (1999b), p. 25. One might take this qualification to indicate that Raz holds a disjunctive view according to which all intentional actions are done either for reasons in the sense of 'facts in virtue of which the action is good,' or for reasons in the sense of 'considerations that are good in the eyes of the agent.' This does not seem to be Raz's view. But if it were, the argument of this chapter would go equally well against it, since an agent can act for a reason that meets neither of these considerations.

[46] Raz (1999b), p. 174.

'distraction' into thinking that there is a *conceptual* connection between unbidden sensations and a diminished ability to concentrate, even if a good definition of 'distraction' is 'difficulty in concentrating caused by unbidden sensations.' For the presence of a word in the language can reflect a contingent regularity in human nature: 'grief' and 'jealousy' are good examples of this. Admittedly, this entire discussion concerns the intelligibility of an inability to concentrate. This is quite different from the intelligibility of a desire or an action. But if brute regularity can make the effect of hunger psychologically intelligible, it is at least more plausible that the same sort of regularity can make desire and action intelligible also.

The second example of brute regularity contributing to psychological intelligibility is the case of contrariness: acting because reasons point in the opposite direction. Raz admits that some intentional action is the result of contrariness, and that such action is more or less intelligible.[47] How is it rendered intelligible? "[C]ontrariness is an established psychological phenomenon."[48] Although Raz does not acknowledge it, the appeal here boils down to brute regularity. Perhaps it is true, as Raz notes, that we cannot understand contrariness except as a degenerate case: a rebellion against the normal case of being motivated by reasons. But even if this is true, it also remains true that it is the brute regularity of contrariness that makes contrary action intelligible, to the extent that it is. What else could it be? It cannot *just* be the fact that contrariness is a relatively simple variation on the standard case. For we could imagine a whole host of other such variations. If those invented variations were not actually part of the standard human repertoire, then appeal to such a variation, even if it could be established to have taken place in some odd agent, would not render that agent's action intelligible in any way. We understand contrariness as we understand that hunger makes concentration difficult: through experience of the brute regularities in which it is manifested.

A final example of choice being rendered intelligible by brute motivational regularity involves the role of personality and character in choice. Raz has the view that the available reasons typically underdetermine

[47] Raz holds that such actions are not completely intelligible. For they are irrational, and he holds that irrational actions cannot be completely intelligible. But he also admits that irrational actions can be 'understandable' in a fairly robust sense: we know what they feel like, we can predict them, they may not even appear irrational from the agent's perspective. In the light of these admissions, his assertion that all irrational actions are unintelligible to some degree seems theoretically motivated. See Raz (1999b), p. 35.

[48] Raz (1999b), p. 33.

choice, leaving a wide field of options, all of which are rationally eligible.[49] This view plays a major role in his account of the nature of the will, of the explanation of moral supererogation, and in other places. The point here is not to dispute the underdetermination claim, which is extremely plausible. Rather, the question is the following: if the available reasons do not determine a choice between A and B, how is the choice of A over B to be rendered intelligible? It is not enough to say that there was an undefeated reason in favor of A. For there was an undefeated reason in favor of B also. The reason in favor of A does, admittedly, render the choice intelligible to a certain degree. Nor does Raz claim that the appeal to the reason in favor of A *does* render this choice intelligible, as a choice of A *over* B. But still, we do want an answer to this further question: Why did the agent chose A *and not B*? If choices are only rendered intelligible by reasons, then in cases of underdetermination by reasons there is no resource that could possibly render the choice intelligible. If, as Raz plausibly holds, most choices are in fact underdetermined, this would seem to imply that most human action is to some degree unintelligible. This conclusion is false, and Raz does not endorse it. When a choice is underdetermined by the available reasons, explanation is still possible. Such explanation will appeal to our personality and character, among other things. As Raz puts it:

> when we face conflicting adequate reasons for action, the explanation of why we followed the reasons we did will involve more than the invocation of our rationality. It will allude to our tastes, predilections, and much else besides.[50]

But Raz holds that our tastes and predilections (saving in some odd cases) do not provide reasons. How then do they help explain? How can my personality make my choice intelligible? Raz does not say how. The suggestion I would like to make is that personality and taste provide explanations for action because they are names for certain brute regularities, and because these sorts of brute regularities are explanatory. Imagine that two people are in sufficiently similar situations that the same reasons apply to their choice. Assume that these reasons underdetermine the choice, and that among the rationally eligible choices are a rather selfish one, and a

[49] Raz (1999b), pp. 46–66. Raz's explanation for this fact is that there is widespread incommensurability of reasons, which makes a determination of the balance of reasons impossible. The distinction between justifying and requiring strengths explains the same underdetermination.

[50] Raz (1999b), p. 117. See also Raz (1999b), p. 66.

rather generous one. Further, suppose that one of the agents is quite selfish while the other is quite generous. If the generous agent chooses the generous action, and someone asks why, we can provide a sufficient answer by citing the generous character of that agent. Moreover, even if we admit that the selfish action is rationally eligible, we will be puzzled if the generous agent chooses the selfish action, and we will press for further explanation. Yet there was presumably an adequate reason to act in this selfish way, since it was also rationally eligible. This shows that merely citing an adequate reason is *not* what makes the generous agent's generous action intelligible. What we want explained is the agent's choosing the selfish action *over* the generous one. This explanation cannot simply consist in citing the reason that favors the selfish action.

