Embodied Minds in Action

ROBERT HANNA
AND MICHELLE MAIESE

Embodied Minds in Action

This page intentionally left blank

Embodied Minds in Action

Robert Hanna and Michelle Maiese

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2009

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Hanna, Robert, 1957–
Embodied minds in action / Robert Hanna and Michelle Maiese.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 978–0–19–923031–0
1. Philosophy of mind. 2. Cognitive neuroscience. 3. Mind and
body. 4. Causation. 5. Act (Philosophy) I. Maiese, Michelle. II.
Title.
BD418.3.H35 2009
128—dc22
2008036356
Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire
ISBN 978–0–19–923031–0
10 9 8 7 6 5 4 3 2 1

To MTH and ETH
There is no love, there are only proofs of love (R.H.)

To AAM and MLM
The heart has reasons of its own **(M.M.)**

This page intentionally left blank

Preface and Acknowledgments

This book began life as another book and also as a Ph.D. dissertation. In the early 2000s, Robert Hanna and Evan Thompson started a book together on the mind–body problem and mental causation. Shortly after that, Michelle Maiese began her Ph.D. dissertation project under Hanna's direction at the University of Colorado at Boulder, on mental causation, the emotions, and intentional action. This three-way collaboration proved to be highly fruitful. Maiese successfully completed her Ph.D. in 2005. Hanna and Maiese then wrote *Embodied Minds in Action*, and in the meantime Thompson wrote another book on his own, which he had begun in the mid-90s with the late Francisco Varela—*Mind in Life: Biology, Phenomenology, and the Sciences of the Mind*. The two books are independent projects, but complementary. They jointly offer a new and unified approach to consciousness, intentionality, the mind–body relation, mental causation, and intentional action. Our particular focus in the present book is a unified treatment of three fundamental philosophical problems arising from these intimately-related topics: What accounts for the existence and specific character of conscious, intentional minds like ours in a physical world? What accounts for the causal relevance and causal efficacy of conscious, intentional minds like ours in a physical world? And what accounts for the categorical difference between the things we consciously and intentionally do, and the things that just happen to us?

Our unified treatment of these fundamental problems rests on two basic claims. The first is that conscious, intentional minds like ours are *essentially embodied*. This entails that our minds are irreducible to our brains, *not* because they are in any way immaterial properties or facts, but instead because they are necessarily and wholly spatially spread throughout our living, organismic, material bodies and belong to their complete neurobiological constitution. The second claim is that essentially embodied minds are *self-organizing thermodynamic systems*. This entails that our mental lives consist in the possibility and actuality of moving our own living organismic bodies through an egocentrically-centered, orientable (i.e., intrinsically directional) space and in thermodynamically irreversible time,

by means of our conscious desires. Otherwise put, our two core ideas, which we call *the Essential Embodiment Theory*, are these:

(1) Conscious, intentional minds like ours are the irreducible and truly global—or inherently dominating—intrinsic structures of motile, neurobiologically complex, situated, forward-flowing living organisms.
(2) Nature basically includes complex dynamic organismic life, and essentially embodied minds like ours are alive. So because organismic life is basically causally efficacious, then essentially embodied minds like ours are basically causally efficacious too.

We are extremely grateful, of course, to Evan Thompson—without whom we would not have started this project. We are also extremely grateful to the following people for their very helpful comments on and criticisms of earlier drafts and presentations of various parts of this book, or for discussions of its central topics over the last eight years: Luc Bovens; Heather Demarest; Ton Derksen and the 03–04 Mind & Cognition Research Group at the University of Nijmegen, Netherlands; Lark Fleming; Robert Harrison; Jane Heal and the 03–04 Philosophy of Mind Discussion Group at the University of Cambridge, UK; Sara Heinämaa; Daniel Korman; Marc Moffett, Edward Sherline, and the other participants in a colloquium talk at the University of Wyoming, Laramie, USA in October 07; Graham Oddie; Robert Pasnau; Josh Rasmussen; David Robb; Teed Rockwell; Jean-Michel Roy and the 03–04 Seminar in the Epistemology of the Cognitive Sciences at the École Normale Supérieure in Lyons, France; the 05–06 Philosophy of Mind Group at the University of Colorado, Boulder (Brandon Bogardus, briefly Nic Damnjanovic, David Ivy, Kristin Demetriou, and Brian Robinson); the 07–08 Cognition, Content, & Consciousness Group at CU-Boulder (Leonard Boonin, Kristin Demetriou, Walt Gorsuch, Ann Howry, Ronald Le Bel, Robert Rupert, and Andrew Winters); the late Robert Solomon; Michael Zerella; and three anonymous referees for Oxford University Press.

Robert Hanna would also like to give the warmest of all possible thanks to the Center for Consciousness Studies at the University of Arizona, Tucson, for research support in 00–02 when initial drafts of parts of some of the chapters were written up; to Fitzwilliam College, University of Cambridge, for a visiting research fellowship in Michaelmas term 06 (and especially to Michael Potter, who organized it), which made it possible

for the penultimate version of the book to emerge from the genteel and intellectually stimulating patterns of Cambridge college life; and finally to the members of his Spring 08, 4300/5300 Philosophy of Mind class at CU-Boulder for their philosophical enthusiasm and good-natured tolerance as the final typescript was being prepared.

All of these people and institutions have greatly helped us in our work, and in many different ways. But none of them is to be held responsible for the views we develop and defend in this book. If the Essential Embodiment Theory is correct, then it subverts the traditionally opposed and seemingly exhaustive categories of Dualism and Materialism, and presents a new paradigm for contemporary mainstream research in the philosophy of mind and cognitive neuroscience. And that in turn will also have revisionary implications for action theory, and the metaphysics of free will and moral responsibility. If we are correct, then the natural world basically includes conscious, intentional, deeply free, rational, and morally responsible complex dynamic living organisms, who substantively enrich and extend nature through their spontaneous mental activities and their intentional body movements, without being in any way reducible to the merely non-living, mechanical, deterministic, or stochastic parts of nature. In other words, we are essentially *minded animals* who help to create the natural world through our own agency. That is a truly radical idea.

The italicized phrase under the dedication to MTH and ETH is RH's translation of a sentence in the brilliant 1945 film, *Les Dames du Bois de Boulogne*, directed by Robert Bresson, with dialogue by Jean Cocteau, which was based on the anonymous 1793 novella, *Exemple Singulier de la Vengeance d'une Femme*, and was translated back into French from Friedrich Schiller's 1785 German translation of a part of Denis Diderot's extended, rambling *Tristram Shandy*-ish novel, first published in 1796, *Jacques le Fataliste et son Maître*. So dedications can have complex dynamic histories too.

This page intentionally left blank

Contents

Introduction

> There is nothing that my own nature teaches me more vividly than that I have a body. . . . Nature also teaches me that I am not merely present in my body as a sailor is present in a ship, but that I am very closely joined and, as it were, intermingled with it, so that I and the body form a unit.
>
> René Descartes[1]

The purpose of this book is to provide the rudiments of a unified treatment of three fundamental philosophical problems: the mind–body problem, the problem of mental causation, and the problem of intentional action. As we are construing it, the mind–body problem is this: What explains the existence and specific character of conscious, intentional minds like ours in a physical world? Correspondingly, the problem of mental causation is this: What explains the causal relevance and causal efficacy of conscious, intentional minds like ours in a physical world? And finally the problem of intentional action is this: What explains the categorical difference between the things we consciously and intentionally do, and the things that just happen to us?

Whether there are minds significantly *unlike* ours—ghostly minds, non-spatiotemporal minds, infinite minds, omniscient minds, omnipotent minds, etc.—and if so, what their nature is, are questions we will not seriously consider here. Our fundamental interest lies in trying to understand precisely what it is to be *a creature with a conscious, intentional* ***mind like ours***, and then in working out the most salient implications of this understanding for a unified treatment of the three fundamental problems. Our resolution to concentrate

[1] Descartes, *Meditations on First Philosophy* 56, AT 80–81.

almost exclusively on creatures with conscious, intentional "minds like ours"—or, for terminological convenience, "$minds_{lo}$"—is not an arbitrary or trivial one. We believe that mainstream contemporary philosophy of mind and cognitive neuroscience (in a broad sense that includes cognitive psychology, medical neurology, neurophysiology, and neurobiology) have been significantly distorted and misled by the modern-classical idea, deriving from Descartes, that disembodied counterparts of our own minds (a.k.a. "spirits") and mindless counterparts of our own living bodies (a.k.a. "zombies") are logically possible. Spirits and zombies are, indeed, *logically* possible, and the concepts of them can play a certain specialized role in certain recherché lines of reasoning in the philosophy of mind and other parts of metaphysics. But it does not follow that concepts about spirits and zombies are in any way explanatorily or metaphysically relevant to the nature of *our* minds or *our* living bodies. Indeed, we hold that spirits and zombies alike are actually *strongly metaphysically a priori impossible* (see especially sections 1.3 and 7.4). This is a type of *essential* impossibility, which flows directly from the kind of *creature* or *being* that we are.[2] So it is the metaphysics of *our kind of* minds and *our kind of* living bodies that we are fundamentally interested in, not the metaphysics of some *essentially different kind* of minds and living bodies, of which we have only the thinnest conceptual and logical grasp. This is a book about the philosophy of *minded animals*, not a book about the philosophy of spirits or zombies.

Our unified treatment of the three fundamental problems, which we call the *Essential Embodiment Theory*, rests on two basic claims. First, $minds_{lo}$ are *essentially embodied*. This means that our minds are necessarily and wholly spatially spread through our entire living bodies and all their vital systems, vital organs, and vital processes—including the higher brain, brain stem, limbic system, nervous system, endocrine system, immune system, and cardiovascular system—right out to the skin. On our view, $minds_{lo}$ necessarily *include* our brains but also are necessarily *not restricted* to our brains. This entails that $minds_{lo}$ are irreducible to our brains, *not* because they are in any way immaterial properties or facts, but instead because they are necessarily and wholly spatially spread throughout our living organismic bodies and belong to their complete neurobiological constitution.

[2] See Fine, "Essence and Modality."

Second, essentially embodied minds are *self-organizing thermodynamic systems*. Self-organizing thermodynamic systems are unified collections of material elements in rule-governed or patterned motion, involving heat and other forms of energy, that also have *dissipative structure* and *natural purposiveness*. A dissipative structure is how the natural energy loss or entropy in a thermodynamic system is absorbed and dispersed (hence "dissipated") by the systematic re-introduction of energy and matter into the system, via a non-static causal balance between the inner states of the system and its surrounding natural environment. And natural purposiveness is how a thermodynamic system with dissipative structure self-generates forms or patterns of order that determine its own causal powers, and in turn places constraints on the later collective behaviors, effects, and outputs of the whole system, in order to maintain itself. The prime example of a self-organizing thermodynamic system is a living organism.

Now, according to the notion of essential embodiment, necessarily all creatures with minds_{lo} are living organisms. If correct, then when combined with the thesis that essentially embodied minds_{lo} are self-organizing thermodynamic systems, this entails that *by virtue* of our having essentially embodied mental lives, we are *also* inherently capable of making intentional body movements. So as we put it in the Preface, the two core ideas of the Essential Embodiment Theory are these:

(1) Conscious, intentional minds_{lo} are the irreducible and truly global or inherently dominating intrinsic structures of motile, neurobiologically complex, situated, forward-flowing, living organisms.
(2) Nature basically includes complex dynamic organismic life, and essentially embodied minds_{lo} are alive. So because organismic life is causally efficacious, then essentially embodied minds_{lo} are basically causally efficacious too.

When we talk about "the mind–body problem," what do we mean by the notions of "mind" and "body"? By the notion of "mind," as we have said, we mean specifically a mind_{lo}. In turn, for a creature to have a mind_{lo} is for that creature to have both $\textit{consciousness}_{lo}$ and $\textit{intentionality}_{lo}$. To say that a creature has $\text{consciousness}_{lo}$ is to say that a creature is either currently enjoying, or has a capacity for, *subjective experience*. And to say that a creature has $\text{intentionality}_{lo}$ is to say

(i) that a conscious creature is either currently enjoying, or has a capacity for, directing itself *at* or *towards* objects, actions, locations, events, other conscious creatures or itself (i.e., its intentional *targets*),

and

(ii) that a conscious creature has mental states that are *about* something or another, by virtue of the *mental content* of those states—which can include
  (i) the targets of those states *themselves*,
  (ii) *how* the conscious creature *refers* to those targets (e.g., anticipations, demonstrations, intuitions, ostensions, direct perceptions, etc.),

or

  (iii) *how* the conscious creature *describes* those targets (e.g., concepts, senses, ''modes-of-presentation,'' propositions, etc.).

For example, both consciousness$_{lo}$ and intentionality$_{lo}$ are manifest in caring of all sorts, salient drives of all sorts, inclinations of all sorts, liking and disliking of all sorts, love and hate, lust and disgust, moods of all sorts, passions of all sorts, pleasures and pains of all sorts, feelings of all sorts, and sensations of all sorts. They are also manifest in what we call *primitive bodily awareness*—proprioception (the sense of the relative positioning of one's own body parts and limbs, at rest or in movement), orientation and balance (the proprioceptive spatial senses of bodily location and locating), kinaesthesia and motility (the proprioceptive temporal senses of bodily movement and movability), bodily pleasures and pains, tickles and itches, the feeling of pressure, the feeling of temperature, the feelings of vitality or lethargy, and so on. Again, they are manifest in the external perceptual modes of touch, smell, taste, hearing, and vision. And finally they are manifest in thinking and reasoning of all sorts too. All the conscious, intentional creatures we actually know about, and seemingly could *ever* know about, are motile neurobiologically complex living organisms. So by the notion of ''body'' in ''the mind–body problem'' we mean the motile, neurobiologically complex, egocentrically-centered and spatially oriented, thermodynamically irreversible living organismic physical body of any creature having a conscious, intentional mind$_{lo}$.

The two core ideas of the Essential Embodiment Theory—(1) that conscious, intentional $minds_{lo}$ are the irreducible and truly global or inherently dominating intrinsic structures of motile, neurobiologically complex, situated, forward-flowing living organisms, and (2) that because organismic life is basically included in nature and is basically causally efficacious, then our minds are basically causally efficacious too—when adequately elaborated, enable us to offer a radically revisionary explanation of mental causation and intentional action. This is, as we have just said, a *radically revisionary* explanation in relation to contemporary mainstream philosophy of mind and cognitive science, but also one that is not *wholly* historically unprecedented. As we will see in a moment, it is significantly related to Aristotle's metaphysics. But perhaps even more significantly, although the Essential Embodiment Theory is radically opposed to Cartesian Dualism and mechanism alike, it is also ironically true that Descartes's own passing remarks about the "intermingling" of mind and body into a single "unit" strongly anticipate our core idea. There is, indeed, nothing that our own nature, and nature itself, teach us more directly and vividly than that we have living organismic bodies and that we are "very closely joined" to them.

We believe that this is no mere metaphysical accident, and that the embodiment of our minds *necessarily* extends to *all* the vital systems, vital organs, and vital processes of our living bodies. This is not to say that we are always or even usually conscious *of* our living bodies and their vital systems, organs, or processes. Indeed, this is relatively rare, as, e.g., when I become single-mindedly and vividly attentive to the pounding of my heart and the heaving of my lungs after running up a flight of stairs. But it is indeed to say that $minds_{lo}$ are always and necessarily conscious and intentional *with*, or *in-and-through*, all the vital systems, organs, and processes of our living bodies. If we are correct that $minds_{lo}$ are always and necessarily conscious and intentional with or in-and-through our living bodies and their basic neurobiology, then it follows that $minds_{lo}$ are necessarily and completely incarnated, situated, forward-flowing, alive, and causally efficaciously engaged with the natural world.

Precisely how does the Essential Embodiment Theory relate to the mind–body problem and the problem of mental causation? As we said at the beginning, the mind–body problem, as we are understanding it, is how to give an adequate account of the existence and specific character

of conscious, intentional minds$_{lo}$ in a physical world; and the problem of mental causation is how to give an adequate account of the causal relevance and causal efficacy of conscious, intentional minds$_{lo}$ in a physical world. So formulated, the mind–body problem and the problem of mental causation date back to the seventeenth century and the emergence of modern natural science, and more specifically to the sixth of Descartes's *Meditations on First Philosophy*. Here he claims that mind and physical matter—and correspondingly, individual minds and individual physical bodies—are two essentially distinct kinds of substance, and that they causally interact despite being only contingently related to one another. This Cartesian doctrine, familiarly known as *Dualism*, also quickly gave rise to some now all-too-familiar questions: How is it possible for a non-extended, immaterial substance to cause physical bodily movements without undermining the mechanistic (whether deterministic or probabilistic) laws of physics? And on the other hand, if all the motions of our own physical bodies can be completely and mechanistically explained by physics, then what causal work is left for the mind?

Strictly speaking, Dualism comes in two distinct flavors: *Substance Dualism* and *Property Dualism*. This distinction, in turn, is usually interpreted as[3] the distinction between the *Interactionist Substance Dualism* described by Descartes in the sixth of the *Meditations*, and *Property-Dualism-Without-Substance-Dualism*.[4] Interactionist Substance Dualism says:

(i) that mind and body are essentially distinct existing kinds of stuff or things, and as a consequence
(ii) that mental properties are not necessarily coextensive with physical properties, and it is possible for both disembodied minds (i.e., spirits) and mindless counterparts of our living bodies (i.e., zombies) to exist,

and also

[3] There are, however, some other possible ways of interpreting it. For example, it is possible to defend a *non-Cartesian* Substance Dualism of *person* and body. See, e.g., Lowe, *An Introduction to the Philosophy of Mind*, 15–21. To keep things relatively simple, in the main text we will not explicitly consider non-Cartesian Substance Dualism. But it seems clear, in any case, that just like Cartesian Interactionist Substance Dualism, it would not be able to provide an adequate solution to the problem of mental causation—see chapter 6 below.

[4] See, e.g., Bealer, "Mental Properties"; and Jackson, "Epiphenomenal Qualia." Strictly speaking, Property-Dualism-Without-Substance-Dualism is also consistent with non-reductive materialism. See note 7 below.

(iii) that minds and bodies nevertheless causally interact with one another.

Property-Dualism-Without-Substance-Dualism, by contrast, while it shares only the *second* of these three theses with Cartesian Interactionist Substance Dualism, *also* says:

(ii*) that only physical stuff and physical things actually do exist, but some of those physical things have accidental or extrinsic mental properties,

and

(ii**) that these accidental or extrinsic mental properties are either epiphenomenal (i.e., caused by physical properties but without any causal powers of their own) or else they have autonomous causal powers with some sort of "downward" causal impact on the physical properties of things.

These days, neither Interactionist Substance Dualism nor Property-Dualism-Without-Substance-Dualism has many supporters, and most contemporary philosophers of mind opt for some form of *Materialism*.[5] These include:

(1) *Eliminative Materialism*, which outright denies the existence of everything mental, including minds, mental states, mental events, mental processes, and mental properties, and asserts that there are nothing but brains and other purely physical things in a purely physical world,
(2) *Reductive Materialism* or *Physicalism*, which identifies mental properties with certain physical properties, and also identifies mental states, events, or processes with certain physical states, events, or processes,[6]

[5] For a good survey of the various types of materialism, see Chalmers, "Consciousness and its Place in Nature."

[6] See, e.g. Kim, *Philosophy of Mind*; Kim, *Mind in a Physical World*; Kim, *Physicalism, or Something Near Enough;* and Kim, *Supervenience and Mind*. It is arguable that the necessitation involved in the materialist supervenience relation must be *logical* necessity in order for both the upwards dependence of the mental on the physical and its co-variation with the physical to be knowable a priori and satisfy the demands of explanatory reduction—see Chalmers, *The Conscious Mind*, ch. 2, and Braddon-Mitchell and Jackson, *Philosophy of Mind and Cognition: An Introduction*, ch. 1. This is very plausible. So for the purposes of our discussion it is conceptually economical to think of Reductive Materialism as including both classical Physicalism, or the mind-brain identity thesis, and also Reductive Functionalism, or

and

(3) *Non-Reductive Materialism*,[7] which rejects the identity of mental properties and physical properties, and thereby accepts the independent existence of mental properties even if it accepts the identity of particular mental states, events, or processes with particular physical states, events, or processes, but also claims:

(a) that as a matter of fact there are not any disembodied minds,

and

(b) that all mental properties (naturally or nomologically) strongly supervene on fundamental physical properties.[8]

A prime example of Eliminative Materialism is the claim that commonsense or "folk" psychology is nothing but a pseudo-science that must be fully replaced by the cognitive neurosciences (i.e., cognitive psychology, medical neurology, neurophysiology, and neurobiology).[9] A prime example of Reductive Materialism is the claim that the mind is identical with the brain, and that phenomenology—i.e., consciousness$_{lo}$ and intentionality$_{lo}$,

the identity of mental properties with certain second-order physical (i.e., functional) properties, and as specifically requiring the logical strong global supervenience of mental properties on fundamental physical properties. For the definition of logical strong global supervenience, see Section 1.1 below.