Moreover, once the explanatory roles of personality and taste are recognized, it becomes much less plausible that desires and actions that *are* determined by the available reasons need any implicit normative judgments to make them intelligible. Rather, we should say that, as a matter of brute fact, the overwhelming majority of human beings share certain motivational characteristics, such as an aversion to pain, death, and injury, and that human personalities also generally overlap in certain unsurprising ways. We do not need two different forms of explanation: one, for desires and actions that are determined by reasons, and another for those that are not. A defender of Raz might say that though we do not need these two forms of explanation, we nevertheless have them, and that a view such as Raz's helps explain the difference between them. That is, one might hold that Raz is simply correct that in some cases our explanation of action makes reference to reasons, and in other cases it makes reference to tastes and personalities. But the point here is not to deny this. Rather, the point is that while there are important differences between these two sorts of explanation, there is also one extremely important similarity: both get their explanatory force by appeal to brute regularity. Reasons-explanations are explanatory because of the brute fact that human beings are almost exclusively motivated by considerations that are, in fact, reasons – whether they are recognized as reasons or not. Personality-explanations are explanatory because of the brute fact that *this human being* is typically motivated in certain ways. In neither case is it necessary to postulate that the agent saw a reason *as a reason*. He may simply have noticed that it was beginning to rain, and opened his umbrella because he wanted to avoid getting wet.

Again, on both Raz's view and on my own, typical intentional actions are actions about which there is a story to tell. What commits Raz to

the judgment thesis is the further claim that this story "is of what the agents took to be facts which show the action to be good, and which therefore constitute a reason for its performance, making it eligible."[51] Raz's objector holds that there are some features of action that make them intelligible objects of choice, but that are not good-making features, and that do not even appear to be good-making to the agents who perform those actions. Thus, much of Raz's argument for the judgment thesis consists in showing that the search for such features is doomed to failure.

Part of the problem for Raz's objector is that Raz forces him to assume that the notion of a reason is only used in one way, either as an explanatory notion, or as a normative one. Call this view 'the univocality of reasons.' Thus, the objector makes the following claim:

> Once we draw a clear distinction between features which show an act to be an intelligible object of choice and ones which show it to be good or of value we will see . . . that reasons belong with the first, and not with the second.[52]

The univocality of reasons will certainly land the objector in difficulties if we can produce claims in which reasons are functioning in an unambiguously normative way. But why cannot some reasons be explanatory without being (even perceived as) normative, just as some reasons can be normative without being explanatory? Because the objector assumes the univocality of reasons, he is forced into an implausible view according to which explanatory reasons have features that are plausibly held only by normative reasons: universality, for example, and the ability to justify actions. This is why the objector wrongly concedes that if the fact that an action will hurt another person is a reason that makes it intelligible, then everyone would have such a reason.[53] It is true that at one point the objector tries to deny that reasons must *pro tanto* justify actions. But Raz's response is that restricting the role of reasons in this way fails to preserve the normativity of reasons, and this – surprisingly – is enough to silence the objector. But if *some* of the things we call reasons are not normative, this is no problem. The objector never raises this point. It is true that Raz is not making these assumptions without realizing it. He explains that it is "central to [his] approach that the same concept is crucial both for intelligibility and to justification (and therefore also to evaluation)."[54] This thought is expressed more clearly a few sentences later, when Raz

[51] Raz (1999b), p. 24. [52] Raz (1999b), p. 25.
[53] Raz (1999b), pp. 26, 28. [54] Raz (1999b), p. 31.

explicitly gives the following test for a consideration possibly making an action an intelligible object of choice:

> the same consideration must also show that the act is justified or at least that there is something to be said in its justification, something that in the absence of contrary considerations makes it justified.

But this is exactly the point of dispute, and Raz provides no defense for it. The view offered in this chapter is that we humans are imperfectly rational in the following sense: we are motivated by a wide range of considerations, most but not all of which are normative reasons. There is no need to hold that a consideration can only belong to the former class if it is regarded by the agent as belonging to the latter.

Scanlon

Of the four philosophers considered in this chapter, Scanlon is perhaps the most explicit in his endorsement of the judgment thesis. For example, he writes:

> [i]n order for a consideration to be an operative reason for me, I have to believe it. In addition, I have to take it to be a reason for the attitude in question.[55]

> The only source of motivation lies in my taking certain considerations – such as the pleasures of drinking, of eating, of hearing from a friend – as reasons.[56]

> [what are usually called desires] are not a matter of preconceptual appetite but involve at least a vague appeal to some evaluative category.[57]

In fact, Scanlon's judgment thesis concerns not only actions, desires, and intentions, but beliefs, anger, and so on. Scanlon characterizes these all as 'judgment-sensitive attitudes,' by which he means that an ideally rational person would have or lose these sorts of attitudes depending upon their judgments of the sufficiency of reasons.[58]

Unlike Raz, Scanlon is happy to appeal to regularity in order to make claims about what sorts of things explain human actions and other attitudes. One central way in which Scanlon argues for the judgment thesis is by

[55] Scanlon (1998), p. 56. These particular claims helpfully distinguish two senses in which judgment might be necessary for a reason to motivate an agent: judgment about a matter of fact, and judgment about a normative matter. Again, this chapter is only challenging the necessity of the latter.

[56] Scanlon (1998), p. 35.

[57] Scanlon (1998), p. 65.

[58] Scanlon (1998), p. 21. Actions are included because they are so closely linked to intentions, which are attitudes.

appeal to a regular connection between normative judgments and attitudes. That is, he repeatedly points out regularities of the following sort:

> [A] rational person who judges there to be compelling reason to do A normally forms the intention to do A.[59]

> Judgments of there being good reason will correlate with action in a rational person.[60]

In context, these claims are intended to do two things. The first is to form part of an argument against the idea that brute desires are the primary source of intentional action in humans. That is, Scanlon is pointing out that there is a regularity between something cognitive on the one hand, and action on the other. Scanlon might admit that desires figure in action-explanation somehow; but if they do, these desires are responses to something else. Thus, an action is best explained by reference to this something else, and not by reference to antecedent desires. The second thing these claims are intended to do is to suggest that the 'something else' is a normative judgment about the sufficiency of reasons. That is, these claims are part of an argument that intentional action is normally the result of the agent making certain normative judgments. Scanlon expresses this particular interpretation of the correlations by characterizing intentions as 'judgment-*sensitive* attitudes,' and not, for example, 'judgment-producing attitudes' or 'judgment-constituting attitudes.'[61]

Let us grant that desire is normally explained by something else, so that this something else ought to be regarded as the real source of intentional action. Why should we believe that this 'something else' is a normative judgment? Scanlon's correlations not only do not settle the issue, they do not support the judgment thesis at all. If the correlations Scanlon asserts really obtain, this is just as likely to be because *both* the action *and* the judgment are explained by a third something: the agent's having noticed the facts about the situation that provide the reasons that favor it. For example, the agent notices that someone has fallen and needs help getting up. That the agent notices this nonnormative fact (which happens to provide a reason) could explain both the subsequent normative judgment 'I have a reason to help this person get up,' and the motivation to do so. One reason to favor this explanation over Scanlon's is that normative judgments are in fact very rarely part of the phenomenology of desire, intention, or action.

[59] Scanlon (1998), pp. 33–34. See also Scanlon (1998), p. 66. [60] Scanlon (1998), p. 61.
[61] See also Scanlon (1998), pp. 23–24 for the assumption of this direction of explanation.

Why does Scanlon choose normative judgments as the 'something else,' instead of mere awareness of facts that are themselves reasons? One possible explanation might be termed 'the philosophical fallacy.' This fallacy occurs when a philosopher tries to become clear on the nature of some general phenomenon, but examines it primarily as it occurs in the context of philosophical discussion or reflection. Thus, when Scanlon examines his own desire for ice cream, he notices that there is something cognitive behind his desire: a belief about the taste and sensations of the ice cream. But because he is also doing philosophy, he *also* notes that this fact seems to provide a reason for eating the ice cream, and that he himself makes this very judgment.[62] The same thing occurs when he considers other cases. Thus, in all of these cases the one constant factor is a normative judgment that he has a reason to do the action. It may be because of this philosophically induced regularity that Scanlon then concludes that it is judgments of this sort that are doing the motivational work. To a philosopher this unifying thesis is more attractive than the claim that in one case it is the thought about the taste that motivates, while in another it is the thought that a person needs help, and in another it is some other thought. But since Scanlon recognizes a plurality of values and reasons in any case, this increased unity is really an illusion.

Scanlon may also have fallen into the philosophical fallacy when he introduced the notion of normative reasons by claiming that they are answers to 'why should' questions.[63] Suppose, for example, that an agent is in a situation in which she has the following reason to ϕ: by ϕ-ing, she will make her sister happy. If, in this context, the agent actually asks a 'why should' question, there will not only be this reason, but there will be the nonnormative belief 'ϕ-ing will make my sister happy,' and the further normative judgment 'making my sister happy provides me with a reason.' In this reflective context it will become more difficult to determine whether it is the reasons themselves, or beliefs about facts that happen to be reasons, or normative judgments, that are doing the relevant motivational work. But in typical cases no 'why should' question precedes action, decision, or desire. In these typical cases there are nevertheless still reasons *for* decisions, actions, and desires, and reasons *why* we in fact decide, act, and desire. If one examines reasons in a more typical and less philosophically reflective context, it is easier to avoid holding that it is *judgments* about reasons that get one to decide, act, or desire.

[62] Scanlon (1998), p. 35. [63] Scanlon (1998), p. 18.

It is true that Scanlon admits that "[j]udgment-sensitive attitudes can arise spontaneously, without judgment or reflection." But he also thinks that in such cases "the formation of these attitudes is generally constrained by general standing judgments about the adequacy of reasons."[64] This qualification, taken together with the bulk of the claims cited above, suggests that normative judgments, whether conscious or unconscious, *play a role in the formation* of almost all of our desires and intentions. Nevertheless, it may be safest to attribute a slightly more modest version of the judgment thesis to Scanlon – a version according to which desires and intentions merely *involve or encompass* normative judgments – for at one point he almost explicitly denies the claim that "all desires arise from prior evaluative judgments of some kind"; he says of this claim that it "seems clearly false."[65] And immediately after this apparent denial he makes the weaker claim that "having what is generally called a desire involves having a tendency to see something as a reason."[66] But one should not interpret this 'tendency' as merely reporting a contingent statistical fact. For soon after these remarks, Scanlon claims that "when a person *does* have a desire [in one very common sense of that word] and acts accordingly, what supplies the motive for this action is the agent's perception of some consideration as a reason."[67] And when he later explains what he was doing in the section from which these last few passages have been taken, he writes the following:

> [t]here is such a thing as a consideration *seeming* to be a reason for a certain course of action. . . . I argued in Section 8, in effect, that such "seemings" are the central element in what is usually called desire.[68]

It is perhaps because Scanlon gives such pride of place to normative judgments about reasons, rather than to reasons themselves, or to beliefs in nonnormative facts that also happen to be reasons, that he defines rationality as relative to the agent's own normative judgments. For Scanlon, it is impossible to act irrationally unless one is acting against one's own normative judgment.[69] Thus all irrational action conforms to the pattern

[64] Scanlon (1998), p. 24. [65] Scanlon (1998), pp. 38–9. [66] Scanlon (1998), p. 39.