[7] Chalmers carefully distinguishes his own view, *Naturalistic Dualism*, which says that mental properties are nomologically but not logically supervenient on physical properties, from Materialism. But just as it is conceptually economical for us to think of Reductive Materialism as including both the mind-brain identity theory and Reductive Functionalism by way of the logical strong supervenience thesis, so too it is correspondingly economical to make the class of non-reductive materialist theories large enough to include all views based on nomological or natural supervenience. To be sure, what counts as "materialism" is somewhat stipulative. And correspondingly there has been a fair bit of controversy and wrangling about the fairly subtle differences between Non-Reductive Materialism on the one hand, and Property-Dualism-Without-Substance-Dualism on the other. But Kim usefully defines "minimal physicalism" as committed to (i) mind–body strong supervenience (physical indiscernibility entails mental indiscernibility), (ii) the anti-Cartesian principle (no disembodied minds allowed), and (iii) mind–body dependence (a thing's mental properties are necessarily determined by its physical properties). See Kim, *Philosophy of Mind*, ch. 1. As we are using the notion, then, Non-Reductive Materialism = minimal physicalism + the non-identity of mental and physical properties. Now mind–body dependence entails that the causal powers of something's mental properties are inherited from the causal powers of its physical properties. But if a non-reductive materialist *also* asserts that something's mental properties can have an autonomous "downward" causal impact on its physical properties, e.g., by saying that "mental properties can make a causal difference," then Non-Reductive Materialism begins to merge very confusingly with Property-Dualism-Without-Substance-Dualism.

[8] Facts are instantiated properties. The phrases 'fundamental physical property', 'fundamental mental property', '*X* is fundamentally physical' and '*X* is fundamentally mental' are technical terms that we define in Section 1.1 below.

[9] See, e.g., Churchland, "Eliminative Materialism and the Propositional Attitudes"; Churchland, *Matter and Consciousness*; and Churchland, *Neurophilosophy*.

as experienced and described by the subject herself—is nothing but a set of concepts for describing neural processes.[10] Another prime example of Reductive Materialism is the claim that the conscious, intentional mind is nothing but a computer program, or some other kind of causal-functional organization, implemented on different kinds of brain hardware. This is Reductive Functionalism. But if one were also to hold that even though the *intentional* mind is nothing but a computer program or some other kind of causal-functional organization, nevertheless the *conscious* mind involves some sort of non-physical properties, precisely because the human organism also has some quite interesting and (so far) irreducible raw phenomenal feels, then that would be a prime example of Non-Reductive Materialism.[11] So it is entirely possible to be at once a Reductive Functionalist (with respect to intentionality$_{lo}$) and a Non-Reductive Materialist (with respect to consciousness$_{lo}$).

But setting aside for a moment the subtle metaphysical details and intensely contentious differences between the various forms of Dualism and Materialism, we can find a single bottom line. Dualism seems clearly false. Both the causal interaction of essentially different mental and physical substances, as well as the "downward" causal impact of the accidental mental properties of things on their intrinsic physical properties, are metaphysically mysterious. And on the other hand, the epiphenomenality or causal inertness of mental properties seems equally mysterious: how and why would something with causal powers ever produce something *without* any causal powers? It would appear then that the only other option is some form of Materialism. Materialists all argue that a conscious, intentional mind is neither a mysteriously empowered spiritual substance nor a mysteriously disempowered shadow of a material substance—a mere "ghost in the machine," as Gilbert Ryle famously describes it[12]—but rather can be explained in fundamentally physical and mechanistic terms. So materialists are all telling us that a conscious, intentional mind$_{lo}$ is really nothing but *another machine* in the machine, whether that other machine

[10] See, e.g., Place, "Is Consciousness a Brain Process?"; and Smart, "Sensations and Brain Processes."

[11] See, e.g., Kim, *Physicalism, Or Something Near Enough*. In light of what we said in note 7 above about the malleability of the notion of Non-Reductive Materialism, it is quite possible that Kim would not accept the non-reductive materialist label. But the crucial point is that he accepts the basic claims of Reductive Functionalism about intentionality, the Multiple Realizability Argument for non-identity, and also the nomologically or naturally strongly supervenient existence of raw phenomenal feels.

[12] Ryle, *The Concept of Mind*, ch. 1.

is something as concrete as a brain or something as abstract as a digital computer program.

Yet supposing that Materialism is true, it seems to violate our most firmly-held commonsense beliefs and feelings about intentional action. Again, the problem of intentional action is how to give an adequate account of the categorical difference between the things we intentionally do—e.g., raising my arm in order to wave to a friend—and the things that just happen to us—e.g., the uncontrollable rising of Peter Sellers's arm into a Nazi salute, in Stanley Kubrick's 1964 satirical sci-fi masterpiece, *Dr Strangelove: Or How I Learned to Stop Worrying and Love the Bomb*. Now intentional acts result from our choosing something, which in turn is the result of our consciously desiring or wanting something. But if our consciously desiring or wanting something did not *efficaciously, freely,*[13] *irreducibly, and literally* cause our intentional body movements, then it seems that our deepest ordinary working assumptions and attitudes about intentional action would turn out to be false. For in that case, I never really choose or do anything *myself*, and no choice or body movement I make is ever really *up to me*. Thus it is no advance over Dualism if instead of being ghosts-in-machines, we are then nothing but second-order machines. Am I a ghost-in-the-machine, or a machine-in-the-machine? It is obviously a Hobson's Choice, and *I want to return my ticket*: I seem to be doomed to theoretical and practical self-alienation and self-stultification no matter what.

But if we are to be *neither* dualists *nor* materialists, then we must account for mental causation and intentional action in some distinctively different way. Otherwise put, we hold that the mind–body problem, the problem of mental causation, and the problem of intentional action are all essentially

[13] 'Freely' should be taken here in the perfectly ordinary sense that includes both negative freedom (a person's ability to choose or act without preventative hindrance, and without internal or external compulsion), positive freedom (a person's ability to choose and act as she wants), and causal or moral responsibility. As the contemporary debate about free will and responsibility shows, this characterization is neutral as between the various competing metaphysical conceptions of freedom, since the metaphysical problem of free will is just this: How can persons choose or act with negative freedom, positive freedom, and moral responsibility in a deterministic or indeterministic world? See, e.g., Kane, *A Contemporary Introduction to Free Will*. So it seems that everyone, even a hard determinist or a hard incompatibilist, agrees that *if there were* free will, then it would at the very least include negative freedom, positive freedom, and moral responsibility. In this book, to keep things somewhat manageable, we avoid any direct discussion of free will. But it should be clear enough that nearly everything we say will have some sort of *bearing* on the free will problem. The Essential Embodiment Theory supports the doctrine of a *deep* or efficacious freedom of the will that is still fully embedded in nature. But that's a long story for another day.

the *same* problem, and require a unified treatment. In turn, our unified approach to these problems *reverse-engineers* a theory of the mind–body relation and mental causation by designing it to fit our basic intuitions about intentional action, including both our prima facie or pre-theoretical intuitions, and also the refined intuitions we develop by critical analysis of other theories of action.

This strategy has led us to an equally *non-dualist* and *non-materialist* view of the mind–body relation—the Essential Embodiment Theory. This view entails that creatures minded like us are *neither* ghosts-in-machines *nor* machines-in-machines. On the contrary, creatures with $minds_{lo}$ are essentially embodied minds and self-organizing thermodynamic systems. And because nature basically includes causally efficacious living organisms, then $minds_{lo}$ are causally efficacious in the same basic way. Or, in other words, creatures with $minds_{lo}$ are motile, neurobiologically complex, situated, forward-flowing living organisms that are truly globally intrinsically structured by irreducible $consciousness_{lo}$ and $intentionality_{lo}$, and are thereby inherently capable of performing intentional body movements under favorable endogenous and exogenous conditions. Or, in still other words, and not so longwindedly: *Because $minds_{lo}$ are alive it necessarily follows that, with a little bit of luck, creatures with $minds_{lo}$ can intentionally move their own living organismic bodies when they want to*. If the Essential Embodiment Theory is correct, then it is definitely something worth writing home about.

The Essential Embodiment Theory also has some broader implications. Essential embodiment and its self-organizing thermodynamics jointly entail a metaphysically *liberal* or tolerant conception of physical nature. On our view, conscious, intentional $minds_{lo}$ exist only *in* nature, and nature is everywhere and everywhen physical, but nature is not everywhere and everywhen *mechanistically* or *narrowly* physical. Believing the contrary entails an *illiberal* or intolerant conception of physical nature: that is, a *reductive* conception. But for us, some parts of nature at some times—the parts that are identical with the inner and outer mental lives of motile, suitably neurobiologically complex, situated, thermodynamically irreversible living organisms—are *essentially mental-and-physical*.

So in sharp contrast to both Dualism and Materialism alike, the Essential Embodiment Theory entails what we call *Mind–Body Animalism*, according to which the fundamental mental properties of conscious, intentional $minds_{lo}$ are at once

(a) non-logically or strongly metaphysically a priori necessarily reciprocally intrinsically related to corresponding fundamental physical properties in a living animal's body (the thesis of *mental-physical property fusion*),[14]

and also

(b) irreducible truly global or inherently dominating intrinsic structures of motile, egocentrically-centered and spatially oriented, thermodynamically irreversible living organisms of a suitable degree of neurobiological complexity (the thesis of *neo-Aristotelian hylomorphism*).[15]

What do these two very esoteric-sounding theses mean? We will explain them in detail in chapters 7 and 8. However, for the time being, they can be explicated quite simply in the following way. It seems clear that

(1) Dualism,
(2) Materialism,
(3) Idealism,

and

(4) the Dual Aspect Theory

exhaust all the basic logically possible metaphysical options for relating mind and body. *Dualism* asserts the mutual independence of mind and body. It says that mind and body are essentially separate but accidentally combined. *Materialism* asserts the asymmetric (i.e., one-way) necessary dependence of mind on body. It says that body strictly determines mind. And *Idealism* asserts the asymmetric necessary dependence of body on mind. It says that mind strictly determines body. Dualism, Materialism, and Idealism make up the *classical* menu of options in the history of modern philosophy

[14] For us, two properties are *fused* if and only if they are (a) non-logically or strongly metaphysically a priori necessarily co-extensive, (b) non-identical, and also (c) reciprocally intrinsic properties of any substance in which they are instantiated. We borrow the very useful term "property fusion" from Paul Humphreys. See Humphreys, "Aspects of Emergence"; Humphreys, "How Properties Emerge"; and Humphreys, "Emergence, Not Supervenience." Humphreys's notion of property fusion is importantly similar to ours, but also importantly different. See Sections 7.1 and 8.2 below for details.

[15] See Section 8.1 below. There are some interesting similarities between Mind–Body Animalism, and animalism in the debate about personal identity, although they are not strictly equivalent. See, e.g., Olson, *The Human Animal*, esp. 126.

and science running directly from the seventeenth century forward to the twenty-first century.

But by sharp contrast to all the options on the the classical menu, *the Dual Aspect Theory* asserts the mutual *inter*dependence of mind and body. Now Mind–Body Animalism is a special version of the Dual Aspect Theory. Mind–Body Animalism says that mind intrinsically requires body, that body intrinsically requires mind, and that they jointly constitute the minded animal. More precisely, we hold that the mutual interdependence of body and mind is essentially the same as the mutual interdependence of the material composition of a motile, suitably neurobiologically complex, situated, forward-flowing, organismic body and its *biological life*. This is because we think that conscious, intentional minds_{lo} *necessarily are alive*, by virtue of their necessary and complete neurobiological embodiment. Indeed, we hold that minds_{lo} and biological life are *continuous* in the robust two-part sense that

(i) conscious, intentional minds_{lo} non-logically or strongly metaphysically a priori necessarily entail biological life,

and

(ii) everything that is non-logically or strongly metaphysically a priori required for minds_{lo} is already present in biological life, although it is not always causal-dynamically organized or structured in the right way for minds_{lo},

hence

(iii) biological life non-logically or strongly metaphysically a priori necessarily entails the *strong metaphysical possibility* of minds_{lo}, although many living things are not themselves actually minded,

and most importantly of all

(iv) insofar as organismic life basically belongs to nature and is causally efficacious, then minds_{lo} are causally efficacious in the same basic way, allowing also for the significantly greater dynamic complexity of minded animals in comparison to other sorts of living organisms.

In this way, Mind–Body Animalism much more closely resembles Aristotle's pre-modern, biologically-oriented "hylomorphic"—matter/form

or stuffing/structure—metaphysics of the mind–body relation than it resembles modern theories of mind in the wake of Descartes's Interactionist Substance Dualism and his mechanism. Our approach is also *neo*-Aristotelian however, precisely because it fuses Aristotle's pre-modern, biologically-oriented hylomorphic metaphysics with the notion of essential embodiment, with modal dualism (the thesis that there are two essentially different types of necessity and necessary truth), and with contemporary dynamic systems theory.

Otherwise put, we are saying that some physical things—namely, motile, suitably neurobiologically complex, situated, thermodynamically irreversible living organisms—have irreducibly mental properties as a non-logically or strongly metaphysically a priori necessary consequence of their physical nature, *and furthermore* that this physical nature is intrinsically connected with those irreducible mental properties, *which in turn* are, again by non-logical or strong metaphysical a priori necessity, essentially embodied and causal-dynamically empowered. These motile suitably neurobiologically complex living organisms are just the creatures with $minds_{lo}$. So in motile, suitably neurobiologically complex, situated, forward-flowing, living organisms with $minds_{lo}$, the physical and the mental play, as it were, a metaphysical and causal-dynamic game of *loop-the-loop*: on the one hand, our fundamental physical properties non-logically or strongly metaphysically a priori necessarily, spatiotemporally, and causally loop through our neurobiology into corresponding fundamental mental properties; and then on the other hand, our fundamental mental properties also non-logically or strongly metaphysically a priori necessarily, spatiotemporally, and causally loop back through our neurobiology into corresponding fundamental physical properties. It is precisely this special metaphysical, inherently spatiotemporal, and inherently causal-dynamic "loop-the-loop" relation that is captured by the conjoined theses of mental-physical property fusion and neo-Aristotelian hylomorphism.

As we already noted, it is truly ironic that Descartes himself anticipated the core idea of the Essential Embodiment Theory. Indeed it seems that Descartes was ultimately pulled in two contradictory directions: on the one hand, towards Interactionist Substance Dualism and mechanism, and on the other, towards a dual-aspect, neo-Aristotelian metaphysics of the

mind–body relation like the Essential Embodiment Theory.[16] But even if Descartes himself was philosophically conflicted, we are *more than happy* to have him—or at least to have him in one of his philosophical guises—on board with us. So our view does not in any way exist in a philosophical void. Indeed, in addition to its Cartesian anticipation, the Essential Embodiment Theory has also been significantly influenced by several pre-modern, modern, recent, and contemporary thinkers and doctrines, including of course Aristotle, but also Kant, existential phenomenology (especially later Husserl, early Heidegger, early Sartre, and Merleau-Ponty), Whitehead's "philosophy of organism," later Wittgenstein, dynamic systems theory, embodied cognition theory, enactive cognition theory, the strong continuity of mind and life thesis developed by Evan Thompson, and the volitional theories of action developed by Harry Frankfurt and Brian O'Shaughnessy.

In any case, the two core ideas of the Essential Embodiment Theory, reformulated now as a general philosophical claim or proposition, are that:

> Intentional agency is possible if and only if (i) creatures with irreducible conscious, intentional $minds_{lo}$ are essentially embodied minds and self-organizing thermodynamic systems, and (ii) $minds_{lo}$ are basically causally efficacious because they are alive and organismic life is basically causally efficacious.

More precisely stated, however, the Essential Embodiment Theory has six central theses:

(1) **The Essential Embodiment Thesis:** Creatures with conscious, intentional $minds_{lo}$ are necessarily and completely neurobiologically embodied.

(2) **The Essentially Embodied Agency Thesis:** Basic acts (e.g., raising one's arm) are intentional body movements caused by an essentially embodied mind's synchronous *trying* to make those very movements and its active *guidance* of them.

(3) **The Emotive Causation Thesis:** Trying and its active guidance, as the cause of basic intentional actions, is primarily a pre-reflective, desire-based emotive mental activity and only derivatively a self-conscious or self-reflective, deliberative intellectual mental activity.

[16] See, e.g., Brown, *Descartes and the Passionate Mind*.

(4) **The Mind-Body Animalism Thesis:** The fundamental mental properties of conscious, intentional $minds_{lo}$ are (a) non-logically or strongly metaphysically a priori necessarily reciprocally intrinsically connected to corresponding fundamental physical properties in a living animal's body (mental-physical property fusion), and (b) irreducible truly global or inherently dominating intrinsic structures of motile, suitably neurobiologically complex, egocentrically-centered and spatially oriented, thermodynamically irreversible living organisms (neo-Aristotelian hylomorphism).

(5) **The Dynamic Emergence Thesis:** The natural world itself is neither fundamentally physical nor fundamentally mental, but is instead essentially a causal-dynamic totality of forces, processes, and patterned movements and changes in real space and real time, all of which exemplify fundamental physical properties (e.g., molecular, atomic, and quantum properties). Some, but not all, of those physical events *also* exemplify irreducible biological properties (e.g., being a living organism), and some but not all of those biological events *also* exemplify irreducible fundamental mental properties (e.g., $consciousness_{lo}$ or $intentionality_{lo}$). And both biological properties and fundamental mental properties are *dynamically emergent* properties of those events.

(6) **The Intentional Causation Thesis:** A mental cause is an event or process involving both $consciousness_{lo}$ and $intentionality_{lo}$, such that it is a necessary proper part of a nomologically jointly sufficient essentially mental-and-physical cause of intentional body movements. In so being, it is a dynamically emergent structuring cause of those movements. Then, under the appropriate endogenous and exogenous conditions, by virtue of synchronous trying and its active guidance, conscious intentional essentially embodied $minds_{lo}$ are mental causes of basic acts from their inception in neurobiological processes to their completion in overt intentional body movements.

The first two chapters make a direct case for thesis (1). The third and fourth chapters make a direct case for thesis (2), and also an indirect case for thesis (1). The fifth chapter offers an argument for thesis (3). And the sixth, seventh, and eighth chapters develop a series of metaphysical arguments for theses (4), (5), and (6). Taken together, these six theses collectively say *that*

in this causal-dynamic natural world, under the right endogenous and exogenous conditions, creatures with irreducible conscious, intentional minds$_{lo}$ can intentionally move their own neurobiologically complex, situated, thermodynamically irreversible living organismic bodies, by means of the synchronous desire-based emotions that constitute trying and its active guidance. Or as we said above, and much more simply put: Because minds$_{lo}$ are alive it necessarily follows that, with a little bit of luck, creatures with minds$_{lo}$ can intentionally move their own living organismic bodies when they want to.

Stepping outside the hothouse atmosphere of contemporary philosophy of mind and action for a moment, it is perhaps difficult to believe that anyone could *fail* to accept the Essential Embodiment Theory. It seems overwhelmingly obvious that like us, you are a rational being, capable of moral responsibility and free agency—a person. And it *also* seems equally overwhelmingly obvious that you are a conscious, intentional, neurobiologically complex, situated, forward-flowing living organism. Within a certain fairly limited range of natural and accidental variations, you are *corporeally outfitted and shaped* just like the rest of us, you *live and move* just like the rest of us, and you will also eventually *die* just the like the rest of us. So what Shakespeare's Shylock very poignantly said of the plight of Jews is undeniably true of us and you too:

> Hath not a Jew eyes? Hath not a Jew hands, organs, dimensions, senses, affections, passions; fed with the same food, hurt with the same weapons, subject to the same diseases, heal'd by the same means, warm'd and cool'd by the same winter and summer, as a Christian is? If you prick us, do we not bleed? If you tickle us, do we not laugh? If you poison us, do we not die?[17]

Now in order to read, understand, and then feel the emotional impact of these lines, you self-consciously, self-reflectively, and deliberatively (i.e., as a result of deliberation involving reasons) performed the intentional body movement of scanning the page with your eyes. But you also non-self-consciously, pre-reflectively, and non-deliberatively changed your body posture slightly, scratched your forehead, tapped your fingers, or wiggled your toes—*just because you desired to do it.* Nothing was preventing or forcing you. It was *up to you*, and *you alone.* Nothing else did it, and nobody else did it. That very bodily movement *would not have happened*

[17] Shakespeare, *The Merchant of Venice*, act III, scene I, ll. 58–66.

if you had not desired to do it and if you had not intentionally done it. So you yourself did those seemingly trivial things intentionally and freely, although neither self-consciously nor self-reflectively nor deliberatively. And if, counterfactually, you were to have suddenly discovered that you were paralyzed, and therefore that you had tried to move your body but failed, it would then have been obvious that in the *actual* case you did all these things by *trying* to do them. This set of pre-theoretical, manifest facts about essentially embodied intentional action is the intuitive starting point for all the philosophical analyses and arguments to follow in this book.