[67] Scanlon (1998), pp. 40–41. The very common sense here is 'desire in the directed-attention sense.' According to Scanlon's account of it, this is a sense of 'desire' that fits well with a way the term 'desire' is frequently used, and that "captures the familiar idea that desires are unreflective elements in our practical thinking" (p. 39). So for Scanlon even actions based on these unreflective desires are motivated by the perception of some consideration *as a reason*.

[68] Scanlon (1998), p. 65. [69] Scanlon (1998), p. 25.

of weakness of will. But this view has bizarre consequences if Scanlon is understood as offering an account of rationality in the 'mental functioning' sense. Consider, as Scanlon does, a person who is completely indifferent to the prospect of even the most excruciating pain on future Tuesdays. As long as this person judges – based, perhaps, on some extremely strange theory of well-being – that such indifference is warranted, Scanlon has "no hesitation in saying he is not irrational, just seriously mistaken."[70] In a similar vein, Scanlon writes that:

> A person who believes, on general theoretical grounds, that her future interests should be sharply discounted [or ignored altogether], and who acts accordingly, may be making a mistake about the reasons she has, but this does not make her irrational, any more that it does a person who accepts a fallacious argument or makes some other mistake about the reasons she has.[71]

I too have no hesitation in saying that such an agent is seriously mistaken. But she is also irrational, in a very important and intuitive sense. Terminologically, it is very unhelpful to place this irrational agent in the same basket as someone who overlooks a mistake in a mathematical proof. In real life, no matter what explanations such a person offered for her attitude, we would regard her as irrational, and try to get her some psychological help.[72] We would not believe that she was simply in the grip of a false theory, and try to argue her into a better view.

Moreover, I do not think that the problem here can be dismissed as a *mere* difference in terminology. Scanlon has no term that captures a very important kind of failure in practical agency. He has his own "irrational," which necessarily involves acting against one's own normative judgments. He has "what we have most reason to do," which is analogous to my "objectively rational," and which has very little to do with mental functioning at all, since many of the relevant reasons may be unknown, and unknowable, by anyone, at the time of action. He has "mistaken," and "open to rational criticism," which indifferently include culpable and nonculpable mistakes about matters of fact, and mistakes about what counts as a reason. And he has "unreasonable," which is always relativized to some goal given by the context, which goal may not be shared by the unreasonable person,

[70] Scanlon (1998), p. 29. [71] Scanlon (1998), p. 31.

[72] Being indifferent to such pain is not the same as having sufficient reasons to suffer it willingly. For example, if the agent justifiably believed that his suffering pain on Tuesdays would have some tremendous benefit for himself or other people, then it might be rational to be willing to suffer the pain. But even in such a strange case, the agent is not indifferent to the pain.

so that there may be no failure in practical mental functioning involved in her unreasonable behavior.[73] What we want is a term that captures the following: sometimes the agent is aware of facts, or should be aware of them, and these facts provide reasons that make a certain action irrational – in a sense different from Scanlon's. If the agent performs the action anyway, he is not properly responsive to reasons. It is this notion, and not Scanlon's formal one, that is most relevant to real-life questions of moral responsibility, freedom of the will, disabilities of the will such as phobias, compulsions, and addictions, competence to give consent, and so on. It is this notion that includes, as subtypes ordered roughly by degree, 'silly,' 'dumb,' 'stupid,' 'wrong-headed,' 'crazy,' 'insane,' and perhaps the colloquial understanding of 'irrational.'[74] Scanlon has robbed himself of this important notion by making action not primarily a response to facts that may or may not be reasons, but to judgments about reasons.

RATIONAL ANIMALS

It may seem that the picture of rational action presented in this chapter implies that the actions of dogs and mice are also, at least generally, rational. After all, dogs and mice are disposed to avoid pain and death, to seek food and sex, and so on; there are reasons why they act as they do, and many of them seem to be adequate reasons for acting in those ways. That is, not only do they act in a goal-directed way, they respond to the available reasons in a way that is appropriate, given the normative significance of those reasons. At this point it may seem that the implicit normative judgments of the judgment thesis are part of what is needed to distinguish humans and other rational beings from arational animals.[75] For quite often this distinction is made by reference to something called 'the will,' which is represented as precisely the power to distance ourselves from our desires in a way that mice cannot, and to choose which desires to act upon, in the light of judgments about the values of the ends towards which those desires are pointed.[76] Whether or not this is the correct explanation, surely some relevant difference must be found between mice and human beings.

When considering the differences between mice and human beings, it is a theoretical advantage to be able to represent those differences as ones

[73] See Scanlon (1998), pp. 32–33 and 191–92.

[74] Though again, the question of how untutored people use the word 'irrational' is neither here nor there. See p. 143.

[75] See Scanlon (1998), p. 23.