1

Consciousness$_{lo}$ and Essential Embodiment I: The Basics

> The soul is really joined to the whole body, and . . . we cannot properly say that it exists in one part of the body to the exclusion of the others. For the body is a unity which is in a sense indivisible because of the arrangement of its organs, these being so related to one another that the removal of any one of them renders the whole body defective. And the soul . . . is related solely to the whole assemblage of the body's organs.
>
> René Descartes[1]

> One must be conscious in order to choose, and one must choose in order to be conscious. Choice and consciousness are one and the same thing.
>
> Jean-Paul Sartre[2]

1.0 Introduction

Let us begin at the beginning—with consciousness$_{lo}$. According to our view, every creature with a consciousness$_{lo}$ is an *essentially embodied mind*. In turn, an essentially embodied mind is a *minded animal,* or more precisely, a conscious, intentional, motile, suitably neurobiologically complex, egocentrically-centered and spatially oriented, thermodynamically irreversible living organism. I have a consciousness$_{lo}$. You have a consciousness$_{lo}$. Your newborn baby, your cat, your dog, your horse, the mouse in your yard, the squirrel in the tree in the park, and the rational human animals living in the house next door all have consciousness$_{lo}$. You have conscious

[1] Descartes, *Passions of the Soul*, 339, AT 351. [2] Sartre, *Being and Nothingness*, 595.

feelings. They have conscious feelings. You spontaneously move your own living body because of your feelings, and by means of those feelings, and you thereby express those feelings. And they do so too, because of their feelings, and by means of their feelings, and in some way or another that is expressive of those feelings. We perceive this, and affectively respond with some further bodily movements of our own. To the extent that we are all doing this, we are all *empathically mirroring each other*. Indeed, there is even a specialized neuronal system set up within us for empathic mirroring, closely associated with learning natural languages.[3]

Thus we are each directly acquainted with consciousness$_{lo}$ in two different but fully complementary ways: first, just by *being* a minded animal and thereby having an essentially embodied conscious, intentional mind; and second, just by *living together* with other minded animals in the natural world.[4] The mental life of an animal with consciousness$_{lo}$ is not an epistemic or metaphysical "mystery."[5] On the contrary, the mental life of an animal with consciousness$_{lo}$ is an *irreducible, complex natural fact*, just like the weather and organismic life, and just as utterly *un*mysterious.

To be sure, the weather, organismic life, and minded animals are all self-organizing thermodynamic systems with emergent truly global or inherently dominating intrinsic structure, and not mere mechanisms like a can-opener or a digital computer (see Chapters 7–8). But something can be non-mechanistic and have emergent truly global intrinsic structure without being in any sense epistemically or metaphysically mysterious. The difference between mere mechanisms and self-organizing systems is a categorical difference between *kinds of*

[3] Arbib and Rizzolatti, "Neural Expectations: A Possible Evolutionary Path from Manual Skills to Language"; Arbib, "From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics"; and Gallese, "The 'Shared Manifold' Hypothesis: From Mirror Neurons to Empathy."

[4] See, e.g., Thompson, "Empathy and Consciousness." In effect, the classical Cartesian Interactionist Substance Dualist "other minds problem"—the skeptical worry that asks: How can I ever know the existence or specific character of another mind if I am only ever directly acquainted with the contents of my own consciousness or with the surfaces of external bodies?—evaporates if creatures with a consciousness$_{lo}$ are essentially embodied minds. For if we are essentially embodied, then we are directly acquainted with our own and other minds just by being directly acquainted with our own minded animal bodies and with other minded animal bodies. The evaporation of the other minds problem does *not*, of course, guarantee that you will never be self-deceived or mistaken about the thoughts, feelings, and intentions of others. Alas! But at least you won't have to wonder (seriously) whether your cat and the people living next door are robots.

[5] For a "mysterian" view of consciousness, see McGinn, "Can We Solve the Mind-Body Problem?"

natural dynamics, but not a categorical difference in *ontological levels* (see Section 7.2).

The particular goal of this chapter and the next is to provide *a neurophenomenological analysis* of consciousness$_{lo}$. By the notion of a "neurophenomenological analysis"[6] we mean the following three-part project:

(i) to describe consciousness$_{lo}$ (including its various mental acts, contents, and targets) as it appears to the first person,

and

(ii) to frame some necessary a priori claims about conscious minds$_{lo}$ on the basis of those first-person descriptions,

but also

(iii) to make these claims cohere as closely as possible with empirical evidence from the cognitive neurosciences, including cognitive psychology, medical neurology, neurophysiology, and neurobiology.

In Sections 1.1 to 1.3, and in Sections 2.1 to 2.4, we neurophenomenologically unpack the nature of a consciousness$_{lo}$ by describing some of its basic types and some of its necessary structures. This leads us to the general thesis that creatures with consciousness$_{lo}$ and intentionality$_{lo}$ are *essentially embodied*, and in turn, to the two sub-theses that the primary manifestation of our essential embodiment is the subjective experience of *desire-based emotion*, which in turn is originally given in *primitive bodily awareness*. In other words, creatures with consciousness$_{lo}$ are, essentially, *desiring minded animals*.

If true, then this entails what we will dub the Essentially Embodied *Cogito: I desire, therefore I am*. Of course, some desiring minded animals are also thinkers, or rational animals: every reader of this book, for instance. Thus the classical *Cogito* is true too. But the classical *Cogito* is misleading if construed as stating both a sufficient *and* necessary condition for being a minded animal, for not all minded animals have a capacity for rationality. For example, all or most non-human animals, third trimester human fetuses, human infants, human toddlers, the insane, many or most victims of Down's syndrome, many or most victims of Alzheimer's disease in its later phases, and so on, are all either *non-rational* or *proto-rational* minded animals.

[6] See, e.g., Hanna and Thompson, "Neurophenomenology and the Spontaneity of Consciousness."

On the other hand, even rational human animals must be able to desire as a necessary foundation of their capacity for thinking. Indeed, in general, it seems that we think only because we have a *felt need* to do so. As Aristotle pointed out in the *Metaphysics*, we inherently desire to know. But in order to know, we must think; and in order to think, we must *want* to think. It also seems very plausible to hold that the particular topics of thought which result from acts of rational conscious attention are selected by our short-term and long-term interests. Of course, thoughts will sometimes come unbidden into the mind! Nevertheless, it seems that those are either the result of ordinary non-self-conscious, pre-reflective, non-deliberative desires, and so nothing to be terribly alarmed about; or else, in extreme and unfortunate cases, the result of mental illness and psychopathology, which presumably generate involuntary or perverse desires. So in either case it seems that all thinking like ours has its origin in some sort of desire-based emotion.

1.1 Some Preliminaries

In order to be as clear as possible, we will start with a few brief definitions, terminological explications, and methodological remarks. These preliminaries are crucial for many of the arguments, explanations, and formulations coming up later. Readers wanting to go directly to something slightly less abstract, however, can jump to the next section and then return to this section later as needed.

It is plausible to hold that the nature of the natural or physical world is revealed to us by three *basic* natural sciences—i.e., *physics* (including astrophysics and cosmology, and also molecular, atomic, and quantum physics), *chemistry*, and *biology*. Together they provide a comprehensive theory of *natural thermodynamics*: matter, energy, motion, force, elementary processes, and organismic life. All the other natural sciences presuppose this triad of natural sciences. In turn, the three basic natural sciences go together to provide a comprehensive theoretical picture—sometimes called "the Scientific Image"[7]—of the causal-dynamic natural world.

[7] See, e.g., Oppenheim and Putnam, "Unity of Science as a Working Hypothesis"; and Sellars, "Philosophy and the Scientific Image of Man." For two different critical reflections on this philosophical picture, see Hanna, *Kant, Science, and Human Nature*; and Van Fraassen, *The Scientific Image*.

In light of this, we will say that a property *P* is a *fundamental physical property* of something *X* if and only if *P* is a necessary, internal property of *X* and *P* is correctly attributed to *X* by at least one of the three basic natural sciences.[8] Non-fundamental physical properties, in turn, are physical properties that strongly supervene on the fundamental ones. We can then define *the natural or physical world* as what corresponds to the set of all correctly ascribed, necessary, internal physical properties in physics, chemistry, or biology, together with whatever is logically or nomologically strongly supervenient on this set.

For us, a property *P* is an *internal property* of something *X* if and only if the instantiation of *P* in *X* constitutes a *proper part* of *X*.[9] For example, having a finger is an internal property of a human hand. But it is possible to have human hands that lack fingers. A property *P* is then an *intrinsic property* of something *X* if and only if *P* is a necessary, internal property of *X*. For example, having four sides of equal length is an intrinsic property of a square. Intuitively, it is an *inherent* property of a square that it have four equal sides. Understood this way, then, the terms 'intrinsic property' and 'inherent property' are synonyms.

Here it must be especially noted that we use the term 'intrinsic property' to mean *a necessary, internal, non-relational* ***or*** *relational property of something*. Therefore we explicitly allow for both

(i) **intrinsic non-relational** properties of things (e.g., the whiteness of a piece of chalk)

and also

(ii) **intrinsic relational** properties of things (e.g., the right-handedness of a hand).

Unfortunately, this is not the *standard* usage of 'intrinsic' in contemporary analytic metaphysics. In fact, there *is* no absolutely standard usage.[10] But the most common or widespread usage of 'intrinsic', deriving remotely from Leibniz, has it that an intrinsic property is just an *inherent monadic* or

[8] There is obviously some sort of explanatory circularity involved in defining *fundamental* physical properties in terms of the *basic* natural sciences. But in this context, the circularity seems benign.

[9] For a plausible analysis of the part-whole relation, see Koslicki, *The Structure of Objects*.

[10] See Weatherson, "Intrinsic vs. Extrinsic Properties"; and for an influential attempt to pin down the notion of intrinsicness, see also Langton and Lewis, "Defining 'Intrinsic'."

non-relational property of things. So on that usage, the intrinsic vs. extrinsic distinction is the same as the distinction between inherent non-relational properties and relational properties. But this usage misleadingly conveys, without argument and by mere stipulation, the false implication that relational properties must all be *extrinsic* or accidental, external properties, and can never be *inherent* properties. On the contrary, it is plausibly arguable that there is a perfectly real and widespread class of inherent *relational* properties.[11] For example, the properties of *globally orientable spaces*—i.e., comprehensive spaces containing directions like up, down, right, left, behind, in front—are necessary, internal, relational properties of the real material things that are embedded in those spaces, e.g., human body parts like hands. And the same thing goes for real material things that are embedded in globally asymmetric or dynamically irreversible time-relations—i.e., time-relations that imply *time's arrow*—like past, future, before, and after, e.g., living organisms like human beings.

Following on from this crucial point, when intrinsic relational properties are specifically based on globally orientable or dynamically irreversible spacetime structures, we call them *intrinsic* ***structural*** *properties*.

We must also define the concepts of *strong supervenience* and *materialist supervenience*, in view of the highly important roles they have played in contemporary mainstream philosophy of mind.

The main idea behind strong supervenience is that it captures a modal dependency relation between types of properties that is somewhat weaker than identity, hence consistent with the denial of identity between properties of the relevant types, and thereby consistent with Property-Dualism-Without-Substance-Dualism. Roughly and simply put, some property P^1 strongly supervenes on another property P^2 just in case P^2 *necessarily determines* P^1. Or in other words, "fixing" the existence and specific character of P^2 thereby "fixes" the existence and specific character of P^1.

But more carefully and precisely now, we can separate all properties into two broadly distinct classes: the *lower-level* or more basic properties, and the *higher-level* or less basic properties. Call the lower-level properties "*B*-properties" and the higher-level properties "*A*-properties." Then *A*-properties strongly supervene on *B*-properties if and only if

[11] See, e.g., Humberstone, "Intrinsic/Extrinsic."

(1) necessarily anything that has some property *G* among the *A*-properties also has some property *F* among the *B*-properties (or equivalently: no two things can share all their *B*-properties in common unless they also share all their *A*-properties in common; or again equivalently: no two things can differ in any of their *A*-properties without also having a corresponding difference among their *B*-properties),

and

(2) necessarily anything's having *F* is sufficient for its also having *G*.

If we assume that the *B*-properties are fundamental physical properties and that the *A*-properties are fundamental mental properties, then this yields a materialist supervenience. In this context, feature (1) of materialist supervenience is known as the "necessary covariation" of the mental with the physical, and feature (2) is known as the "upwards dependence" of the mental on the physical.

A materialist supervenience can be further qualified either by *modal strength*, of which there are two basic kinds, or by *scope-of-supervenience-base*, of which there are three basic kinds. The two basic kinds of modal strength are

(i) *logical strong supervenience*, according to which the term 'necessarily' in the original definition means *true in every logically possible world*,

and

(ii) *nomological strong supervenience*, according to which the term 'necessarily' in the original definition means *true in every logically possible world having the same set of general causal natural laws as the actual world*.

And the three basic kinds of scope-of-supervenience-base are

(a) *local strong supervenience*, according to which the *B*-properties apply to all individual material objects or substances,
(b) *regional strong supervenience*, according to which the *B*-properties apply to all material domains larger than individuals but smaller than the whole material world,

and

(c) *global strong supervenience*, according to which the *B*-properties apply to the whole material world.

So defined, local strong supervenience entails both regional and global; but regional strong supervenience entails neither global nor local; and global strong supervenience entails neither regional nor local.

Corresponding to what we said in the second paragraph of this section about the nature of the natural or physical world, it is also plausible to hold that the nature of the mental, or mentality, is revealed to us by neurophenomenology and the cognitive neurosciences. Since consciousness$_{lo}$, on our view, is the necessary and sufficient mark of the mental, we hold that a property *P* is a *fundamental mental property* of something *X* if and only if *P* is an intrinsic (whether non-relational or relational) property of *X* and *P*'s being correctly attributed to *X* by neurophenomenology or the cognitive neurosciences, entails *X*'s having consciousness$_{lo}$. All the mental properties of a creature with a consciousness$_{lo}$ are fundamental mental properties of that creature. Non-fundamental mental properties—e.g., the syntactic and semantic properties of the logics and natural languages we rationally cognize, the aesthetic properties of works of art, the exchange value of money, etc.—are mental properties that strongly supervene on the fundamental ones. In short, we hold that rational human animals *cognitively construct* logics and natural languages,[12] and presumably the same goes for works of art and monetary systems. We can then define *the mental world*, or *mentality*, as what corresponds to the set of all consciousness$_{lo}$-entailing correctly ascribed intrinsic properties in neurophenomenology and the cognitive neurosciences, together with whatever is logically or nomologically strongly supervenient on this set.

One last remark, by way of wrapping up the preliminaries. Our working assumption from a methodological point of view is that the philosophy of mind is essentially *the triangulating, comprehensive* ***Science of Minds***$_{lo}$, or alternatively, *the triangulating, comprehensive* ***Science of Minded Animals***, because it simultaneously employs three distinct sub-methods in conjunction:

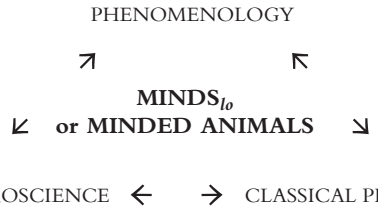
[12] See, e.g., Hanna, *Rationality and Logic*.

(1) phenomenology,
(2) cognitive neuroscience,

and

(3) classical philosophical reasoning, including logic, conceptual analysis, and modal metaphysics.

So, diagrammatically presented, our methodological conception of the philosophy of mind is this:



The special two-sided collaboration between phenomenology and cognitive neuroscience is *neurophenomenology*. The special two-sided collaboration between classical philosophical reasoning and cognitive neuroscience is *neurophilosophy*. And the special two-sided collaboration between phenomenology and classical philosophical reasoning is *subjective philosophy*. Each of these two-sided conceptions is perfectly legitimate, but somewhat restricted in scope. So by sharp contrast with each and all of the special two-sided conceptions, our over-arching *triangulating* conception is motivated by three leading ideas:

(i) that the core fact of minds$_{lo}$ or minded animals is equally accessible to phenomenology, cognitive neuroscience, and classical philosophical reasoning,
(ii) that each contributing sub-method equally needs to be *supplemented* by each of the other two sub-methods if it is to make its concepts and claims fully meaningful,

and

(iii) that all three methods *taken together* are required in order adequately to illuminate and explain the core fact of minds$_{lo}$ or minded animals.

1.2 The Nature of Consciousness$_{lo}$

To say that a creature has a consciousness$_{lo}$ is not to say that it is always *occurrently* conscious. It is obviously possible for a minded animal to be temporarily *un*conscious—e.g., in a fainting fit, in a seizure, drugged, or in a coma. But even a temporarily unconscious minded animal must also have a *capacity, disposition*, or *power* for consciousness$_{lo}$, in the sense that it possesses some properly functioning *natural matrix*—or natural basis—of consciousness$_{lo}$, that can be triggered into occurrent or actualized consciousness$_{lo}$ under appropriate real-world conditions. The destruction or permanent shut-down of this natural matrix entails the non-existence of any sort of consciousness$_{lo}$, and indeed also ends the *personal* life of any minded animal that is also a rational animal. Contrapositively, the continuing existence of this natural matrix also *sustains* the overall life of a person whose mental life is in temporary hiatus during periods of non-persistent unconsciousness or coma. On our view, the natural matrix of consciousness$_{lo}$ necessarily *includes* the brain—this puts the *neuro* into *neuro*phenomenology and into the cognitive *neuro*sciences—but as we shall argue later, it is by no means *restricted* to the brain.

As we just pointed out, creatures with consciousness$_{lo}$ are not always occurrently conscious. But on our view, necessarily whenever a creature with a consciousness$_{lo}$ is in *any* sort of mental state, then it is *also* occurrently conscious in some definite way, even if only minimally. So occurrent consciousness$_{lo}$ penetrates into *every* aspect of our mental lives. Call this *the Deep Consciousness Thesis*.[13] As we shall see in a moment, the Deep Consciousness Thesis has several extremely important implications.

But first, a brief elaboration. By the Deep Consciousness Thesis we do not mean to say that minded animals are always *maximally* conscious, even when they are occurrently conscious. For a minded animal might be in a mental state that is, in a certain sense, *relatively non-conscious*—e.g., dreaming sleep, dreamless sleep, trances, automatism, reflex action, divided attention, peripheral awareness, subliminal awareness, Freudian psychodynamism, cognitive priming, and "tacit" computational information processing—and not *fully* and *paradigmatically* conscious. Nevertheless,

[13] See also Hanna, "Kantian Non-Conceptualism."

these states of relative non-consciousness all necessarily involve some definite degree and some definite structural kind of minor deviation from our normal, full, and paradigmatic condition of attentive, singly-focused, alert, waking consciousness$_{lo}$. Therefore, relative non-consciousness in this sense still implies the existence of occurrent consciousness. So the Deep Consciousness Thesis is saying that even relatively non-conscious states in minded animals are necessarily at least *minimally* occurrently conscious in some *definite* way or another.

Indeed, and perhaps most radically, the Deep Consciousness Thesis entails that in rational human animals all information processing—e.g., visual information processing or linguistic information processing that occurs without any self-conscious or self-reflective deliberative intention to look at something or to understand what someone says or writes—is minimally and definitely occurrently conscious. Obviously we are not usually attending focally to that sort of mental activity, nor is that sort of activity usually very vivid. We generally have bigger and more interesting things to think about. Nor, to the extent that we are usually *not* thinking about that sort of mental activity, do we usually have any occurrent *conceptual* access to it. But it does not follow that in and through information processing we are not first-person aware *at all*. Even holding information content fixed, it simply *feels* very different to receive the same information visually as a picture (say, in a movie) and visually in words (say, in the corresponding novel).

As a case in point, consider Robert Bresson's very moving (and like all of Bresson's films, very *slow*-moving) 1951 film, *Diary of a Country Priest*, based on the same-named novel by George Bernanos. Bresson presents scene after scene in which the protagonist, a young priest dying of tuberculosis, is shown writing out words and saying them simultaneously, while the events he is describing are *also* simultaneously shown on film, sometimes with diaphanous images of the written words layered over them as well. The gradual and combined subjectively experiential effect of the three types of simultaneous information processing over the same information content is cinematically stunning. But this sort of subjective experience would be impossible if this information processing were unavailable to consciousness$_{lo}$.

Furthermore, any sort of breakdown in normal mental processing—say, being temporarily unable to remember something already very well known, such as one's own home telephone number, or more dramatically the

pathological agnosias, e.g., the inability to recognize even very familiar faces (a.k.a. "prosopagnosia")—has an immediate, vivid, and often highly disturbing subjectively experiential character. Consider now a patient with a brain tumor who suddenly loses the ability to recognize a loved one's face, and then just as suddenly recovers that ability and also remembers having lost it. What he consciously experienced was his own temporary inability to recognize faces, *or the actual breakdown of his otherwise smoothly operating mental processing ability*—the lower-level scanning processor running over the loved one's perceived face again and again, without any cognitive purchase, just like those frustrating moments when your PC or Mac freezes or stalls out when you are trying to download or upload something, only it is unimaginably worse, because the stall-out is happening *inside* his own body. Suppose further that he never tells anyone about this experience, and that there are no overt behavioral effects of it. Later, catastrophically, he loses all ability to recognize faces but also, mercifully, he also loses the ability to know that he has lost this ability. Surely, however, the early stages of his prosopagnosia are essentially terrifying for him. It is like something out of Kafka. He experiences what it is like to live in a world without familiar faces—a world in which every face is a mask. The complex and intensely emotionally-charged specific character of these subjective experiences seems to be explained only by holding that necessarily all information processing is minimally and definitely occurrently conscious.

This implies that by holding the Deep Consciousness Thesis we are rejecting what has sometimes been treated as an unquestionable truism of contemporary philosophy of mind and cognitive neuroscience, namely, the existence of a fundamental difference between

(1) the Conscious Mind (or *first-personal, conscious* information processing)

and

(2) the Computational Mind (or *sub-personal, non-conscious* information processing).

Indeed, this difference is so sharply defined that it has yielded a new Cartesian problem—which Ray Jackendoff aptly calls "the Mind–Mind Problem"—of how there can ever be genuine two-way causal or semantic

interaction across the gap between the first-personal and sub-personal levels.[14] But the Deep Consciousness Thesis entails that, on the contrary, all information-carrying and information-constructing processes in minded animals—like all neurobiological processes in minded animals—must be at least minimally and definitely occurrently conscious, *if* conscious, intentional minds$_{lo}$ are essentially embodied.

In this way, essentially embodied consciousness$_{lo}$ goes *all the way down* to the ground floor of information processing, and a new solution to the Mind–Mind Problem is made possible, according to which all information processing in an animal minded like us is not *sub-personal* but in fact still *first-personal and conscious*, although *non-conceptual*[15] and *non-self-conscious or non-self-reflective*. Furthermore, on the reasonable assumption that neurocomputation is causally efficacious, the Deep Consciousness Thesis also entails the causal efficacy of consciousness$_{lo}$ in all neurocomputational processes and states. Consciousness$_{lo}$ does not epiphenomenally float *above* neurocomputation, precisely because it is already necessarily *embedded in* neurocomputation, and beyond, right out to the skin.

The Deep Consciousness Thesis may at first seem shockingly unorthodox. But properly understood, it is much less shocking than it may seem. One fundamental source of philosophical confusion in this area is that the very idea of consciousness$_{lo}$ or subjective experience, i.e., "the first-personal," is deeply ambiguous as between, on the one hand,

(a) *self-consciousness$_{lo}$ or self-reflection*,

which is the ability of a creature with a consciousness$_{lo}$ to have conscious meta-representational states, or conscious thoughts about itself and its own mental states; and on the other hand, what Evan Thompson aptly calls

(b) *sensorimotor subjectivity*,[16]

[14] See, e.g., Jackendoff, *Consciousness and the Computational Mind*.

[15] The thesis of Non-Conceptualism says that representational content is neither wholly nor solely determined by a conscious animal's conceptual capacities, and that at least some contents are both solely and wholly determined by its non-conceptual capacities. The version of Non-Conceptualism that we favor, *essentialist content* Non-Conceptualism, says that the representational content of a state is essentially non-conceptual if and only if its semantic structure and psychological function are inherently different from the structure and function of conceptual content. See, e.g., Gunther (ed.), *Essays on Nonconceptual Content*; Hanna, "Kantian Non-Conceptualism"; and Speaks, "Is There a Problem about Nonconceptual Content"?

[16] See Thompson, "Sensorimotor Subjectivity and the Enactive Approach to Experience."

which is the more primitive ability of conscious, intentional suitably neurobiologically complex, situated, forward-flowing living organisms to have what Thomas Nagel aptly calls a "single point of view."[17] In turn, we hold, this ability of minded animals to have a single point of view is grounded in egocentrically-centered and spatially oriented essential embodiment, and a primitive bodily awareness that includes proprioception (including kinaesthesia, and the sense of orientation and balance), bodily pleasures and pains, tickles and itches, the feeling of pressure, the feeling of temperature, and the feelings of vitality or lethargy.

Here is another way of putting the same distinction. It is one thing for a minded animal to be *a first-person in the high-powered sense of self-conscious or self-reflective agency and rationality*. But it is distinctively another thing for a minded animal to be *a first-person in the lower-powered sense of conscious pre-reflective intentional agency and desire-based volition*.[18] Every minded animal is a first-person in the lower-powered sense, precisely because it is obvious that no minded animal could be a self-conscious or self-reflective and rational agent without *also* being a conscious, intentional agent and capable of desire-based volition. But not every minded animal is a first-person in the higher-powered sense, precisely because it is equally obvious that not every minded animal is a *rational* animal.

The crucial point here is that self-consciousness$_{lo}$ or self-reflection requires pre-reflectively conscious sensorimotor subjectivity, but pre-reflectively conscious sensorimotor subjectivity does not require self-consciousness$_{lo}$ or self-reflection. For example, at least some non-human animals—e.g., Nagel's bat[19]—and all normal human infants have sensorimotor-subjective states that are not also self-conscious or self-reflective. And again, when I am skillfully driving my car but thinking about philosophy, the conscious states that skillfully control my driving are sensorimotor-subjective but not in any way self-conscious or self-reflective. Indeed, *all day long* we are all making simple, actively guided intentional body movements—sitting, standing, walking, stretching, balancing, twiddling our fingers, wiggling our toes, looking this way and that,

[17] Nagel, "What it is like to be a bat?," 166–7.

[18] See also Sartre, *Transcendence of the Ego*.

[19] See note 17 above. Nagel famously argues that since we have no meaningful third-person conception of what it is like to be a bat, i.e., of the specific phenomenal character of a bat's conscious experience, then by the same token we cannot have any meaningful third-person natural scientific conception of conscious experience more generally.

clearing our throats, licking our lips, humming or whistling under our breaths, etc.—which are subjectively experienced as freely willed basic actions but *not* experienced as self-conscious or self-reflective deliberative actions, as we focus attentively on other more exciting and important things. But just because massively many of our consciously freely willed activities are not terribly exciting or important, this does not imply that they do not exist! On the contrary, they are the constant and necessary background hum of our conscious lives as minded animals. So since, presumably, *everyone* would agree that normal human infants and at least some non-human animals are minded animals but not also self-conscious or self-reflective animals, and also that it is possible to drive a car consciously but not self-consciously or self-reflectively, and also that all day long we are making simple, actively guided conscious intentional body movements that are not also done self-consciously or self-reflectively, then at least implicitly everyone *already* concedes a distinction between pre-reflectively conscious sensorimotor subjectivity and meta-representational subjectivity.

Hence it is not so very shocking after all for us to hold that all mental states in a creature with a consciousness$_{lo}$, even so-called tacit computational information processing states, are also at least minimally occurrently and definitely conscious. All we are saying is that even so-called tacit computational information processing involves *pre-reflectively conscious sensorimotor subjectivity*, but not self-conscious or self-reflective meta-representational subjectivity.

Again, it needs to be especially emphasized that by our saying that all information processing in minded animals is minimally occurrently conscious in the sense of sensorimotor subjectivity, we do *not* mean to say that we are always occurrently conscious *of* this information processing, or consciously attentive *to* it—either in the strong sense that we are self-consciously or self-reflectively aware of it (e.g., as a self-directed belief or verbal report), or even in the slightly weaker meta-representational sense that we can at least consciously generate explicit mental models or imagery of it. Instead, on our view *all* forms of our mental activity are fundamentally manifest to us, as conscious sensorimotor subjects, in pre-reflective desire-based emotional feeling and primitive bodily awareness. If so, then all information processing is fundamentally manifest to us in this way as well.

For us then, the notion of something's being so-called tacit information processing only means:

that information processing occurs in a non-focal, non-vivid, non-meta-representational, non-conceptual, and *conatively affective* pre-reflectively conscious process or state in a motile, suitably neurobiologically complex, egocentrically-centered and spatially oriented, forward-flowing living organism,

and does *not* mean:

that information occurs in a mental process or state that excludes all consciousness$_{lo}$ whatsoever.

What we are calling "conatively affective" consciousness$_{lo}$ is the same as *desire-based emotive* consciousness$_{lo}$. So the so-called *tacit* dimension of minds$_{lo}$ is just the pre-reflective desire-based *emotive* dimension of minds$_{lo}$. What has been mistakenly regarded as "the cognitive unconscious"[20] is just *a pre-reflective non-cognitive consciousness*$_{lo}$. We will come back to the seminal notion of conatively affective consciousness$_{lo}$ a few paragraphs below.

Now what, more precisely and more specifically, is a consciousness$_{lo}$? On our view, consciousness$_{lo}$ is *the subjective experience of a suitably neurobiologically complex living organism*. This formulation has three distinct but also complementary components:

(1) the subjective component,
(2) the experiential component,

and

(3) the neurobiological component.

So let us consider each one briefly in turn.

First, to say that consciousness$_{lo}$ is *subjective* is to say that it necessarily involves an egocentrically-centered and spatially oriented embodiment of consciousness$_{lo}$, that it necessarily includes a single point of view, that it is necessarily a pre-reflectively conscious sensorimotor subjectivity, and also that it necessarily is "immanently reflexive"—by which we mean

[20] See Kihlstrom, "The Cognitive Unconscious."

that it *intrinsically contains an immediate sense of self*, which we will discuss more fully in Sections 2.2 and 2.3. By virtue of its being subjective, a consciousness$_{lo}$ also *can*, but need not necessarily, include capacities for meta-representation and self-consciousness or self-reflection. On our view, all and only the *rational* animals with consciousness$_{lo}$ also possess such capacities.

(There is a further subtlety in this connection that is worth mentioning in passing. The subtlety arises from the fact that it seems to be possible for a minded animal to possess a capacity for conscious meta-representation, which involves merely a *consciousness* of its own consciousness$_{lo}$, without also possessing a capacity for self-consciousness or self-reflection, which involves a *belief* or *thought* about its own consciousness$_{lo}$. For example, human toddlers and many non-human animals seem to have a capacity for meta-consciousness—especially in the form of higher-order desires about first-order desires—without also having a capacity for self-consciousness$_{lo}$ or self-reflection. This point is a crucial one for our theory of action in Chapters 3–5, where we will argue that intentional agency does *not* require self-conscious, deliberative intentions.)

Second, to say that consciousness$_{lo}$ is *experiential* is to say that it necessarily involves some or another kind of conatively affective content (based on primitive bodily awareness), sensory content (based on sense perception of the world and acts of thought), and representational content (also based on sense perception of the world and acts of thought). So consciousness$_{lo}$ entails intentionality$_{lo}$.We will come back to this point in a few pages, and also again in Sections 2.2 and 2.3.

Finally, and most importantly, to say that consciousness$_{lo}$ *belongs to a suitably neurobiologically complex living organism* is to say that our mental properties are necessarily instantiated in all the vital neurobiological systems, organs, and processes of our living bodies—including the higher brain, brain stem, limbic system, nervous system, endocrine system, immune system, and cardiovascular system—right out to the skin. This is what we call *the Essential Embodiment Thesis*, and it is important to note that it has two logically distinct parts:

(1) the *necessary* embodiment of conscious, intentional mind$_{lo}$ in a living organism (the Necessity Thesis),

and

(2) the *complete* neurobiological embodiment of conscious, intentional $minds_{lo}$ in all the vital systems, vital organs, and vital processes of our living bodies (the Completeness Thesis).

The Necessity Thesis says that necessarily, $minds_{lo}$ are alive. Negatively formulated, it says that $minds_{lo}$ cannot be dead, disembodied, or purely mechanical.

By contrast, the Completeness Thesis says that $minds_{lo}$ are fully spread out into our living bodies, necessarily including the brain, but also necessarily not *restricted* to the brain. Please note that we are *not* saying that our brains, hearts, livers, or stomachs are either conscious or intentionally directed! On the contrary, according to our view it is only complete *minded animals*, including all *real human persons*, that are conscious or intentionally directed, not their body parts alone, and not even their brains alone. So what we are saying by asserting the Completeness Thesis is that every minded animal, including every real human person, is conscious or intentional *with,* or *in-and-through*, its brain, heart, liver, stomach, or whatever, right out to the skin.

One could, at least in principle, assert the Necessity Thesis and also reject the Completeness Thesis. But we want to assert both of them together. So we hold that the supposed consciousness of a causally detached brain—say, a living brain floating listlessly in a vat, as in Hilary Putnam's famous thought-experiment[21]—even though it seems both conceivable and logically possible, necessarily would not be a consciousness *like ours*. On our view, a $consciousness_{lo}$ necessarily involves a brain that is causal-dynamically coupled with all the other vital systems, organs, and processes of our living body.

The notion of a "causal-dynamic coupling" is crucial. The Necessity Thesis and Completeness Thesis do *not* jointly entail that $consciousness_{lo}$ actually is or ever could be embodied in *any* causally necessary condition of our specific kind of consciousness, which would of course include all sorts of entities and facts outside our living bodies. That is what we will later call *the Embodiment Fallacy* (see Section 8.1). Instead the Necessity and Completeness theses jointly entail that $consciousness_{lo}$ is embodied *only* in a

[21] See Putnam, *Reason, Truth, and History*, ch. 1.

certain kind of fully integrated dynamic system that is both causally necessary and causally sufficient for our specific kind of consciousness—namely, one that has *all the same causal powers as* the vital systems, organs, and processes of our living bodies. Any such living body is the natural matrix, or natural basis, of a consciousness$_{lo}$.

And that point in turn raises another extremely important point that is specifically about the very idea of a "natural matrix" of consciousness$_{lo}$. A natural matrix of consciousness$_{lo}$ is not merely a *compositional material substrate*—a mass of specific bodily stuff and a collection of particular body parts—that necessarily accompanies and supports consciousness$_{lo}$. A natural matrix is instead a system of causal-dynamic relations, embedded in some or another compositional material substrate, awaiting specific activation or actualization. This means that if you significantly modify the shape of your body, or lose a limb or some other body part, *without also replacing it with an equivalent counterpart that has the same relational causal powers*, then you would also correspondingly modify your mind. But the *specific bodily stuff* and the *particular body parts* are not metaphysically important. The mere matter doesn't really matter.

In *Meditations* VI, and while auspiciously wearing his Substance Dualism hat, Descartes makes a similar point:

> Although the whole mind seems to be united to the whole body, I recognize that if a foot or arm or any other part of the body is cut off, nothing has been thereby taken away from the mind.[22]

But our *reason* for making this point is radically different from that of Descartes. In his substance dualist guise, Descartes holds that the mind is an absolutely homogeneous and simple unity, and thereby indivisible. But *our* point is about the metaphysics of living animal bodies like ours, not about the metaphysics of mental substance. Again, what we hold is that the natural matrix of consciousness$_{lo}$ is *not* just a hunk of specific bodily stuff and *not* just a heap of particular bodily parts. Instead, the natural matrix of a consciousness$_{lo}$ is all the vital systems, organs, and processes of our living bodies, *as individuated by their relational causal powers*, that is, by what they can efficaciously do in causal community with each other and with the larger surrounding world. That these vital systems, organs, and processes

[22] Descartes, *Meditations on First Philosophy*, 59, AT 86.

are in fact composed of some or another hunk of specific bodily stuff and also of some or another heap of particular bodily parts—say, specifically human body stuff and particular human body parts—is of course extremely practically important for members of the relevant species made out of that stuff and those parts, but otherwise it is metaphysically trivial. Thus the Essential Embodiment Thesis is a thesis about the *operative neurobiological dynamics* of creatures with consciousness$_{lo}$, and not (except trivially) a thesis about our compositional material substrate.

Assuming, then, that the Completeness Thesis is formulated in terms of the relational causal powers of the vital systems, organs, and processes of our living bodies, and not (except trivially) in terms of their compositional material substrate, there are at least four good reasons for defending Completeness.

First, it seems obvious that if any of the vital systems, organs, or processes in our bodies is destroyed or permanently disabled without a functional replacement that has essentially the same relational causal powers—say, an artificial heart, a liver transplant, etc.—then our consciousness will cease to exist, precisely because the whole organism *dies*. Therefore the *existence* of consciousness$_{lo}$ necessarily depends on its complete neurobiological embodiment.[23]

Second, it seems equally obvious that significant changes made to the relational causal powers of any of our vital systems, organs, or processes normally produce correspondingly significant changes in the specific character of conscious minds$_{lo}$. And this is as true of the *non-brain* systems as it is of the *brain* systems. A thyroid gland malfunction, hormone imbalance, adrenaline surge, or heart attack is apt to cause highly significant changes in consciousness$_{lo}$. Therefore the *specific character* of consciousness$_{lo}$ also necessarily depends on its complete neurobiological embodiment.[24]

To be sure, a lobotomy or a concussive blow to the head is apt to cause *more* fundamental changes in consciousness$_{lo}$ than a thyroid malfunction, hormone imbalance, and so on. And again, to be sure, the brain is

[23] The relevant set of neurobiological properties alone is not a *sufficient* condition of the existence of consciousness$_{lo}$, however. Instead, the existence of consciousness$_{lo}$ is *jointly hylomorphically constituted* by the relevant mental and neurobiological properties. See Section 7.1 below.

[24] Just as in the case of the existence of consciousness$_{lo}$, so too the relevant set of neurobiological properties alone is not a sufficient condition of the *specific character* of a consciousness like ours. Both the existence and specific character of a consciousness$_{lo}$ are jointly hylomorphically constituted by relevant mental and neurobiological properties. Again, see Section 7.1 below.

centrally causally involved in every aspect of normal attentive, singly-focused, alert, self-reflective, waking consciousness$_{lo}$. So we are not in any way denying the necessary and central causal role of the brain in the constitution of our normal attentive, singly-focused, alert, self-reflective, waking consciousness$_{lo}$ and intentionality$_{lo}$. But at the same time, we are also strongly recommending that philosophers of mind and cognitive neuroscientists should not *overemphasize* the causal role of the brain,[25] to the extent that this undermines our recognition of the equally necessary role of the relational causal powers of the rest of our vital systems, organs, and processes. For example, as everyone knows, even fairly minor changes in our *digestive* processes can produce non-trivial changes in our consciousness. Think, e.g., of the striking phenomenological differences between:

(a) feeling very hungry and craving a plate of spaghetti,
(b) feeling as if you ate just the right amount of spaghetti,

and

(c) feeling utterly stuffed with spaghetti.

The brain obviously is centrally causally involved in these normal attentive, singly-focused, alert, waking phenomenological differences, but it seems also equally obvious that the brain does not in and of itself causally *control or determine* these differences. On the contrary, it seems obvious that the "enteric brain"—our guts—is doing much of the causally controlling and determinative work here.[26] And similar points can be made about the other non-brain vital organs, systems, and processes. Each of them can and does play a causally controlling and determining role with respect to some differences in normal attentive, singly-focused, alert, self-reflective waking consciousness$_{lo}$, even if the brain is also centrally causally involved.

Analogously, even if every basic act of a corporation passes directly through its Chief Executive Officer, it does not follow that the CEO controls or determines the specific character of *every* such act, or even *most* of them. In fact, in a great many cases the CEO is just *the chief executive slave* of the controlling determinations of the shareholders (if it is a public

[25] For example, the 29 January 2007 issue of *Time* magazine was entirely devoted to the topic, *The Brain: A User's Guide*, and included supportive articles by or interviews with many leading contemporary philosophers of mind and cognitive neuroscientists.

[26] See, e.g., Gershon, *The Second Brain*.

company), or of the employees (if it is either an employee-owned company or unionized), or of the actual business operations of the company. So too the brain is often just *the central causal slave* of the rest of the living body.

Third, there is neurophysiological empirical evidence that supports the Completeness Thesis. For example, recent work on the neurochemistry of human emotions strongly suggests that the vital systems centrally causally involved with and embodying our basic emotions are gut-based, not brain-based.[27]

But fourth and finally, probably the most compelling empirical evidence for Completeness, precisely because it is the simplest, is the well-known fact that the "arc" of reflex action (say, someone's pulling her hand away from something very hot) operates more quickly than the time it takes for the brain to process information sent to it via the nervous system about the body parts involved in that reflex action (say, that the subject's hand has been seriously burned). If the Deep Consciousness Thesis is true, then this is *also* a sensorimotor-subjective experience, although not of course a normal self-conscious or self-reflective experience. In the example of the burned hand, the subject's hand moves *before* she self-consciously or self-reflectively feels the searing pain of a burn. But we think that reflex action still has a special *phenomenology*, in the classical Nagelian sense that there is a definite something-it-is-like-to-be, and a particular point of view, for a suitably neurobiologically complex living organism with a mind_{lo}, when, e.g., that minded animal is pulling her hand away from something very hot even though the self-conscious searing pain of the burn has not yet emerged. Reflex action necessarily includes a first-order "immanently reflexive" *pre-reflective* $\text{consciousness}_{lo}$, even if it does not necessarily include a higher-order *self-conscious or self-reflective* $\text{consciousness}_{lo}$.