[76] See Korsgaard (1996a), pp. 91, 113.

of degree. And where there is a difference that is not merely a difference in degree, it is a theoretical advantage to be able to represent such a difference as arising from a difference in degree. Human beings *are* animals. At some point in our evolutionary history our ancestors were more similar to dogs and mice than they were to us. It does not seem credible that the characteristics that we currently possess, and in virtue of which our actions can be assessed as rational and irrational, appeared at one go, as the result of a single fortuitous mutation. It is true that in possessing hearts and lungs humans differ from bacteria in more than mere degree. And it may be that humans possess some entirely new and distinct mental organs completely absent in dogs or mice. But we are much, much more similar to dogs and mice than we are to bacteria. Moreover, the weak claim being made here is only that *if* a satisfying account of some psychological difference between mice and human beings can be produced without appeal to an entirely new and distinct psychological organ, so much the better.

My suggestion is that what is of primary importance in explaining why human actions can be assessed as rational and irrational, while those of mice cannot, is that human beings typically see farther into the future, and represent more possibilities of action and their likely results.[77] Since the selection of one of these possible actions is therefore a far more complicated process than the process by which a mouse ends up pressing a bar for food, there are many more ways in which it can go wrong. One way is that an option that would certainly have been represented in a normal human being, and that should have been selected, given the relevant reasons, somehow failed to get represented. Another type of failure is that an option that should have been selected was not selected, even though it was represented. It is precisely these sorts of errors that are typically cited when we explain why we regard someone's action as irrational: 'You should have known that extra helping would make you sick,' 'You knew that yelling at her would just make matters worse.'[78] Because we can make these sorts

[77] See McDowell (1995), pp. 152–53. McDowell imagines a rational wolf, and makes many claims about such a being that resonate strongly with the current suggestion. But McDowell turns reason into an independent psychological faculty when he writes of this wolf that "[h]aving acquired reason, he can contemplate alternatives; he can step back from the natural impulse and direct critical scrutiny at it." This is where McDowell goes astray, assuming that the rational wolf needs something *extra* to explain its ability to choose among the various options presented by its increased imagination.

[78] The point here is only that these simple forms of irrationality provide a sufficient basis for making one important distinction between humans and other animals. That claim is consistent with the fact that acting against one's normative judgments is *another* way in which one can act irrationally.

of mistakes, and because we can be trained, by criticism, to minimize them, we have developed the concepts ('irrational,' 'crazy,' 'stupid') with which to criticize them. It is because these terms apply to some human actions, and not to others, that it is correct to say that human actions are appropriate objects of rational assessment. Because mice do not have such a sophisticated ability to represent (and, hence, to *mis*represent) future contingencies, their actions will never be correctly assessable as irrational. Therefore we do not think it appropriate to classify those actions as rational either. Their actions are outside the sphere of rational assessment, so we call them 'arational.'[79]

It is beyond the scope of this book to provide a complete defense of the above suggestion. Happily, however, it is not crucial to the current project that the suggestion turn out to be correct. What is crucial is only that *some* relevant difference between humans and arational animals be found that does not depend on the assumption that rational action requires agents to make normative judgments. On the above explanation of what makes human action apt for rational assessment, it remains possible to regard such action as the product of the interplay of desires, where desires do not involve any normative assessments.[80] In complex choice situations, perhaps one common desire is the desire to take a moment to reflect. But even this desire does not have any normative judgment at its core. Moreover, during the resulting reflection, what may often happen is simply that a number of options are imagined, and one is acted upon, without any evaluation of it as the best option. Those who are used to thinking of human agency as involving a conceptually separable will, an entity standing somehow above our desires and impulses, may protest that this description of human action is false to the phenomena. They may suggest that we are all aware of the entity they call 'the will,' directing action in the light of normative judgments based on the relevant reasons. But I must confess, I do not have even the faintest awareness of any such entity. My own internal reflection and, more importantly, my memory of actions done without reflection, suggests that the simpler view presented here is much more true to the phenomena. Typically, I just act. Sometimes I consider the

[79] It is true that, for example, brain-damaged mice behave in ways that can, without any abuse of language, be called 'crazy.' But clearly, if one wishes to call the actions of mice rational and irrational, one is not an adherent of the judgment thesis.

[80] Pure cognitivists, and those with views similar to Dancy's, can substitute 'capacity to be motivated by various considerations' for 'desire' here. Nor is anything in the basic picture opposed to the introduction of other psychological entities, such as intentions, decisions, and so on.

options beforehand. Sometimes I consider them carefully. But in the end, I act, and I am not aware of anything else going on.

Of course there is room in the picture for normative judgments. Not only can we make them, but they can be part of the explanation for our actions. There is no limit on the content of our desires that *excludes* explicitly normative ends. As a result of upbringings by decent parents in relatively stable societies, many of us have a desire to do the morally right thing, or at least to avoid doing the morally wrong one. This desire explains why moral judgments sometimes play a role in explaining our actions. And in many nonmoral cases, as in the choice of a career, we may well ask ourselves 'What should I do?' and act on the answer we arrive at. In many cases this question may merely be a prelude to reflection on the options. In such cases the reasons that explain our subsequent action may lie entirely in the ends of our desires. In other cases, however, the desire to *do what we judge we ought to do* may also contribute to the explanation of our action. When this desire plays the right sort of role, we might wish to call the resulting action 'autonomous,' or to identify it in some other way. But if we do this, we should keep in mind that autonomy, in this sense, is not as common as rationality, and is not, for example, a requisite for moral responsibility.[81]