Along with the Deep Consciousness Thesis, a further reason to think that *reflex action* is indeed *reflexively conscious*, although in a pre-reflectively conscious, sensorimotor-subjective way, is that it is possible to train oneself, through biofeedback strategies, to modulate or even suppress such reflexes.

So if the Deep Consciousness Thesis is correct, and if we also take biofeedback data seriously, then even in cases of simple reflex action a pre-reflective $\text{consciousness}_{lo}$ always occurs *with* and *in-and-through* the vital

[27] See, e.g., Damasio, *Descartes' Error*; Damasio, *The Feeling of What Happens*; Damasio, *Looking for Spinoza*; Pert, *Molecules of Emotion*; and Prinz, *Gut Reactions*.

systems that constitute and subserve our intentional body movements, even though by hypothesis the brain is *not* centrally causally involved in the production of these spontaneous, pre-reflective intentional actions. Or in other words, there is compelling empirical evidence that there is a complete neurobiological embodiment of consciousness$_{lo}$ even when the brain is only *peripherally* causally involved.

Moreover, there is a direct metaphysical pay-off from this conclusion. As W.T. Rockwell aptly puts it:

> [I]f the brain does not record certain features of a perception that the mind is nevertheless aware of, this must mean that the mind is not identical with the brain.[28]

But this is not Dualism. For Rockwell, and also for us, a conscious mind$_{lo}$ is not identical to the brain, and thus a conscious mind$_{lo}$ is not reducible to the brain, not because a mind$_{lo}$ is in any way metaphysically separable either from the brain or from the vital systems of the living body more generally, but instead just because the embodiment of consciousness$_{lo}$ *goes much further out into the living body than the brain*.

In this connection, it also needs to be re-emphasized that the necessary, complete neurobiological embodiment of minds$_{lo}$ does *not* entail that we are necessarily or even normally conscious *of* our vital systems and organs or of their dynamic operations. On the contrary, it seems clear that we are only occasionally conscious *of* them, in the sense that they become the objects of our singly-focused, vivid, self-conscious or self-reflective attention—e.g., when listening to my own heartbeat with a stethoscope, or groaning with stomach ache because I ate too much spaghetti, or convulsing in pain because I have seriously burned my hand. But the essential embodiment of minds$_{lo}$ does indeed entail that we are always and necessarily conscious *with* all our vital systems, organs, and processes in their dynamic operations via our sensorimotor subjectivity—that is, we are always and necessarily pre-reflectively and sensorimotor-subjectively conscious *in-and-through* our living animal bodies.

Here is another example to illustrate the same fundamental point. Imagine that you are standing on a platform waiting for a bus or train, and absorbed in reading a paperback novel. Now consider the difference between

[28] Rockwell, *Neither Brain Nor Ghost*, 47.

(1) absent-mindedly shifting your body weight from your normal left leg onto a sore right leg

and

(2) absent-mindedly shifting your body weight from your normal left leg onto a normal right leg.

Only in case (1) do you become conscious *of* your leg as a singly-focused, vivid topic of thought. But if one were inclined to conclude from this that case (2) does not also involve a specific character of consciousness$_{lo}$ with or in-and-through the living body, then one should stop and consider now the difference between case (2) and

(3) absent-mindedly shifting your body weight from your normal left leg onto a *phantom* right leg.

Clearly there is a huge phenomenological (and of course also practical, and neurobiological) difference between the pre-reflectively conscious sensorimotor-subjective experience of *balancing on your normal right leg* and the pre-reflectively conscious sensorimotor-subjective experience of *falling through your phantom right leg*. Indeed, catastrophically, this second sort of case actually happens to some amputees.[29]

One could try to argue against the point we are making by claiming that the phenomenological difference between the two cases is that the former experience is "phenomenally blank," while the latter is not.[30] But if that were true, then how could the former experience have been authentically an *experience* of the conscious subject? Subjective experiences without any phenomenal character would be like events without intrinsic temporal structure, i.e., an impossibility. The upshot is that the Essential Embodiment Thesis is directly supported by neurophenomenological evidence deriving from the pre-reflectively conscious sensorimotor-subjective experience of primitive bodily awareness, or primitive consciousness$_{lo}$-*with* the body, or primitive consciousness$_{lo}$-*in-and-through* the body, which in turn is sharply different from both

[29] See Gallagher, *How the Body Shapes the Mind*, 90.

[30] We owe this helpful objection to one of the anonymous readers for OUP.

(i) bodily self-consciousness$_{lo}$ or bodily self-reflection

and also

(ii) the body image,

which in turn are distinct sub-species of consciousness$_{lo}$-*of* the body. We will return to the crucial neurophenomenological distinction between consciousness$_{lo}$-*with*, or consciousness$_{lo}$-*in-and-through* the body, and consciousness$_{lo}$-*of* the body, in Sections 2.3 and 2.4.

Presumably no one would doubt that consciousness$_{lo}$ is subjective and sensory. Yet as regards the experiential component, perhaps it is somewhat controversial[31] for us to say that consciousness$_{lo}$ necessarily involves representational content, for this tightly ties consciousness$_{lo}$ to the possibility of *intentionality*$_{lo}$—which, as we have said, is the directedness of consciousness$_{lo}$ to objects, actions, locations, events, other conscious creatures or itself (a.k.a. "intentional targets"), or the "aboutness" of a conscious mental state via its content.[32] So in this respect we agree with Terence Horgan and John Tienson, who have argued for what they call the "Intentionality of Phenomenology Thesis," or IP Thesis.[33] Horgan and Tienson's IP Thesis says:

Mental states of the sort commonly cited as paradigmatically phenomenal (e.g., sensory-experiential states such as color-experiences, itches, and smells) have intentional content that is inseparable from their phenomenal character.[34]

[31] *Somewhat* controversial, but not *radically* controversial. This is because there are *first-order representational* theories of consciousness that echo G. E. Moore's famous remarks about the "transparency" of consciousness, and try to explain consciousness entirely in terms of the objects, properties, and relations represented in perceptual or propositional states. See, e.g., Dretske, "Conscious Experience"; and Moore, "The Refutation of Idealism," 37. There are also *higher-order* representational theories of consciousness—see, e.g., Rosenthal, "A Theory of Consciousness"; and Rosenthal, "Two Concepts of Consciousness."

[32] Close readers will also have noticed that this definition makes consciousness$_{lo}$ a necessary condition of intentionality$_{lo}$, and entails what we will later call the "Phenomenology$_{lo}$ of Intentionality$_{lo}$ Thesis," or $P_{lo}I_{lo}$ Thesis. See also Searle, *Rediscovery of the Mind*. In evaluating the $P_{lo}I_{lo}$ Thesis, it should also be remembered that we are concentrating exclusively on consciousness *like ours* and intentionality *like ours*, and also that we are assuming that the Deep Consciousness Thesis is true.

[33] Horgan and Tienson, "The Intentionality of Phenomenology and Phenomenology of Intentionality."

[34] Ibid., 520.

But our version of the IP thesis (which we will call the "Intentionality$_{lo}$ of Phenomenology$_{lo}$ Thesis" or $I_{lo}P_{lo}$ Thesis) is even *stronger* than Horgan and Tienson's thesis. Our $I_{lo}P_{lo}$ Thesis says:

> Necessarily *all* states of a consciousness$_{lo}$, even ones that are neither "paradigmatic" in Horgan and Tienson's sense, nor cited as such, are also characterized by intentionality$_{lo}$.

Part of what enables us to radically strengthen the IP Thesis in this way is our distinction between pre-reflective sensorimotor-subjective consciousness$_{lo}$, and self-consciousness$_{lo}$ or self-reflection. But the other part is our commitment to the thesis that all conscious intentionality in minds$_{lo}$ includes *intrinsic spatiotemporal relations*. For us, consciousness$_{lo}$ is inherently *egocentrically centered* in orientable space and also inherently *flowing forward* in thermodynamically irreversible time.[35] Hence it is possible for consciousness$_{lo}$ simply to be intentionally *there-directed* in orientable space towards some immediate location or another, or intentionally *forward-directed* in thermodynamically irreversible time towards some immediately future event or another (Husserl called this primitive form of temporal intentionality "protention"[36]), without there being any determinate objects as *further* targets of intentionality. We will spell out this crucial point further in Sections 2.3 and 2.4.

Many contemporary theorists of consciousness want, on the contrary, to restrict consciousness to the domain of the non-intentional. This is usually because they believe that intentionality can be adequately explained in functionalist or physicalist terms alone. So, quite reasonably, these philosophers have adopted a divide-and-conquer strategy for giving a complete materialist explanation of the mind:

> *first*, right now and in the near future, given our recent accomplishments in cognitive neuroscience, we will explain intentionality in functionalist or physicalist terms (the "easy" problem of consciousness),

and then

[35] See, e.g., Husserl, *Phenomenology of Internal Time Consciousness*; Ismael, *The Situated Self*; and Pred, *Onflow: Dynamics of Consciousness and Experience*.

[36] See Husserl, *Phenomenology of Internal Time Consciousness*.

second, much later, hoping for large future advances in cognitive neuroscience, we will somehow explain non-intentional phenomenal consciousness (the "hard" problem of consciousness).[37]

But suppose that the $I_{lo}P_{lo}$ Thesis is correct, and thus all consciousness$_{lo}$ necessitates intentionality$_{lo}$. And suppose, further, that Horgan and Tienson's "Phenomenology of Intentionality Thesis," or PI Thesis, is also correct:

Mental states of the sort commonly cited as paradigmatically intentional (e.g., such cognitive states as beliefs, and conative states as desires), when conscious, have phenomenal character that is inseparable from their intentional content.[38]

Finally, suppose even further that *all* intentionality$_{lo}$ necessitates consciousness$_{lo}$. This is what we will call the "Phenomenology$_{lo}$ of Intentionality$_{lo}$Thesis" or $P_{lo}I_{lo}$ Thesis. The $P_{lo}I_{lo}$ Thesis follows directly from Horgan and Tienson's PI Thesis, together with our Deep Consciousness Thesis. Then since the $I_{lo}P_{lo}$ Thesis and the $P_{lo}I_{lo}$ Thesis are both correct, it follows that it is impossible to solve the so-called *easy* problem of consciousness without also solving the *hard* problem of consciousness. If that is right, then the two-stage functionalist or physicalist train of explanation is in fact running round and round a vicious loop, going nowhere.

That is controversial enough. It is, however, perhaps even more controversial for us to say that consciousness$_{lo}$ necessarily involves the possibility of *conative affect*, for this tightly ties the nature of consciousness$_{lo}$ to the possibility of *desire-based emotion*,[39] or "emotion$_d$" for short, by which we mean the total package of overlapping psychological capacities and facts that includes caring of all sorts, salient drives of all sorts, inclinations of all sorts, liking and disliking of all sorts, love and hate, lust and disgust, moods of all sorts, passions of all sorts, pleasures and pains of all sorts, feelings of all sorts, sensations of all sorts, and sentience of all sorts.

[37] See, e.g., Chalmers, *The Conscious Mind*. Chalmers distinguishes between consciousness, which is non-intentional, and *awareness*, which is intentional. See also Kim, *Physicalism, or Something Near Enough*.

[38] Horgan and Tienson, "The Intentionality of Phenomenology and the Phenomenology of Intentionality," 520.

[39] See also: Damasio, *Descartes' Error*; Damasio, *The Feeling of What Happens*; and Damasio, *Looking for Spinoza*.

Obviously it is possible to draw finegrained distinctions between different sorts of emotion—for example, some sorts of emotion have propositional or conceptual content, and some do not. But this does not undermine our more basic point, which is that the possibility of desire-based affect and therefore the possibility of desire-based emotion pervades *every* aspect, *every* species, and *every* sub-species of consciousness$_{lo}$ and intentionality$_{lo}$. This thought is captured by the Essentially Embodied *Cogito*: *I desire, therefore I am*. Now on our view, as we shall argue in detail in chapters 3 to 5, emotion$_{d}$ is the psychological foundation of all choice, volition, or willing. If that is correct, then as Sartre very aptly puts it, (the capacity for) consciousness$_{lo}$ and (the capacity for) choice necessarily entail one another.

By sharp contrast, many and perhaps even most contemporary philosophers of mind want to restrict consciousness narrowly to the domain of what seems to be affectlessly, emotionlessly, passionlessly, passively sensory—e.g., seeing red.[40] But why? One reason, no doubt, is that they do find the "hard" problem of consciousness—i.e., explaining phenomenal consciousness as paradigmatically exemplified in externally-caused sensory states, in purely functional or physicalist terms—to be so very *hard*. Hence they never get beyond working on that particular problem. But another diagnostic hypothesis, quite consistent with the first, is that they also restrict consciousness to affectless, emotionless, passionless, passive, externally-caused sensory states because they have already explicitly or implicitly conceded that consciousness is *epiphenomenal*, or causally dependent on the physical world (and in particular, the brain) for its existence and specific character, yet without any causal powers of its own.[41]

But as we wondered aloud earlier, how and why would something with causal powers of its own produce something that has no causal powers of its own? That seems completely mysterious. On our view, by contrast, all consciousness$_{lo}$—even just seeing red—intrinsically has a desire-based emotive character and a set of efficacious causal powers that are spread out completely into our living bodies and adequately expressed in our abilities to make intentional body movements. This claim cannot be fully defended

[40] See, e.g., Humphrey, *Seeing Red: A Study in Consciousness*.
[41] See, e.g., Jackson, "Epiphenomenal Qualia."

or elaborated until we explicitly address the problem of intentional action in Chapters 3 to 5, and the problem of mental causation in Chapters 6 to 8.

For the moment, however, we need only note that to connect consciousness$_{lo}$ intrinsically with desire-based emotion and choice is to imply that all consciousness$_{lo}$ is inherently *poised for our trying to do something*.[42] Indeed, except in rarefied philosophical contexts, even just seeing red normally involves the experience of some degree of attraction towards and excitement about red objects. Why else would the official color of Valentine's Day be red? Why else would fire engines be red? Why else would stop signs and stoplights be red? Why else would Michael Powell have chosen red to be the color of those dance slippers in his deliriously dreamy 1948 film *The Red Shoes*? And what about Dorothy's magical ruby red slippers in Victor Fleming's classic 1939 musical, *The Wizard of Oz*? Of course there are various sorts of conceptual associations at play here. But even setting those aside, the ordinary subjective experience of just seeing red does seem to go well beyond inertly "beholding a red sense datum," or "sensing reddishly." Even in ordinary speech, the expression 'seeing red' can also mean *being angry as hell and just about to explode into violent movement*.

Finally, it is most certainly *quite* controversial for us to defend the Essential Embodiment Thesis, including both the Necessity Thesis and the Completeness Thesis. This is true even if we leave aside Cartesian Interactionist Substance Dualism and its corresponding Property Dualist Thesis to the effect that the disembodied existence of consciousness$_{lo}$ is possible (the possibility of spirits), and concentrate just on contemporary materialists, who might be initially and reasonably supposed to be quite sympathetic to embodiment theses.

As regards the Necessity Thesis however, and the necessary connection between consciousness$_{lo}$ and biological life, no doubt many contemporary materialist philosophers of mind, and especially those influenced by Reductive Functionalism, will want to deny it. Reductive functionalists type-identify a mind with an abstract lawlike system of computational or causal-theoretical mappings from inputs to the organism or its brain to outputs from the organism or its brain, and often also token-identify a mind with whatever actually plays the role specified by that functional

[42] See, e.g., O'Shaughnessy, "Trying (as the Mental 'Pineal Gland')."

organization.[43] In any case, classical functionalists want to bind consciousness and all other mental facts to essentially inert and mechanical facts, extrinsic relational computational or causal properties, and linear dynamic processes that can be *multiply realized* in different kinds of inert matter or compositional stuff—even if it is true that *local reductions* for, say, human pain, to say, firing C-fibres, are possible.[44] The paradigm analogy for classical functionalists is the operations of a universal Turing machine, or digital computer, which can be implemented in many different sorts of hardware. Correspondingly, the notion of the wide-ranging multiple realizability of minds like ours accounts for the meaning of the slogan, very popular during the heyday of classical Functionalism, that "the mind is compositionally plastic."

But other contemporary materialist philosophers are moving in a decidedly *post-functionalist* direction. They believe that functional organizations can be multiply realized *only* in physical systems with essentially different kinds of relational causal powers. So in order for a functional organization to be realized in two or more ways, each realizer must have an essentially different set of relational causal powers.[45] Trivial differences in compositional material substrate and trivial differences in non-relational or relational causal powers are not sufficient for distinct realization. For example, according to this approach to multiple realization,

(1) a bottle opener painted red is not a distinct realization of the functional kind *bottle opener* from an otherwise identical bottle opener painted blue,
(2) a bottle opener made out of metal-reinforced high strength plastic is not a distinct realization of the functional kind *bottle opener* from an identically-shaped bottle opener made entirely of metal,

and

(3) an ordinary key-shaped bottle opener is not a distinct realization of the functional kind *bottle opener* from an ordinary bottle-top-grabbing opener.

[43] See, e.g., Block, "What is Functionalism?"; Kim, *Philosophy of Mind*, chs. 5–6; and Putnam, "The Nature of Mental States."
[44] See, e.g., Kim, "Multiple Realization and the Metaphysics of Reduction."
[45] See, e.g., Searle, *Rediscovery of the Mind,* chs. 3–5; and Shapiro, *The Mind Incarnate.*

If we understand multiple realization in this way, then it seems very likely that the multiple realizability thesis, as applied to conscious, intentional $minds_{lo}$, is *false*. On the contrary it seems very likely that, as a matter of necessity, conscious, intentional $minds_{lo}$ are *not* multiply realized in physical systems with essentially different kinds of relational causal powers, but rather must be instantiated in the actual world in *only one type of physical thing*, namely, living organisms of a suitable degree of neurobiological complexity. Although we are not materialists, this conclusion is certainly grist for our mill.

While we are considering the similarities between our view and post-functionalist reductive materialism, it is also important to draw attention to another emerging division within the camp of contemporary reductive materialists. Many reductive materialist philosophers of mind who defend the necessary embodiment of mind, either in the sense that minds simply do not exist and there are nothing but brains and other purely physical things in a purely physical world, or in the sense that consciousness is nothing but a brain process—e.g., eliminative materialists like the Churchlands, and classical mind–brain identity theorists like Place and Smart—also assume a *Cartesian Materialist* thesis to the effect that necessary embodiment is necessarily limited to the brain, or perhaps even necessarily limited only to certain parts of the brain.[46] In any case, Cartesian Materialism, as we use that term, entails that $consciousness_{lo}$ is restricted to a causally isolable and decontextualizable brain alone—and in the limit case, to a mere Putnamian brain-in-a-vat. But while we think it that it is quite true that the brain, *along with all the other vital systems, organs, and processes*, is causally and metaphysically necessary for $consciousness_{lo}$, we also think that it simply does not follow that a brain-in-the-vat is either causally or metaphysically *sufficient* for $consciousness_{lo}$.[47] How could a merely *envatted* brain ever have all the same causal powers as the completely neurobiologically *embedded* brain? Obviously a merely envatted brain could not stand in the same causal *relations* to the other vital systems, organs, and processes of the living body as the embedded brain. Hence a merely envatted brain could not possibly have the same set of relational causal *powers* as the embedded brain. If the

[46] "Cartesian Materialism" is a useful critical label that has, however, been used in slightly different ways. See, e.g., Dennett, *Consciousness Explained*; and Rockwell, *Neither Brain Nor Ghost*.

[47] See also Rockwell, *Neither Brain Nor Ghost*, chs. 2–5 and 9; and Shapiro, *The Mind Incarnate*, ch. 6.

Completeness Thesis is true, then consciousness$_{lo}$ includes as a necessary part of its nature that it is instantiated in something that has all the same relational causal powers as the embedded brain. Because our brain is an embedded brain, envatted brains are not *brains like ours*.

So we see no compelling reasons at all for restricting consciousness$_{lo}$ to the non-intentional, to the purely sensory, to the essentially inert and mechanical, to the merely computationally or causal-theoretically functional, or to a causally isolable and decontextualizable (i.e., envatted) brain. Surely the richer and thicker view of consciousness$_{lo}$ as inherently intentional, emotive$_d$, completely neurobiologically embodied, situated, forward-flowing, and causal-dynamically engaged with the natural world is far better supported by evidence supplied by neurophenomenology, the cognitive neurosciences, and neurological clinical medical practice as well.[48] Indeed, it seems to us overwhelmingly obvious that creatures with consciousness$_{lo}$ are *conscious, intentional, desiring, suitably neurobiologically complex, egocentrically-centered and spatially oriented, thermodynamically irreversible living organisms*. So, again, perhaps the most relevant question is: Why would anyone ever have thought *otherwise*?