CONCLUSION

The judgment thesis arises in the course of arguments against the Humean view that desire is the ground of, or a necessary condition for, an agent having a normative reason for action. Nevertheless, it is plausible that the judgment thesis is itself a result of the same theoretical pull that leads Humeans to make their characteristic claims. This is the pull to establish a noncontingent connection between desire and justification. Humeans, famously, yield to this pull by investing desire itself with normative significance. Dancy, Quinn, Raz, and Scanlon may be yielding to it when they claim that the normal case of desire involves the perception of a consideration as a reason. But in rejecting the Humean picture of normative reasons, advocates of the objective reasons thesis have the resources to separate reasons and desires to a more appropriate distance. Such a separation can help proponents of the objective reasons thesis to avoid the criticism

[81] For example, we are morally responsible for many negligent acts that do not involve any desire to do what we judge we ought to do.

that they are committed to 'queer' normative properties, the perception of which is sufficient to generate desire, or the more general problem of explaining how beliefs can motivate independent of antecedent desire.[82] These problematic views are avoided if we simply acknowledge that things such as pain, death, knowledge, pleasure, power, and so on, are extremely common motives for human action, and that these things are *also* normative reasons. These last two claims are of course related. Human language, and the concepts to which it gives rise, have been formed in the light of human nature. So it is no surprise that we have a term such as 'harm,' that collects the unvarying core of human aversions. And it is no surprise that such a term functions in a normative way, to guide and assess action.

Despite the failure of Dancy, Quinn, Raz, or Scanlon to provide any convincing argument in support of the judgment thesis, those who are committed to it may try to claim that normative judgments are somehow implicit in our actions, or are presupposed by them. But in order to argue for such a claim they will have to provide criteria for attributing such normative judgments: criteria that are distinct from the mere disposition to be motivated by certain ends. For without such independent criteria, normative judgments will collapse into desires, as they did in the case of normative appearances. Nor will it be sufficient merely to provide these sort of independent criteria. Proponents of the judgment thesis will also have to establish that these judgments typically play an explanatory role in the genesis of desire or action, and are not themselves to be explained by the same facts that explain desire or action.

[82] See Mackie (1977), ch. 1.

References

Alston, W. 1991. *Perceiving God*, Ithaca, Cornell University Press.

Anscombe, G.E.M. 1995. "Practical Inference" in Hursthouse *et al.* 1995, pp. 1–34.

Arthur, J. and Shaw, W. (eds.) 1991. *Justice and Economic Distribution*, New Jersey, Prentice Hall.

Audi, R. 1985. "Rationalization and Rationality" *Synthese* 65, 159–84.

1997. *Moral Knowledge and Ethical Character*, New York, Oxford University Press.

Baier, K. 1954. "The Point of View of Morality" *Australasian Journal of Philosophy* 32, 104–35.

1965. *The Moral Point of View*, New York, Random House.

1978. "Moral Reasons and Reasons to be Moral" in Goldman and Kim 1978, pp. 231–56.

Balguy, J. 1978. *The Foundation of Moral Goodness* (1728), facsimile edition, New York, Garland Press.

Binkley, R., Bronaugh, R., and Marras, A. (eds.) 1971. *Agent, Action, and Reason*, Toronto, Toronto University Press.

Blackburn, S. 1995. "The Flight to Reality" in Hursthouse *et al.* 1995, pp. 35–56.

Brandt, R. 1979. *A Theory of the Good and the Right*, Oxford, Oxford University Press.

Brink, D. 1986. "Externalist Moral Realism" *Southern Journal of Philosophy* 24, 23–42.

Broome, J. 1999. "Normative Requirements" *Ratio* 12, 398–419.

Cahn, S. and Haber, J. (eds.) 1995. *Twentieth Century Ethical Theory*, New Jersey, Prentice Hall.

Chang, R. (ed.) 1997 *Incommensurability, Incomparability, and Practical Reason*, Cambridge, Harvard University Press.

Clarke, R. 1994. "Doing What One Wants Less: A Reappraisal of the Law of Desire" *Pacific Philosophical Quarterly* 75, 1–10.

Clarke, S. 1978. *Works of Samuel Clarke*, vol. 2 (1738), facsimile edition, New York, Garland Press.

Cohen, G.A. 1996. "Reason, Humanity, and the Moral Law" in Korsgaard, 1996a, pp. 167–88.

Cohon, R. 1986. "Are External Reasons Impossible?" *Ethics* 96, 545–56.

Copp, D. 1995. *Morality, Normativity, and Society*, New York, Oxford University Press.