1.3 Essential Embodiment and the Cartesian Mistakes

By way of an answer to that question, our diagnostic hypothesis is that what makes so many classical, recent, and contemporary philosophers and other scientists of the mind think otherwise is what can be called *the set of five classical Cartesian dualist conceivable possibilities*:

(i) that it is clearly and distinctly conceivable and therefore logically possible for a creature to be conscious but incapable of being directed towards anything;
(ii) that it is clearly and distinctly conceivable and therefore logically possible for a creature to be conscious but incapable of having emotions$_d$;

[48] See, e.g., Anderson, "Embodied Cognition: A Field Guide"; Blakesee and Ramachandran, *Phantoms in the Brain*; Clark, *Being There*; Damasio, *The Feeling of What Happens;* Heidegger, *Being and Time*, 67–273; Merleau-Ponty, *Phenomenology of Perception*; Sartre, *Being and Nothingness*, parts 3–4; Rockwell, *Neither Brain Nor Ghost*; Searle, *Rediscovery of the Mind*; Shapiro, *The Mind Incarnate*; Sacks, *The Man Who Mistook his Wife for a Hat*; and Thompson, *Mind in Life*.

(iii) that it is clearly and distinctly conceivable and therefore logically possible for a creature to be conscious but non-living or purely mechanical;
(iv) that it is clearly and distinctly conceivable and therefore logically possible for a creature to be conscious but disembodied;

and

(v) that it is clearly and distinctly conceivable and therefore logically possible for a creature to be conscious but nothing but a brain-in-a-vat.

In fact, there seem to be three important Cartesian mistakes here. The first is the obvious mistake of being entranced by the substance dualist tendencies of the *Meditations* and failing to heed the neurophenomenology of essentially embodied minds. In point of fact, however, even Descartes *himself* does not consistently adhere to Interactionist Substance Dualism and the five classical Cartesian dualist conceivable possibilities, for as we saw in the famous "mind–body union" text we cited in the Introduction, he is sometimes strongly drawn towards a very different metaphysical picture of the mind–body relation. This picture is deepened in *Passions of the Soul*, as the epigraph for the present chapter shows, to a *complete* embodiment thesis that connects the existence and nature of the mind to the organismic unity of the whole body.

This alternative Cartesian mind–body doctrine is clearly metaphysically inconsistent with Interactionist Substance Dualism, which requires the *necessary mutual substantial independence* of mind and body, and allows only for *contingent* mind–body connections. Nevertheless for Descartes what "my nature teaches me" and what "nature teaches us," in contradistinction to his own official Dualism, is that mental–physical connections in creatures minded like us are *necessarily mutually interdependent*, by way of what he called "the passions," which we would re-describe as the set of our capacities for having emotive$_d$, volitional, living, completely embodied, and causal-dynamically engaged states of consciousness$_{lo}$ and intentionality$_{lo}$. Indeed it is even arguable that Descartes himself, as opposed to the classical Interactionist Substance Dualist tradition, is ultimately committed to this *passionate* conception of the mind and correspondingly to the necessary mutual interdependence thesis of the Dual-Aspect Theory and

our Mind–Body Animalism, and not to the classical and official Cartesian necessary mutual independence thesis.[49] If so, then paradoxically enough the real Descartes is *not* a card-carrying Cartesian in every one of his guises. In fact, in at least one of his guises, the real Descartes is really playing on *our* team.

The second Cartesian mistake is closely connected with the first. We think that it is crucial to distinguish between conscious minds *unlike* ours (ghostly minds, purely abstract minds, angelic minds, divine minds, etc.), and conscious minds *like ours* (including, of course, all minded human and non-human animals). Cartesian dualism seems to be the natural consequence of attempting to run these two sharply different categories of minds together under the umbrella of a single undifferentiated theory. More precisely, the substance dualist line of thought seems to be that because we are normally capable of some degree of rationality, and because we place a high normative value on rationality, and because ideal rationality is supposedly divine, angelic, and unaffected by bodily desires or emotions, then our minds must be assimilated to essentially *disembodied* minds. In this way, Interactionist Substance Dualism effectively occludes the crucial concept of a mind_{lo}.

But as we see it, home is where the heart, brain, and the rest of the living body *all* are, and this is where specifically *our* kind of consciousness and *our* kind of intentionality necessarily are too. Thus the philosophy of mind is first and foremost the philosophy of the essentially embodied Cartesian *passionate* mind, and not first and foremost the philosophy of the substance dualist Cartesian *dispassionate* mind. Indeed, for us, logical rationality like ours and practical rationality like ours are *also* essentially embodied.[50] As we mentioned above, on our view the philosophy of mind is *the triangulating, comprehensive* ***Science of Minds***$_{lo}$ or *the triangulating, comprehensive* ***Science of Minded Animals***.

The third Cartesian mistake—a subtler mistake than the first two, but perhaps even more deeply-rooted, pernicious, and stubborn for that very reason—is the false assumption that *logical* possibility determines *metaphysical* possibility. Logical possibility is the formal consistency of a proposition with the laws of logic. By contrast, metaphysical possibility is the semantic

[49] See. e.g., Brown, *Descartes and the Passionate Mind*.
[50] See, e.g., Hanna, *Rationality and Logic*, ch. 7, 231.

consistency of a proposition with the ontological constituents of any possible world in which that proposition is true. To be sure, contemporary critics of the famous Cartesian *Meditations* VI argument for Substance Dualism—an argument that more recently has been resuscitated by Saul Kripke and David Chalmers as a modal argument for Property-Dualism-Without-Substance-Dualism[51]—have vigorously attacked the more or less explicit inferential step from clear and distinct conceivability to logical possibility. We hold that this inferential step is fully acceptable, and shall have much more to say in Section 6.3 about why we hold that.

But the point we are making here is a different one. What we are criticizing is the different assumption, made by virtually *every* participant in the contemporary debate, whether dualist or materialist, to the effect that there is one and only one basic kind of necessity (and correspondingly, one and only one basic kind of possibility), namely *logical* necessity (and correspondingly, *logical* possibility). This assumption is the thesis of *Modal Monism*.

What, more precisely, does Modal Monism look like? Logical necessity is the truth of a proposition according to the laws of logic alone.[52] And logical possibility is the consistency of a proposition with the laws of logic alone. So logical necessity is the truth of a proposition in every logically possible world, and logical possibility is the truth of a proposition in at least one logically possible world. Conditional (a.k.a. "hypothetical" or "relative") necessity is then standardly defined in terms of logical necessity and the material conditional arrow '$\rightarrow$', as just the logical entailment (i.e., logically necessary material conditional implication) of a proposition that is the consequent of such an implication, by a set of non-logical axioms or postulates included as the antecedent of that very implication.[53] So if we symbolize logical necessity as "L," and an arbitrarily-chosen set of non-logical axioms or postulates as "Γ", then the conditional necessity of a proposition P is standardly defined as:

$$L(\Gamma \rightarrow P).$$

[51] See Kripke, *Naming and Necessity*, 148–55; and Chalmers, *The Conscious Mind*, ch. 4.

[52] To be sure, there has been a vigorous debate about what *logic* is. See Hanna, *Rationality and Logic*, esp. ch. 2. For our purposes in this chapter however, the definition of logical necessity and logical possibility is officially neutral as between different kinds of logic.

[53] See, e.g., Montague, "Logical Necessity, Physical Necessity, Ethics, and Quantifiers"; and Smiley, "Relative Necessity."

When Γ is the set of actual causal laws of nature, then *P* is *nomologically* or *physically* necessary. And similarly, with appropriate changes, for logical possibility, conditional possibility, and nomological or physical possibility. Finally then, we can say that Modal Monism is the doctrine that all forms of necessity (or possibility) are either logical necessity (or logical possibility) itself, or else definable in terms of logical necessity (or logical possibility) in the standard way.

Modal monism can in turn be elaborated by *Two-Dimensional Modal Semantics*, which maps meaningful terms from logically possible worlds to extensions in two different ways, according to two different intensions (a.k.a., the "primary intension" or "1-intension," and the "secondary intension or "2-intension"). The 1-intension (e.g., the necessarily true proposition "Water is the watery stuff") is knowable a priori and maps from egocentrically-centered worlds "considered as actual," to extensions. By contrast, the 2-intension (e.g., the necessarily true proposition "Water is H_2O") is knowable only a posteriori and maps from possible worlds "considered as counterfactual"—i.e., logically possible non-actual variants on the actual world, each of which contain some stuff bearing an identity relation to some microphysically-defined stuff in the actual world—to extensions.[54] But both intensions are defined over *the total class of all logically possible worlds,* and if necessarily true, are always *logically* true. The metaphysical necessity of such propositions is thus at best *weakly* metaphysically necessary.

But suppose for a moment that, contrary to Modal Monism, there are *two* essentially different basic kinds of possibility, namely:

(1) logical or *weak* metaphysical a priori possibility (i.e., consistency with the laws of logic alone),

and

(2) non-logical or *strong* metaphysical a priori possibility (i.e., consistency with the laws of logic and *also* with all and only the universal intrinsic relational properties of the actual world, especially including its spatiotemporal structure, its global causal architecture, and its mathematical structure).

[54] See Chalmers, *The Conscious Mind*, ch. 2; and Chalmers, "The Foundations of Two-Dimensional Semantics."

And similarly, with appropriate changes, for necessity. This is the thesis of *Modal Dualism*. If Modal Dualism is correct, then the mere logical or weak metaphysical a priori possibility of a non-intentional, non-emotive, non-living, non-spatial, atemporal, causally isolated and disengaged (whether altogether disembodied, or merely brainy) consciousness will *not* entail that it is non-logically or strongly metaphysically a priori possible for a creature with minds_{lo} to exist in any of these ways. On the contrary, it seems to us non-logically or strongly metaphysically a priori necessary that all creatures with a $\text{consciousness}_{lo}$ are also capable of $\text{intentionality}_{lo}$, emotive_{d}, volitional, alive, necessarily and completely embodied, egocentrically-centered and spatially oriented, forward-flowing, and causal-dynamically engaged with the world. We elaborate and defend the notions of non-logical or strong metaphysical a priori possibility and necessity in Section 7.4.

Nevertheless, granting us for the moment at least the *intelligibility* of a significant modal dualist difference between logical possibility and necessity on the one hand, and non-logical or strong metaphysical a priori possibility and necessity on the other hand, it then seems that the mere logical or weak metaphysical a priori possibility of an exact physical counterpart of one of us that altogether *lacks* a $\text{consciousness}_{lo}$—a zombie in the philosophical sense[55]—will *not* entail that zombies are non-logically or strongly metaphysically a priori possible. Indeed, on our Essential Embodiment Theory of the mind–body relation, zombies are logically possible but also non-logically or strongly metaphysically a priori *impossible*,[56] since any exact physical counterpart of one of us would also be neurobiologically fully identical to that real human person, which would non-logically or strongly metaphysically a priori necessitate its also having a $\text{consciousness}_{lo}$. Indeed, we hold that $\text{consciousness}_{lo}$ and its complete neurobiological embodiment are non-logically or strongly metaphysically a priori necessarily equivalent.

But this is not an *identity* thesis because, as Kripke demonstrated in "Identity and Necessity," an identity of individuals (as named by rigid designators) is a *logically* necessary identity, and also an identity of properties or types (again, as named by rigid designators), which entails their *logical* equivalence—or more precisely, it entails their co-extensiveness across all

[55] Cf. Chalmers, *The Conscious Mind*, chs. 1–5.

[56] See also Kirk, *Zombies and Consciousness*.

logically possible worlds.[57] So our thesis of the non-logical or strongly metaphysically a priori necessary equivalence of mental properties and certain physical properties is not a *reductive* metaphysical thesis, as in the classical Mind–Brain Identity Theory.[58] We do not "downwardly identify" mental properties with certain physical properties of the brain. Neither do we "downwardly identify" mental properties with any *other* kind of physical properties. Nor, indeed, do we hold that mental properties are in any way asymmetrically necessarily dependent on—in the sense of being "upwardly determined" by, or strongly supervenient on (whether logically or even just nomologically)—any kind of physical properties. This is because we hold that consciousness$_{lo}$ and its complete neurobiological embodiment are *both non-identical and also reciprocally inherently related* via the reciprocal intrinsic relatedness of fundamental mental properties and corresponding neurobiological properties. This is the thesis of *mental-physical property fusion*, which says that fundamental mental properties and certain fundamental physical properties are both mutually irreducible to one another and also non-logically or strongly metaphysically a priori necessarily interdependent in all and only creatures minded like us, and thus necessarily jointly constitutive of all and only creatures minded like us. So mental properties are not "nothing but" certain physical properties but instead are *non-logically or strongly metaphysically a priori necessarily intertwined with* certain fundamental physical properties in all and only animals of a suitable level of neurobiological complexity.

These claims may sound abstruse and strange at first. But really, they provide only a carefully-formulated modal metaphysical interpretation and elaboration of what Descartes says in that celebrated passage in *Meditations* VI and again in *Passions of the Soul*:

> I am not merely present in my body as a sailor is present in a ship, but . . . I am very closely joined and, as it were, intermingled with it, so that I and the body form a unit.
>
> The soul is really joined to the whole body, and . . . we cannot properly say that it exists in one part of the body to the exclusion of the others. For the body is a unity which is in a sense indivisible because of the arrangement of its organs, these

[57] See Kripke, "Identity and Necessity"; and Kripke, *Naming and Necessity*. The standard or "textbook" interpretation of Kripke's theory is provided by two-dimensional modal semantics. See note 54 above.

[58] See, e.g., Place, "Is Consciousness a Brain Process?"

being so related to one another that the removal of any one of them renders the whole body defective. And the soul . . . is related solely to the whole assemblage of the body's organs.

In other words, irreducibly conscious, intentional $minds_{lo}$ are necessarily inherently reciprocally connected to living animal bodies like ours. So if what Descartes says here makes perfect sense to you, then so does what we are saying. To be sure, we explicitly add two further basic elements to Descartes's Passionate Mind picture of the mind–body relation: a theory of non-logical or strong metaphysical a priori necessity, and neo-Aristotelian hylomorphism. But neither of these is *ruled out* by anything Descartes says in this particular connection, and all things considered, they seem to augment and complete his Passionate Mind picture rather beautifully. So this is *not* Dualism, and it is *not* Materialism. It is a distinctively and radically *different* option—the Essential Embodiment Theory and its Mind–Body Animalism.

This page intentionally left blank

2

Consciousness$_{lo}$ and Essential Embodiment II: Types and Structures

> [T]he mind is utterly indivisible. For when I consider the mind, or myself insofar as I am merely a thinking thing, I am unable to distinguish any parts within myself; I understand myself to be something quite single and complete.
>
> René Descartes[1]

> [L]ife is the subjective condition of all our possible experience.
>
> Life without the feeling of the corporeal organ is merely consciousness of one's existence, but not a feeling of well- or ill-being, i.e., the promotion or inhibition of the powers of life; because the mind for itself is entirely life (the principle of life itself), and hindrances and promotions must be sought outside it, though in the human being himself, hence in combination with his body.
>
> Immanuel Kant[2]

2.0 Introduction

In *Meditations* VI and while prominently wearing his Interactionist Substance Dualism hat—as opposed to his Passionate Mind hat—Descartes asserts the *utter indivisibility* of the conscious human mind. It seems clear to us, however, that a conscious mind$_{lo}$ is neither homogeneous, nor single and complete in itself, nor indeed a "thing." On the contrary, a conscious

[1] Descartes, *Meditations on First Philosophy*, 59, AT 86.

[2] Kant, *Prolegomena to Any Future Metaphysics*, 76 (Ak 4: 335); and Kant, *Critique of the Power of Judgment*, 159 (Ak 5: 278).

$mind_{lo}$ is a *typed, structured, holistic set of spontaneous capacities or powers in a situated, forward-flowing, living organismic body of a suitable degree of neurobiological complexity*. In these respects, Descartes's sharply contrasting Passionate Mind conception is an important anticipation of Kant's biological, dynamic, and conatively affective conception of the conscious human mind in the *Critique of the Power of Judgment*. But leaving aside the historical provenances of the Essential Embodiment Theory, let us now consider some of the basic types of $consciousness_{lo}$.

2.1 Ten Types of $Consciousness_{lo}$

(1) Phenomenal $consciousness_{lo}$.
(2) Access $consciousness_{lo}$.
(3) Waking creature $consciousness_{lo}$.
(4) Non-waking creature $consciousness_{lo}$.
(5) State $consciousness_{lo}$.
(6) Intransitive $consciousness_{lo}$.
(7) Transitive $consciousness_{lo}$, $consciousness_{lo}$-of, or $intentionality_{lo}$.
(8) First-order transitive $consciousness_{lo}$.
(9) Higher-order transitive $consciousness_{lo}$.
(10) Immanent reflexivity, or the immediate sense of self.

What follows are brief descriptions of these ten types of $consciousness_{lo}$. We make no pretence of completeness here. There are almost certainly several other important and importantly different kinds of $consciousness_{lo}$—e.g., spatial $consciousness_{lo}$,[3] temporal $consciousness_{lo}$,[4] imaginational $consciousness_{lo}$,[5] and intersubjective or social $consciousness_{lo}$.[6] But by general contemporary philosophical consensus, the ten kinds are among the central ones. So we do need to acknowledge them before moving on to a more finegrained neurophenomenological analysis of $consciousness_{lo}$ in Sections 2.2 and 2.3.

[3] See, e.g., Eilan et al. (eds.), *Spatial Representation*. See also Section 2.3 below.
[4] See, e.g., Husserl, *The Phenomenology of the Consciousness of Internal Time*. See also Section 2.3 below.
[5] See, e.g., Sartre, *The Psychology of Imagininations*.
[6] See, e.g., Schutz, *The Phenomenology of the Social World*.

(1) *Phenomenal consciousness$_{lo}$*. Phenomenal consciousness$_{lo}$, according to Nagel's canonical formulation, is the *subjective character* of experience:

> [F]undamentally an organism has conscious mental states if and only if there is something it is like to *be* that organism—something it is like *for* the organism. We may call this the subjective character of experience.[7]

Similarly, Chalmers says that "a mental state is conscious if it has a *qualitative feel*—an associated quality of experience."[8] So phenomenal consciousness$_{lo}$ is the subjective qualitative feel of a mental state. We agree with Chalmers and Nagel that subjective qualitative feel is a necessary and sufficient condition of consciousness$_{lo}$. But we also strongly disagree with Chalmers and Nagel if they are further claiming, or at least implying, that phenomenal consciousness$_{lo}$ is the *solely* or *uniquely* necessary and sufficient condition of consciousness$_{lo}$—for we hold that subjective qualitative feel is only *one* aspect of subjective experience, and indeed an aspect that is fully *embedded*, in the sense that it cannot be either neurophenomenologically or metaphysically detached from a much larger and essentially richer set of factors.

For us, the neurophenomenological and metaphysical ground of all consciousness$_{lo}$ is *primitive bodily awareness*, not phenomenal consciousness$_{lo}$. To be sure, all primitive bodily awareness contains a subjective qualitative feel, and thus contains phenomenal consciousness$_{lo}$. But since this "feel" is always bound up with a sense of our whole living body, its egocentrically-centered spatial orientation, its forward-flowing thermodynamics, and its desire-based emotions, it is never correctly describable as a merely *raw* feel or a *thin* feel, that is, a merely *sensory* feel. On the contrary, primitive bodily awareness is always a deeply *robust* or *thick* feel. For example, you might attempt to detach a "bourbon-ish quale" from the experience of downing a neat shot of Buffalo Trace, and feeling the sudden burn as it hits the back of your throat, together with the lovely, warmly spreading way it encounters your stomach. This detachment is cognitively possible by an act of narrowly concentrated attention. But to consider the attentional tag-end of that intensely robust, thick, essentially embodied experience to be something independently real is to be on the verge of a highly misleading neurophenomenological and metaphysical abstraction.

[7] Nagel, "What is it like to be a bat?," 166.

[8] Chalmers, *The Conscious Mind*, 4.

(2) *Access consciousness$_{lo}$*.[9] According to the characterization offered by Ned Block, access consciousness$_{lo}$ applies to all and only those mental states whose contents are poised for (or in a later formulation, "broadcast for") thought, verbal report, and the control of self-conscious deliberative action. It seems possible for mental states to lack certain subjectively qualitative characters, and yet remain accessible for verbal report and poised for the control of action—as, e.g., in the phenomenon of blindsight, in which persons report the absence of visual sensations, and yet also are able to point with some accuracy to objects in the self-professedly blind parts of their visual fields.[10] Conversely, it also seems possible for mental states to be conscious but lack, at that moment, any determinate contents poised for (or broadcast for) thought, verbal report, or self-conscious, deliberative action-control—as, e.g., in the ordinary phenomenon of "spacing out" or "zoning out."