1997. "Belief, Reason, and Motivation: Michael Smith's *The Moral Problem*" *Ethics* 108, 33–54.
Cullity, G. and Gaut, B. (eds.) 1997. *Ethics and Practical Reason*, New York, Clarendon Press.
Dancy, J. 2000. *Practical Reality*, New York, Oxford University Press.
Darwall, S. 1983. *Impartial Reason*, Ithaca, Cornell University Press.
1990. "Autonomist Internalism and the Justification of Morals" *Noûs* 24, 257–68.
1992. "Internalism and Agency" in J. Tomberlin (ed.), *Philosophical Perspectives 6, Ethics*, Atascadero, Ridgeview, pp. 155–74.
1994. "From Morality to Virtue and Back" *Philosophy and Phenomenological Research* 54, 695–701.
1999. "Ethical Intuitionism and the Motivation Problem" presented at *Jornadas Internacionales de Etica y Derecho*, Universidad Torcuato Di Tella, Buenos Aires, June 23, 1999.
Davidson, D. 1963. "Actions, Reasons, and Causes" *Journal of Philosophy* 60, 685–99.
1980. "Freedom to Act" in his *Essays on Actions and Events*, Oxford, Clarendon Press, pp. 63–81.
Davies, M. and Humberstone, L. 1980. "Two Notions of Necessity" *Philosophical Studies* 38, 1–30.
Deigh, J. 1996. "Reason and Motivation" in J. Deigh, *The Sources of Moral Agency*, Cambridge, Cambridge University Press, pp. 133–59.
Doris, J. 1998. "Persons, Situations, and Virtue Ethics" *Noûs* 32, 504–30.
Dreier, J. 1990. "Internalism and Speaker Relativism" *Ethics* 101, 6–25.
Duggan, T. and Gert, B. 1967. "Voluntary Abilities" *American Philosophical Quarterly* 4, 127–35.
1979. "Free Will as the Ability to Will" *Noûs* 13, 197–217.
Edgley, R. 1965. "Practical Reasoning" *Mind* 74, 174–91.
Foley, R. 1991. "Rationality, Belief and Commitment" *Synthese* 89, 365–92.
1992, "The Epistemology of Belief and Degrees of Belief" *American Philosophical Quarterly* 29, 111–24.
Foot, P. 1978a. "Morality as a System of Hypothetical Imperatives" in Foot 1978c, pp. 157–73.
1978b. "Reasons for Action and Desires" in Foot 1978c, pp. 148–56.
1978c. *Virtues and Vices*, Oxford, Basil Blackwell.
Gert, B. 1998. *Morality: Its Nature and Justification*, New York, Oxford University Press.
Gert, J. 2002a. "Expressivism and Language Learning" *Ethics* 112, 292–312.
2002b. "Avoiding the Conditional Fallacy" *Philosophical Quarterly* 52, 88–95.
Gibbard, A. 1990. *Wise Choices, Apt Feelings*, Cambridge, Harvard University Press.
Goldman, A. and Kim, J. (eds.) 1978. *Values and Morals*, Boston, Reidel Publishing Co.
Good, I.J. 1952. "Rational Decisions" *Journal of the Royal Statistical Society*, Ser. B. 14, 107–14.

Greenspan, P. 1975. "Conditional Oughts and Hypothetical Imperatives" *Journal of Philosophy* 72, 259–76.
Hardin, C.L. 1993. *Color for Philosophers*, expanded edition, Indianapolis, Hackett Publishing Co.
Hare, R.M. 1971. "Wanting: Some Pitfalls" in Binkley *et al.* 1971, pp. 81–97.
Harman, G. 1982. "Review of *A Theory of the Good and the Right* by Richard Brandt" *Philosophical Studies* 42, 119–39.
Heath, J. 1997. "Foundationalism and Practical Reason" *Mind* 106, 451–73.
Herman, B. 1993. *The Practice of Moral Judgment*, Cambridge, Harvard University Press.
Hobbes, T. 1994. *Leviathan* (1651), E. Curley (ed.), Indianapolis, Hackett.
Hume, D. 1978. *A Treatise of Human Nature*, L.A. Selby-Bigge and P.H. Nidditch (eds.), Oxford, Clarendon Press.
Hursthouse, R., Lawrence, G., and Quinn, W. (eds.) 1995. *Virtues and Reasons: Philippa Foot and Moral Theory*, New York, Oxford University Press.
Isen, A.M. and Levin, H. 1972. "The Effect of Feeling Good on Helping: Cookies and Kindness" *Journal of Personality and Social Psychology* 21, 384–88.
Johnson, R. 1999. "Internal Reasons and the Conditional Fallacy" *Philosophical Quarterly* 49, 53–71.
Kagan, S. 1989. *The Limits of Morality*, New York, Oxford University Press.
Kane, R. 1996. *The Significance of Free Will*, New York, Oxford University Press.
Kant, I. 1988. *Fundamental Principles of the Metaphysics of Morals* (1785), T.K. Abbott (trans.), Amherst NY, Prometheus Books.
Korsgaard, C. 1996a. *The Sources of Normativity*, Onora O'Neill (ed.), Cambridge, Cambridge University Press.
1996b. "Skepticism about Practical Reason" in her *Creating the Kingdom of Ends*, Cambridge, Cambridge University Press, pp. 311–34.
1997. "The Normativity of Instrumental Reason" in Cullity and Gaut 1997, pp. 215–54.
Lawrence, G. 1995. "The Rationality of Morality" in Hursthouse *et al.* 1995, pp. 89–147.
Lewis, D. 1989. "Dispositional Theories of Value" *Aristotelian Society*, Suppl. 63, 113–37.
Mackie, J.L. 1977. *Ethics: Inventing Right and Wrong*, Harmondsworth, Penguin.
McDowell, J. 1995. "Two Sorts of Naturalism" in Hursthouse *et al.* 1995, pp. 149–79.
Mele, A. 1989. "Motivational Internalism: The Powers and Limits of Practical Reasoning" *Philosophia* 19, 417–36.
1998. "Motivational Strength" *Noûs* 32, 23–36.
2000. "Goal-Directed Action: Teleological Explanations, Causal Theories, and Deviance" *Philosophical Perspectives* 14, 279–300.
2003. *Motivation and Agency*, New York, Oxford University Press.
Mill, J.S. 1979. *Utilitarianism* (1861), George Sher (ed.), Indianapolis, Hackett.
Millikan, R.G. 2000. *On Clear and Confused Ideas*, New York, Cambridge University Press.

Morris, C.W. and Ripstein, A. (eds.) 2001. *Practical Rationality and Preference*, New York, Cambridge University Press.
Nagel, T. 1970. *The Possibility of Altruism*, Princeton, Princeton University Press.
Nozick, R. 1993. *The Nature of Rationality*, Princeton, Princeton University Press.
Pattit, D. 1986. *Reasons and Persons*, Oxford, Oxford University Press.
1997. "Reasons and Motivation" *Aristotelian Society*, Suppl. 71, 99–131.
2001. "Bombs and Coconuts, or Rational Irrationality" in Morris and Ripstein 2001, pp. 81–97.
Pettit, P. 1991. "Realism and Response-Dependence" *Mind* 100, 587–626.
Pettit, P. and Smith, M. 1993. "Practical Unreason" *Mind* 102, 53–79.
Philips, M. 1987. "Weighing Moral Reasons" *Mind* 96, 367–75.
Quinn, W. 1995. "Putting Rationality in its Place" in Hursthouse *et al.* 1995, pp. 181–208.
Railton, P. 1986. "Facts and Values" *Philosophical Topics* 14, 5–31.
Rawls, J. 1971. *A Theory of Justice*, Cambridge, Harvard University Press.
Raz, J. 1975. "Permissions and Supererogation" *American Philosophical Quarterly* 12, 161–68.
1985–6. "Value Incommensurability: Some Preliminaries" *Proceedings of the Aristotelian Society* 86, 117–34.
1999a. *Practical Reason and Norms*, New York, Oxford University Press.
1999b. *Engaging Reason*, New York, Oxford University Press.
Rescher, N. 1987. "Rationality and Moral Obligation" *Synthese* 72, 29–43.
1994. "Replies to Commentators" *Philosophy and Phenomenological Research* 54, 441–57.
Ripstein, A. 2001. "Preference" in Morris and Ripstein 2001, pp. 37–55.
Rosati, C. 1995. "Persons, Perspectives, and Full Information Accounts of the Good" *Ethics* 105, 296–325.
Ross, D. 1939. *The Foundations of Ethics*, New York, Oxford University Press.
Scanlon, T. 1998. *What We Owe to Each Other*, Cambridge, Harvard University Press.
Scheffler, S. 1995. *The Rejection of Consequentialism*, Oxford, Clarendon Press.
Sen, A. and Williams, B. (eds.) 1982. *Utilitarianism and Beyond*, Cambridge, Cambridge University Press.
Sidgwick, H. 1981. *The Methods of Ethics* (1907), Indianapolis, Hackett.
Singer, M. 1996. "Reason, Humanity, and the Moral Law" in Korsgaard 1996, pp. 167–88.
Singer, P. 1972. "Famine, Affluence, and Morality" *Philosophy and Public Affairs* 1, 229–42.
Skorupski, J. 1999. "Irrealist Cognitivism" *Ratio* 12, 436–59.
Slote, M. 1984. "Morality and Self-Other Asymmetry" *Journal of Philosophy* 81, 179–92.
1992. *From Morality to Virtue*, New York, Oxford University Press.
Smart, J. 1991. "Distributive Justice and Utilitarianism" in Arthur and Shaw (eds.) 1991, pp. 106–17.
Smith, M. 1994. *The Moral Problem*, Cambridge, Blackwell.
1995. "Internal Reasons" *Philosophy and Phenomenological Research* 55, 109–31.

1996. "Normative Reasons and Full Rationality: Reply to Swanton" *Analysis* 56, 160–68.

2002. "Bernard Gert's Complex Hybrid Conception of Rationality" in R. Audi and W. Sinnott-Armstrong (eds.), *Rationality, Rules, and Ideals: Critical Essays on Bernard Gert's Moral Theory*, Boston, Rowman and Littlefield, pp. 109–23.

Sobel, D. 2001. "Subjective Accounts of Reasons for Action" *Ethics* 111, 461–92.

Stampe, D. 1987. "The Authority of Desire" *Philosophical Review* 96, 335–81.

Stocker, M. 1994. "Self-Other Asymmetries and Virtue Theory" *Philosophy and Phenomenological Research* 54, 689–94.

Strang, C. 1995. "What if Everyone Did That?" in Cahn and Haber 1995, pp. 378–85.

Svavarsdóttir, S. 1999. "Moral Cognitivism and Motivation" *Philosophical Review* 108, 161–219.

Thalberg, I. 1985. "Questions about Motivational Strength" in E. LePore and B. McLaughlin (eds.), *Actions and Events*, Oxford, Basil Blackwell, pp. 88–103.

Tilley, J. 1997. "Motivation and Practical Reasons" *Erkenntnis* 47, 105–27.

Velleman, D. 1996. "The Possibility of Practical Reason" *Ethics* 106, 694–726.

Weirich, P. 2001. "Risk's Place in Decisions Rules" *Synthese* 126, 427–41.

Wiggins, D. 1998. *Needs, Values, Truth*, 3rd edn., New York, Oxford University Press.

Williams, B. 1981. "Internal and External Reasons" his in *Moral Luck*, Cambridge, Cambridge University Press, pp. 101–13.

Wittgenstein, L. 1953. *Philosophical Investigations*, New York, Macmillan.

Index