On our view, however, and according to the Deep Consciousness Thesis, it is not correct for Block to cite blindsight as a case of access consciousness$_{lo}$ without phenomenal consciousness$_{lo}$. For according to the Deep Consciousness Thesis, since *every* mental state in a minded animal is at least minimally occurrently conscious, then every state is also minimally occurrently *phenomenally* conscious, including blindsighted states. Indeed, using the distinction we noted in Section 1.2 between pre-reflectively conscious sensorimotor subjectivity on the one hand, and self-consciousness$_{lo}$ or self-reflection on the other, then blindsight is explicable in a way that is perfectly consistent with the Deep Consciousness Thesis. For we can then say that blindsight is guided by *pre-reflectively conscious sensorimotor-subjective vision*, even though blindsighters lack *self-conscious or self-reflective vision* for that cognitive and practical task. This pre-reflectively conscious capacity includes not only the roughgrained sensorimotor ability manifest in actual blindsight, but also the finegrained or hyper-finegrained—respectively, in the thought-experimental cases of what Block calls *superblindsight* and *superduperblindsight*[11]—sensorimotor connection between what blindsighters perceive in space, and their ability to point to it, discriminate it, or track it.

Otherwise put, we are saying that in blindsight the frontline information-processing mechanisms of the eyes and related areas of the wider brain–body

[9] See Block, "Concepts of Consciousness"; Block, "On a Confusion about a Function of Consciousness"; and Block, "Paradox and Cross Purposes in Recent Work on Consciousness."

[10] See Weiskrantz, *Blindsight*.

[11] Block, "Concepts of Consciousness," 211.

system are undamaged (blindsighters, after all, have their eyes open and are working under well-lit conditions) and continue to transmit sensorimotor-subjective visual information, even though the corresponding downstream mechanisms for processing self-conscious or self-reflective visual information have broken down. Blindsighters would then be best characterized as *sighted* in one sense of conscious vision, but *blind* in another sense of conscious vision. If that is correct, then blindsighters experience self-conscious or self-reflective *blindness* via the more sophisticated downstream processing mechanisms of the brain–body system, but also experience pre-reflectively conscious sensorimotor-subjective *sight* via the simpler processing mechanisms of the eyes. The notion of a divided consciousness is already theoretically familiar from well-known experiments involving divided attention tasks, and the dissociated cognitive abilities of patients who have undergone neo-commissurotomy—i.e., the recent surgical severing of the corpus callosum, the main connection between the right and left hemispheres of the brain[12]; so it should not therefore be very difficult to extend the same general idea to blindsight.

This in turn would neatly avoid the obvious paradox that in blindsight brute, non-conscious, non-unified, purposeless mental processing somehow exerts roughgrained, finegrained, or hyper-finegrained control over our conscious cognition and intentional body movements. It seems to us very implausible to hold that blindsighted people are *mere robots* in the blind areas of their self-conscious or self-reflective visual fields. On the contrary, it seems to us far more plausible that blindsighted people are still genuinely *visually conscious* in those blind areas, but in a way that is in some respects intrinsically phenomenologically, semantically, and neurobiologically different from the visual consciousness$_{lo}$ of normal self-consciously or self-reflectively sighted people.

(3) *Waking creature consciousness$_{lo}$*.[13] Waking creature consciousness$_{lo}$ is the subjective experience of an organism taken as a whole, insofar as it is awake. The opposite of waking creature consciousness$_{lo}$ is a creature's being *un*conscious.

(4) *Non-waking creature consciousness$_{lo}$*. Non-waking creature consciousness$_{lo}$ includes various non-waking but also still non-unconscious subjective

[12] See, e.g., Nagel, "Brain Bisection and the Unity of Consciousness."
[13] See Rosenthal, "A Theory of Consciousness"; Rosenthal, "Two Concepts of Consciousness."

experiences of the whole organism, such as dreaming, dreamless sleep, sleep-walking, trances, hypnotic states, fugue states, hysteria, etc.

Dreamless sleep is a particularly interesting case of consciousness$_{lo}$, in which there is a forward-flowing temporal conscious intentionality$_{lo}$ without any determinate *objects* of intentionality$_{lo}$. It is often held that dreamless sleep is non-conscious. But on the contrary, and consistently with the Deep Consciousness Thesis, we hold that dreamless sleep is just an affectively and emotively serene, very low-intensity, non-focused form of consciousness$_{lo}$. It is not as though one's mental life *stops* during dreamless sleep, and it always makes sense to ask someone *how well he or she slept*, even over and above the obvious intention to ask how he or she feels upon waking. It is also sometimes held that the fact of dreamless sleep shows that there can be consciousness$_{lo}$ without intentionality$_{lo}$. But this will be true only if one narrowly restricts the range of possible *targets* of intentionality$_{lo}$ to *objects*. Apart from objects, intentionality$_{lo}$ can also be directed to the intentional subject herself, to actions, to locations, and to *events*. Dreamless sleep, in particular, anticipates immediately future events of continuing serene, very low-intensity, non-focused subjective experience. And that is why when one is awakened by someone else, by a nearby sound or light, or by some unexpected internal bodily disruption, it comes as a vividly rude shock. "Being awakened" is something that merely *happens* to us. "Waking up," by contrast, other things being equal, is something we intentionally *do*. Waking up, other things being equal, is *up to me*. So, perhaps surprisingly, on our view, not only is dreamless sleep a form of conscious intentionality$_{lo}$, but also waking up from sleep (as opposed to being awakened from sleep) is a form of *spontaneous or pre-reflective intentional action*. We will come back to these and other closely-related points in Section 2.4 and Chapters 3 to 5.

(5) *State consciousness$_{lo}$*. State consciousness$_{lo}$, as opposed to either waking or non-waking creature consciousness$_{lo}$, is the subjective experience of a creature under specific conditions and as individuated by its phenomenal character or representational content. At any given moment and over any given stretch of waking or non-waking creature consciousness$_{lo}$, I am in some conscious, intentional state or another. Moreover, this state is dynamically complex. For example, right now I am peering at the screen of my Dell Latitude D810 laptop computer and leaning forward in my typing chair, banging away at the keyboard in a highly idiosyncratic, fairly

silly-looking 2.5 finger search-and-smash typing style, and simultaneously subjectively experiencing the computer, my chair, my intentional body positioning and intentional body movements, while also concentrating on what I am writing, what I have just written, and what I will be writing next. The overall mental life of a minded animal is thus composed of a (normally[14]) continuous sequence of occurrent and dynamically complex instances or tokens of state consciousness$_{lo}$.

(6) *Intransitive* consciousness$_{lo}$. Intransitive consciousness$_{lo}$ is any form of consciousness$_{lo}$ that is not object-directed, not action-directed, not location-directed, not event-directed, or not self-directed. As we have already mentioned in Section 1.2, we hold that there is no form of consciousness$_{lo}$ that is *absolutely* or *essentially* intransitive, or non-intentional. But it is perfectly consistent with this claim that there are forms of consciousness$_{lo}$ that are *relatively* or *accidentally* intransitive or non-intentional with respect to one or another of the basic classes of targets of intentionality$_{lo}$—e.g., objects. Thus temporarily spacing out or zoning out, dreamless sleep, and free-floating moods, are all cases of "objectless" transitive consciousness$_{lo}$.

(7) *Transitive* consciousness$_{lo}$, consciousness$_{lo}$*-of, or intentionality*$_{lo}$. This is any conscious state that is object-directed, action-directed, location-directed, event-directed, or reflexive. These objects, actions, locations, events, or oneself are *intentional targets*. It should be particularly noted that intentional targets need not always be *objects*—i.e., thinkable (syntactically and semantically well-formed) complexes of relatively determinate properties.[15] Moreover intentional targets, whether objects or otherwise, need not always actually *exist*. And also it is possible for the same target to be consciously intended in different ways, so the mapping from intentional contents to intentional targets is many-to-one.

In Section 1.2, we argued for radically strengthened versions of the Intentionality of Phenomenology (IP) and Phenomenology of Intentionality (PI) theses defended by Horgan and Tienson, which we dubbed the $I_{lo}P_{lo}$ Thesis and the $P_{lo}I_{lo}$ Thesis respectively. John Searle, Horgan, and Tienson have also argued that even the states of affairs represented by means of *propositional*

[14] There may also be some periods of unconsciousness or coma during which a minded animal's mental life goes into temporary hiatus even though the *person* continues to exist, precisely because the capacity for consciousness$_{lo}$ continues to exist in its natural matrix. See Section 1.2.

[15] See Meinong, "The Theory of Objects."

mental contents have phenomenal character.[16] This follows directly from the $P_{lo}I_{lo}$ Thesis, and obviously we accept it. Searle aptly calls the phenomenal character of propositional mental contents "aspectual shape."[17] The Aspectual Shape Thesis, to be sure, is controversial. Many philosophers of mind hold that propositional attitude states have no phenomenal character whatsoever. But a simple phenomenological example seems to show just the opposite. Say or think to yourself "The tiger is on the mat," and then say or think to yourself "The feline is on the mat." They feel *quite* different, and yet the concept FELINE is merely a determinable concept of the determinate concept TIGER, and so merely an analytic consequence of the latter.

Now suppose that, as we also believe, there is an inherent difference in kind between conceptual mental content and non-conceptual mental content.[18] Given this example and an indefinitely large number of similar examples in which propositions differ only in purely logical ways[19] it follows that *either* the phenomenal character of the subjective experience of these judgments strongly supervenes on conceptual content, *or else* it is determined non-conceptually—without being either wholly or solely determined by our conceptual capacities. We believe that it is determined non-conceptually, since it is quite possible for someone to love tigers in particular and hate felines in general, or hate tigers in particular and love felines in general. This, in turn, will determine the phenomenal character of those judgments for her, but *that* has nothing to do with their conceptual determinable-determinate relations. Whether phenomenal character is determined conceptually or non-conceptually however, it still follows that every mental state involving propositional intentionality$_{lo}$ has *some phenomenal character or another*, although not necessarily the same phenomenal character for each state involving the same propositional content, and therefore that the Aspectual Shape Thesis is correct, as a specification of the $P_{lo}I_{lo}$ Thesis.

(8) *First-order transitive consciousness*$_{lo}$.[20] First-order transitive consciousness$_{lo}$ is the unmediated, direct, and non-conceptual subjective awareness

[16] See Searle, *The Rediscovery of the Mind*, ch. 6; and Horgan and Tienson, "The Intentionality of Phenomenology and the Phenomenology of Intentionality."

[17] See Searle, *Rediscovery of the Mind*, ch. 7. [18] See Ch. 1 above, note 15.

[19] Even "I believe that the tiger is on the mat," "I do not believe that the tiger is on the mat," "I believe that the tiger is not on the mat" all feel *quite* different from one another, and yet they differ only in classical logical operations.

[20] See. e.g., Dretske, "Conscious Experience."

internal to a representational mental state—usually, but not necessarily, a perceptual state. For example, while working on your laptop computer and intensely focused on what you are writing, you also consciously feel your own body parts and limbs, in their relative positions, movements, and changing orientations, and thus subjectively experience one aspect of your own single point of view without any conscious meta-representation, concepts, or self-directed thoughts. This is pre-reflective consciousness$_{lo}$ or sensorimotor subjectivity. In a few paragraphs we will argue that pre-reflective consciousness$_{lo}$ or sensorimotor subjectivity is not only logically independent of, but also *presupposed by*, all self-conscious, self-reflective, and meta-representational subjective awareness.

(9) *Higher-order transitive* consciousness$_{lo}$.[21] Higher-order transitive consciousness$_{lo}$ is the relation between a first-order mental state and a distinct higher-order mental state (e.g., an introspection, a self-directed thought, or a higher-order desire for or against a certain first-order desire—and much more on this latter crucial notion in Chapters 3 to 5) that is either directly referred to, or descriptively referred to, but in any case is *about*, the first state. Or in other words, it is a *meta-representational* consciousness$_{lo}$. Higher-order transitive consciousness$_{lo}$ need not always involve propositional or conceptual thoughts. Some kinds of non-human animals, and also normal pre-linguistic human children, it seems, have conscious states that are complex in this way, yet do not involve propositional or conceptual thinking, and are merely desire-based or imaginational. But even at the level of rational human consciousness$_{lo}$, e.g., it seems entirely possible for a struggling alcoholic

(a) to intensely want to drink alcohol,

and also

(b) to intensely want not to want to drink alcohol,

without forming any thoughts or concepts about this at all. To be sure, he also has a large repertoire of other concepts and thoughts, some of which are sometimes directed to his alcoholism. All we are saying is that it is entirely possible that at that moment his mind might be focused entirely on

[21] See Botterill and Carruthers, *The Philosophy of Psychology*, ch. 9.

the desire-based emotive consciousness$_{lo}$ of a certain awful kind of thirst, and nothing else.

(10) *Immanent reflexivity or the immediate sense of self*.[22] Immanent reflexivity or the immediate sense of self is the first-order, direct, non-conceptual, non-propositional self-awareness of an essentially embodied mind, whether rational or non-rational, and it is manifest fundamentally via conscious desire-based emotions. Immanent reflexivity is not the same as *self-consciousness*$_{lo}$, which requires the animal's possession of a concept of itself, together with the ability to make judgments about itself and form beliefs about itself. Nor is immanent reflexivity the same as *self-reflection*, which requires, in addition to self-consciousness$_{lo}$, an ability of the animal to think about its own life as a whole.

Immanent reflexivity is inherently less structured than either self-consciousness$_{lo}$ or self-reflection, both of which are meta-representational states, and yet it is also presupposed by both. This in turn is because immanent reflexivity is intrinsically connected with the single egocentrically-centered standpoint that constitutes our essentially embodied occupation of actual space and time. This standpoint determines our representations of oriented directions in space (right, left, up, down, backwards, forwards, etc.) and thermodynamically irreversible directions in time (past, present, future), both for ourselves and also for objects co-embedded with us in that space and time. In a word, immanent reflexivity belongs essentially to pre-reflective consciousness$_{lo}$ or sensorimotor subjectivity; and sensorimotor subjectivity, as we have seen, is originally consciously given in primitive bodily awareness. We will see in Section 2.4 that primitive bodily awareness need not be either singly-focused or vivid. But even if it is non-focused, multiply-focused, or non-vivid it necessarily accompanies and expresses the finegrained pre-reflectively conscious sensorimotor control of our own living animal bodies in perception and intentional movement.

As we mentioned in Section 1.2, primitive bodily awareness must be carefully distinguished both from

(i) *bodily self-consciousness*$_{lo}$ or *bodily self-reflection*, which is the singly-focused, vivid, thoughtful awareness we have *of* our body (including its several parts, systems, organs, or processes)

[22] See Frankfurt, "Identification and Wholeheartedness," 160–3; Sartre, *The Transcendence of the Ego*; Searle, *Mind: A Brief Introduction*, 101; and Wider, *The Bodily Basis of Consciousness*.

and also

(ii) *the body image*, which is the explicit cognitive, affective, and practical mental model or pictorial map we have of our own bodies.

In this connection, Shaun Gallagher very usefully opposes the notion of a *body schema* to bodily self-consciousness$_{lo}$ and the body image alike:

I defined body image as a . . . system of perceptions, attitudes, beliefs, and dispositions pertaining to one's own body. It can be characterized as involving at least three aspects: body percept, body concept, and body affect. Body schema, in contrast, is a system of sensory-motor processes that constantly regulate posture and movement—processes that function without reflective awareness or the necessity of perceptual monitoring. Body schemas can also be thought of as a collection of sensory-motor interactions that individually define a specific movement or posture, including elementary (relatively defined) movements, such as the rotation of a wrist within a larger movement or the movement of hand to mouth.[23]

On our view, what Gallagher calls the body schema is the *intentional content*—including non-conceptual representational content, sensory-affective content, and practical-motile or "how-to-move" content—of primitive bodily awareness. On our view, this intentional content is *essentially* non-conceptual in that its semantic structure and psychological function are essentially different in nature from the structure and function of conceptual content. What is this essential difference? On the one hand, conceptual content determines our allocentric or third-personal and indirect *descriptions* of objects, and provides for objective, truth-evaluable, logically-governed, linguistically communicable information about objects with which subjects may not be directly acquainted. But by sharp contrast and on the other hand, non-conceptual content determines our egocentric or first-personal and direct *acquaintance* with objects and with ourselves in orientable space and thermodynamically irreversible time, and guides the accurate or inaccurate finegrained sensorimotor control of our body movements as we endeavor to uniquely locate and track worldly objects and ourselves, in order to carry out perceptual cognitions and basic intentional actions.[24] So the body schema is more cognitively and action-theoretically basic than conceptual content. It is crucial to remember, moroever, that according to

[23] Gallagher, *How the Body Shapes the Mind*, 37–8.
[24] See Hanna, "Kantian Non-Conceptualism."

the Deep Consciousness Thesis, although the non-conceptual intentional content of the body schema is relatively non-conscious in a certain sense, it is nevertheless necessarily also minimally and definitely occurrently conscious in a pre-reflectively conscious sensorimotor-subjective sense.

Here is another way of elaborating the same points. Nagel's canonical description of consciousness has it that

> an organism has conscious mental states if and only if there is something it is like to *be* that organism—something it is like *for* the organism.[25]

But it is also true that an organism has conscious states if and only if there is something it is like to be that organism *itself*—something it is like for the organism *itself*. A conscious organism, in feeling things, also *immediately and pre-reflectively feels its own situated living bodily presence and dynamic capacities for forward-flowing cognitive activity and intentional body movement, right here and now*. This immediate feeling, furthermore, is inherently emotive_d in nature. A conscious organism, in feeling things, and in caring about things, and in feeling its own bodily presence, and in feeling on the verge of moving its own body towards new adventures in cognition and intentional action in accordance with its desires, inherently also *cares about itself*. Its own virtually motile, situated, forward-flowing mental states *matter to itself*. This is its immediate, essentially embodied sense of self. And this immediate sense of self, it seems, cannot be removed from any minded animal without simply extinguishing its $\text{consciousness}_{lo}$.

To be sure, apart from philosophical intuitions based on neuro-phenomenological descriptions, knockdown proof of this thesis is difficult to find. But it does also seem to be well supported by empirical evidence from studies in cognitive ethology and human fetal development—e.g., the fact of quickening, or spontaneous fetal movement, which normally begins to happen between 13 to 18 weeks after conception, so in the second trimester—which strongly suggest that sentience somewhat outruns fully-constituted $\text{consciousness}_{lo}$ in both the non-human and human animal world. If so, then there are animals—e.g., insects, reptiles, normal human fetuses in the second trimester, etc.—that are capable of *feeling* the world in some ways, without also having an essentially embodied

[25] Nagel "What is it like to be a bat?," 166.

first-person awareness.[26] Then the real-time transition between mere sentience and conscious, intentional sentience in a living organism would be constituted by the emergence of a motile, egocentrically-centered and spatially oriented, thermodynamically irreversible, immanently reflexive, emotive$_d$, sensory, and non-conceptually representational animal mind from the psychic blur or cacaphony of mere sentience, together with a corresponding complete neurobiological embodiment. In normal humans this transition seems to occur between 22 and 26 weeks after conception, so roughly at the beginning of the third trimester.[27] In any case, it is empirically known that less than three months after this fundamental transition, normal neonates are actually able to imitate gestures and respond to faces, which is a basic form of empathic mirroring that surely entails the actual existence of a motile, situated, forward-flowing, immanently reflexive, emotive$_d$, sensory, and non-conceptually representational living organismic mind.[28]

As we have emphasized, there is a basic neurophenomenological distinction between the situated, forward-flowing, and immanent reflexive pre-reflectively conscious sensorimotor subjectivity of a consciousness$_{lo}$ (originally expressed as primitive bodily awareness, and having as its content a body schema) and a *body image*. The latter is an explicit cognitive, affective, and practical mental model or pictorial map of our own body, which is intimately connected with how we think and feel about ourselves, how we present ourselves to others, and how we plan our intentional movements. No doubt the generation of a fully explicit body image is closely connected with our encounters with mirrors and other reflective surfaces, although it probably begins to arise as soon as we are able to recognize that others are seeing *us*. But the fascinating phenomenon of *unilateral neglect*,[29] whereby stroke patients can fail to attend to or perceive one side of their body, even while still being able to make effective, skilled intentional body movements—e.g., being able to dress themselves—also provides compelling empirical evidence for our thesis that body schema and body image are sharply distinct. For in cases of unilateral neglect we clearly have a complete and uncompromised primitive bodily awareness and

[26] See, e.g., DeGrazia, *Taking Animals Seriously*; and Dennett, "Animal Consciousness: What Matters and Why."

[27] See *British Parliamentary Office of Science and Technology Notes* 94 (1997). URL = <http://www.parliament.uk/post/pn094.pdf>.The relevant study was done by Prof. Maria Fitzgerald of the Dept. of Anatomy and Developmental Biology at UCL in 1995.

[28] See Gallagher, *How the Body Shapes the Mind*, ch. 3.

[29] Ibid., 40.

correspondingly uncompromised body schema, along with an incomplete or compromised body image.

There is also significant empirical evidence for the converse phenomenon—a normally functioning body image along with a compromised primitive bodily awareness and body schema—in the strange case of Ian Waterman. The victim of a catastrophic illness at age 19, Waterman lost certain crucial aspects of his primitive sense of proprioception below his neck, although he did retain the capacity for normal proprioceptive experience above his neck, and especially in the facial area. Consequently, like Merleau-Ponty's unfortunate Schneider, Waterman has severe apraxia, i.e., an inability to perform simple body movements without great effort.[30] Indeed, because Waterman must *see* his own limbs in order to move them, he collapses in the dark. It would be a mistake, however, to conclude from this case that primitive bodily awareness and its body schema are not more fundamental than a body image. For, perhaps even more strangely, Waterman can skillfully drive an automobile, often without having to look at or specifically think about his limbs, and finds it immensely easier to drive three or four hundred miles, than to stop and fill up his car with gas.[31] It seems reasonable to conclude that the practice of driving somehow temporarily reinstates aspects of Waterman's primitive bodily awareness and body schema—perhaps by non-conceptually mapping them directly from his proprioceptive experience of his facial area to the rest of his body?—and temporarily partially restores his egocentrically-centered and spatially oriented, thermodynamically irreversible, and immanently reflexive pre-reflectively conscious sensorimotor subjectivity.

In rational animals like us, our primitive bodily awareness together with our conceptual and other discursive capacities yield a hybrid capacity for generating the uniquely self-locatory and self-tracking propositional representation *I am here now*. This representation describes the pre-reflective immanent reflexivity of sensorimotor subjectivity, and it conceptually expresses the core element of the body schema—to the extent that this is possible, that is, since the content of the body schema is itself

[30] See Gallagher, *How the Body Shapes the Mind*, 43–64; and Merleau-Ponty, *Phenomenology of Perception*, 103–47.

[31] Gallagher, *How the Body Shapes the Mind*, 58.

essentially non-conceptual.[32] More precisely, however, the propositional representation *I am here now* has three basic features:

(i) it is necessarily true (because it is true in every possible context of conscious framing or utterance),
(ii) it is immune to error through misidentification (because it holds even if the subject holds false beliefs or no beliefs at all about who, what, where, or when that subject itself actually is),[33]

and

(iii) it is presupposed by any *further* act of higher-order cognition, and especially any further act of self-consciousness$_{lo}$ or self-reflection.[34]

Generally then, we hold that *all* first-order transitive consciousness$_{lo}$ (i.e., consciousness$_{lo}$ of type (8)) intrinsically includes immanent reflexivity or the immediate sense of self, along with a motile, situated, forward-flowing pre-reflectively conscious sensorimotor subjectivity, its primitive bodily awareness, and its body schema as foundational factors—even when those experiences *also* include conceptual, propositional, meta-representational, self-conscious or self-reflective, logically rational, or practically rational elements. So on our view, all rationality in minded animals is essentially embodied too.[35]

2.2 Eight Structures of Consciousness$_{lo}$

If consciousness$_{lo}$ is the subjective experience of a suitably neurobiologically complex living organism, then precisely what *inherent forms* does this subjective experience take? This is the same as the question: "What are

[32] To say that content is essentially non-conceptual is to say that it has an inherently different kind of semantic structure and psychological function from the structure and function of conceptual content, and thus that a concept cannot express *everything* that a non-conceptual content expresses. But this does not mean that a concept cannot be used to express *anything* that a non-conceptual content expresses. So it is possible to form *parasitical* concepts of essentially non-conceptual contents, e.g., the concept of haecceity or thisness. In this way, the proposition *I am here now* is thus a proposition logically built up out of the parasitical concepts of the ego, hereness, and nowness.

[33] See Evans, *Varieties of Reference*, 179–91, and ch.7.

[34] See, e.g., Bermúdez, *The Paradox of Self-Consciousnes*; Campbell, *Past, Space, and Self*; and Hurley, *Consciousness in Action*, ch. 2.

[35] See Hanna, *Rationality and Logic*, ch. 7.

the intrinsic structures of consciousness$_{lo}$?'' By way of an answer to this question, and by (it seems to us) appealing implicitly to what we are calling ''neurophenomenological analysis,'' Searle has recently provided two lists of the structures of consciousness$_{lo}$—one in *The Rediscovery of the Mind*, and the other in *Mind: A Brief Introduction*. These lists provide a very useful starting point for the next phase of our discussion. Here is the list from *Rediscovery*:

(1) Finite modalities
(2) Unity
(3) Intentionality
(4) Subjective feeling
(5) The connection between consciousness and intentionality
(6) The figure-ground, gestalt structure of conscious experience
(7) The aspect of familiarity
(8) Overflow
(9) The center and the periphery
(10) Boundary conditions
(11) Mood
(12) The pleasure/unpleasure dimension.[36]

And here is the list from *Mind*:

(1) Qualitativeness
(2) Subjectivity
(3) Unity
(4) Intentionality
(5) Mood
(6) The distinction between the center and the periphery
(7) Pleasure/unpleasure
(8) Situatedness
(9) Active and passive consciousness
(10) The sense of self.[37]

One thing should, however, be noticed about Searle's two lists before we go on to the details of our own account. In *Rediscovery* he says that they are ''gross, structural features of normal, everyday consciousness'' and that

[36] Searle, *Rediscovery of the Mind*, 128–41. [37] Searle, *Mind*, 93–101.

"often the argument I will use for identifying a feature is the absence of the feature in pathological forms."[38] And in *Mind* he says that they are "central features of human, and presumably animal, consciousness."[39] These formulations are somewhat ambiguous as to the precise intended force and scope of his analysis. Nevertheless, we think that it is philosophically fruitful to construe Searle as asserting the substantive thesis that *at least some* of the structures yielded by a neurophenomenological analysis are *inherent or intrinsic (i.e., necessary, internal) structures of every* consciousness$_{lo}$. But in any case, with the specific aim of testing that substantive thesis, we will present an eight-entry neurophenomenological list of our own—in some ways similar to Searle's two lists, but also in at least three ways importantly different.

Eight Structures of Consciousness$_{lo}$

(1) *Affectivity*: phenomenal character and conative affectivity.
(2) *Egocentricity*: immanent reflexivity, as originally expressed by primitive bodily awareness.
(3) *Spatiality*: orientability and balanceability in proprioception.
(4) *Temporality*: spontaneity, motility, and kinaesthesia in proprioception.
(5) *Embodiment*: the immediate sense of a unique continuing essential embodiment.
(6) *Intentionality$_{lo}$*: directedness and aboutness.
(7) *Focus*: single-focus/multi-focused/non-focused.
(8) *Intensity*: degrees of experience.

The first and most obvious important difference between our list and Searle's is our strong emphasis on essential embodiment. The second is our similarly strong emphasis on the conatively affective and desire-based emotive character of consciousness$_{lo}$. And the third important difference is our equally strong emphasis on the spatiotemporal—and more specifically, on the situated (egocentrically centered, and spatially oriented) and thermodynamically irreversible (forward-flowing)—features of our mental lives. In the next two sections we will unpack the eight entries on our list, and attempt to demonstrate that each entry is an intrinsic structure of every consciousness$_{lo}$.

[38] Searle, *Rediscovery of the Mind*, 128. [39] Searle, *Mind*, 93.

2.3 Affectivity, Egocentricity, Spatiality, and Temporality

(1) *Affectivity: phenomenal character and conative affectivity*. Phenomenal character is one specific element of what we called the "experiential" aspect of consciousness$_{lo}$. Phenomenal character is also, in at least one important respect, the same as what Nagel calls the "subjective character" of consciousness, what Searle calls the "subjective feeling" or "qualitativeness" of consciousness, and what Chalmers calls the "qualitative feel" of consciousness. For states of consciousness$_{lo}$ to have phenomenal character, is also for them to have some or another irreducibly sensory property necessarily instantiated in that state; and this is true of "subjective character," "subjective feeling," and "qualitative feel" as well.

Nevertheless it is an extremely important and open question whether such irreducibly and necessarily instantiated sensory properties are inherent or intrinsic *non-relational* features of states of consciousness$_{lo}$ and furthermore whether (the instances of) these properties are, as Dennett puts it, "ineffable," "private," and "directly or immediately apprehensible."[40] In other words, it is an important and open question whether phenomenal characters are *qualia* in the classical sense or not. Our view, shared with Dennett, but for reasons somewhat different from his,[41] is that *there are no such things as qualia in the classical sense*. So like Dennett we are *qualia eliminativists*.

But we also strongly agree with the later Wittgenstein that although a phenomenal character is not a quale, and thus "not a *something*," it is not a *nothing* either.[42] We hold that the subjective experience of creatures minded like us is primitively real and physically irreducible. So we are at once *qualia eliminativists* and also *freaks about consciousness*$_{lo}$. More precisely, then, on our view phenomenal characters are

(i) intrinsic and also structural properties, i.e., necessary, internal, relational properties that are inherently bound up with the

[40] See Dennett, "Quining Qualia," 229.

[41] Dennett's reasons are mainly behaviorist and verificationist, whereas ours are based on neurophenomenological analysis and philosophical intuitions about the metaphysics of the mind–body relation.

[42] Wittgenstein, *Philosophical Investigations*, 102e, §304.

spatiotemporal neurobiological dynamics of our living organismic bodies,

(ii) effable, i.e., communicable to another essentially embodied subject who is suitably egocentrically positioned in orientable space and thermodynamically irreversible time, even if not conceptually describable to that subject,

(iii) shareable, at the very least, by means of empathic mirroring of intentional body movements—i.e., as movement-types, although not as tokens of those movement-types,

(iv) directly apprehensible, i.e., available without further cognitive mediation to either pre-reflectively conscious sensorimotor subjectivity or self-conscious, self-reflective introspective subjectivity,

and

(v) fallible, i.e., open to introspective misinterpretation

features of all conscious states like ours. Or in other words, and now focusing on the paradigm case of conscious pain-experience,[43] we think

(i*) that conscious pain necessarily happens *in and through* our entire living animal bodies, which we aptly capture in natural language by saying that I am *in* pain,[44]

(ii*) that I can meaningfully convey the character of my pain to you, whether non-linguistically or linguistically,

(iii*) that insofar as you are able empathically to mirror, by emulation or simulation, the essentially embodied conditions of my experience, it is thereby possible for someone else to feel the same *type* of pain that I am feeling, although not possible for someone else to "live my pain" or to feel just the same pain-*token*,

(iv*) that I am directly aware of my own pain through either primitive bodily awareness or else self-conscious or self-reflective introspection,

[43] See, e.g., Grahek, *Feeling Pain and Being in Pain*.

[44] Obviously pain is also normally localized by reference to the specific bodily causal source of inner or outer damage or disruption to vital organs, systems, or processes to which conscious pain is the response. So it would be more accurate, although also obviously more of a mouthful, when I have just smashed my thumb with a hammer (and finished cursing and hopping up and down), to say "I'm in terrible pain because of damage to my hand" than to say "My hand hurts!"

and

(v*) that even though I feel my own pain just about as directly and intimately as anything can ever be felt, I may still occasionally be wrong in my conceptual and self-conscious or self-reflective characterizations of it—e.g., I may occasionally mislabel my pain as pleasure, etc.[45]

In any case we are going to proceed as if *qualia in the classical sense* do not exist but *phenomenal characters* do exist, and focus now on the further question of whether all conscious states of minded animals must have phenomenal character in the sense just described.

Our answer to that latter question is *yes*, but with the crucial qualification that such characters are originally displayed in *primitive bodily awareness*, and are fundamentally *conatively affective*, hence emotive$_d$. On our view, as we mentioned earlier, consciousness$_{lo}$ necessarily involves the possibility of conative affect or desire-based emotion. But we also hold the even stronger view that consciousness$_{lo}$ is fundamentally experienced by us as emotion$_d$, and moreover that we share this fundamental experience of conative affectivity with all other creatures minded like us, whether human or non-human. If so, then minded animals are essentially animals with a *sensibility*, not merely animals with *sensations*. An animal is minded just insofar as it has *felt needs*, whether real or merely imagined, and insofar as those felt needs constitute "what-it-is-like-to-be" that creature. To say that it has felt needs is to say that it *desires things*, whether positively as a *desire-for* or liking, or negatively as a *desire-against* or disliking. Or in other words, and again: *I desire, therefore I am*.

The neutral state of *desire-suspension* is of course also possible—e.g., when I am hesitating between different options for caring, or when I am temporarily sated and satisfied, or when I am temporarily lethargic or stunned, or when I am dreamlessly asleep—but only as a variant on the normal states of desire-for and desire-against. Even in desire-suspension, I am still poised to desire, or at the very least I am still capable of desiring. A creature that not only does not desire anything, but also *cannot* desire things in *any* sense of that term, and thereby lacks any sort of *sensibility* (or as we

[45] Is sneezing painful or pleasurable? Is scratching an itch painful or pleasurable? Is coughing painful or pleasurable? And so on. It is very hard to say.

shall put it in Chapter 5, lacks any ability to *care*), we think, is simply *not* a minded animal. As a matter of bare conceptual or logical possibility there could be a minded being without any desires or a sensibility—perhaps some sort of alien (Mr Spock? Data?), angel, ghost, or other disembodied spirit—but this creature would not have an inner life in the sense that *we* have an inner life. If this line of thinking is correct, then to the extent that a minded animal exists, necessarily it is always either occurrently desiring for or against, on the verge of desiring, or at least capable of desiring. Otherwise put, for a creature with a consciousness$_{lo}$, everything that is experienced or experienceable *matters* in one way or another. Why, for example, would a creature with a consciousness like ours ever *be attentive* to anything, if it could not desire things and lacked any sort of sensibility? In any case, if this thesis is correct, then it will also directly support the universality and necessity of Searle's features of "mood" and "pleasure/unpleasure."

Moreover the conative affectivity of consciousness$_{lo}$ need not necessarily depend on the *external* senses. Instead, as we have said, on our view all conative affectivity depends ultimately on primitive bodily awareness. As a matter of naïve phenomenology, our pre-reflectively conscious desire-based emotions do certainly seem to arise from the middle of our body and emanate upwards, downwards, and outwards towards our heads, lower extremities, and external sense organs. But, perhaps not surprisingly, this phenomenology also has a direct neurobiological correlate, as recent empirical work on the "enteric brain" and recent philosophical work on emotions as "gut reactions" strongly suggests.[46] So there seem to be good neurophenomenological grounds for closely linking the pre-reflectively conscious desire-based emotions and primitive bodily awareness.

In any case, it seems perfectly possible to think coherently, even within the domain of strong metaphysical possibility, of cases in which someone is anaesthetized for taste, touch, and smell, and is also both profoundly blind and deaf—and so, roughly, is a Helen Keller under various sorts of local anaesthetic—yet remains intensely conscious via her emotions$_d$ and at least some of her bodily senses. But on the other hand, it seems perfectly *impossible* to think coherently of cases in which someone has a consciousness$_{lo}$ and yet is also stripped of *all* her felt needs, emotions$_d$ and *all* modes of primitive bodily awareness. So it seems clearly and distinctly true that an

[46] See ch. 1 above, notes 26–7.

Emotional Zero, or Hollow Man—a creature necessarily devoid of felt needs, emotions$_d$ and all modes of primitive bodily awareness—necessarily would not have a *consciousness*$_{lo}$ (see Section 5.4). And even if, as a matter of bare conceptual and logical possibility, an Emotional Zero or Hollow Man could be *in some sense* conscious (and this seems essentially the same as the thought that an angel, ghost, or other disembodied spirit could be *in some sense* conscious) we do not have the slightest neurophenomenological grip on what *that* sort of consciousness *would be like*. So this would be simply a case in which classical philosophical methods are surreptitiously outrunning the methods of phenomenology and cognitive neuroscience, the other two necessary elements in the methodological triangle constituting an adequate Science of Minded Animals or Minds$_{lo}$ (see Section 1.1).

Indeed, even the subjective experience of *numbness* is itself a special kind of essentially embodied emotion$_d$. Here one can remember or imagine what it is like to be under the influence of an oral anaesthetic when having one's tooth filled or removed, or to be going under or coming out from under the influence of a general anaesthetic in surgery, or what it is like to be very drunk, very disappointed, very shocked, or very surprised. It is most certainly *not* the case that you feel nothing at all. In feeling numb, you directly feel the actual presence or pressure of external things, but without the more or less insistently intense sense of pleasure or pain that such presence or pressure usually brings. So when you are sitting in the dentist's chair, lying on the operating table, totally hammered, totally bummed out, or totally blown away, the external things are still all there for you, and they still all matter to you *somewhat*—only, for the time being, they just do not matter to you so *damned* much.

(2) *Egocentricity: immanent reflexivity, as originally expressed by primitive bodily awareness*. As we mentioned above, to say that all consciousness$_{lo}$ is "subjective" is to say that it is *egocentrically-centered* and *immanently reflexive*. For a conscious state to be egocentrically-centered is for that state to have an "inner" source-point, as opposed to an "outer" derivation or dispersal, and also for the creature in that state to be able to relate everything that is experienced to this inner source-point. This psychic relating can take either an inner→outer direction, or outer→inner direction, which presumably trace the corresponding efferent and afferent directions of neurobiological dynamics in the sensorimotor nervous system. Moreover,

an analogy between subjectivity and Newton's universal gravitational forces of attraction and repulsion is quite illuminating here. Roughly speaking, the *I* of our subjectivity is the inner relatum, in the *centrifugal* sense, of everything else whatsoever in the experienced world (repulsive force, inner→outer direction, efferent neurobiological dynamics); and the *me* of subjectivity is the inner relatum, in the *centripetal* sense, of everything else whatsoever in the experienced world (attractive force, outer→inner direction, afferent neurobiological dynamics). Otherwise put, the *I* is the "subjective subject" of consciousness$_{lo}$ and the *me* is the "subjective object" of consciousness$_{lo}$.

By a subtle but important contrast, for a state of consciousness$_{lo}$ to be immanently reflexive, as we noted above, is for it to include an immediate sense of self, or for it to be directly aware of itself in a wholly first-order sense—that is, to be folded back upon itself, to be directly attentive to itself, and care directly about itself, without any division or opacity between itself and the content of its own experience. Indeed, this wholly first-order lamination of a conscious state like ours back upon itself is so intimate that G. E. Moore called it the "transparency" of consciousness.[47] Early Wittgenstein called the same fact "the microcosm."[48] Sartre called it "the pre-reflective *cogito*."[49] The basic idea shared by Moore, Wittgenstein, and Sartre alike is that in immanent reflexivity the content of consciousness$_{lo}$ entirely fills up my conscious awareness, leaving as a remainder only the primitive solipsistic fact of *my owning*, here and now, the whole world of which I am conscious. As Wittgenstein puts it: "The world is *my* world," or even more centripetally expressed, "I am my world." By contrast, Sartre expresses the very same idea in a centrifugal way as "the transcendence of the ego."

The combination of egocentric centering, spatial orientation, forward-flow, and immanent reflexivity in an essentially embodied mind, as we have also argued, is the primitive bodily awareness of a pre-reflectively conscious sensorimotor subjectivity, which is sharply distinct from both bodily self-consciousness$_{lo}$, or bodily self-reflection, and body image alike. We have further argued that the inherent or intrinsic relational phenomenal characters of primitive bodily awareness are themselves forms of desire-based

[47] Moore, "The Refutation of Idealism."
[48] Wittgenstein, *Tractatus Logico-Philosophicus*, props. 5.62–5.6331, p. 151.
[49] Sartre, *Being and Nothingness*, 9–17.

emotion, and that the essentially non-conceptual intentional content of primitive bodily awareness is the body schema, which guides pre-reflective sensorimotor operations in cognition and intentional action alike, and whose core element is conceptually expressible (to the extent that this is possible) as the propositional representation *I am here now*. All that remains, then, is to argue explicitly for the necessity and universality of primitive bodily awareness and pre-reflectively conscious sensorimotor subjectivity in consciousness$_{lo}$. But this follows directly from our argument for the necessity and universality of conative affectivity. So—for a change!—we can simply cite that argument now as a sufficient reason, and move on.

(3) *Spatiality: orientability and balanceability in proprioception*. To the extent that consciousness$_{lo}$ is necessarily and completely neurobiologically embodied—essentially embodied—it also seems to be necessarily spatialized. In having subjective experiences, my experience necessarily occurs *here*, wherever that might happen to be. But I need not be able to know where I am. I could be asleep in my bed in Colorado or in Massuchusetts, but falsely think that I am running frantically (and also, very frustratingly, as if through clear molasses) to catch a train somewhere in England, as I dreamt last night, in fact. Or I could be the man, famously described by Russell, who dreamt he was making a speech in Parliament, then awoke, and *was* making a speech in Parliament. Or I could be actually awake and just confused or mistaken about my actual whereabouts—something quite easy for those of us not naturally gifted with the powers of a Global Positioning System. But this does not entail that I am not subjectively experiencing myself as *uniquely located*, or *uniquely positioned*. As we noted already, the spatiotemporal uniquely locatory proposition *I am here now*, and thus also the spatial unique locatory proposition *I am here*, are necessary truths precisely because they have their foundation in the nature of essentially embodied consciousness, as expressed in primitive bodily awareness, via its body schema.[50]

Moreover, this necessary spatiality of essentially embodied experience carries with it an intrinsic topology and dynamics. Essentially embodied consciousness$_{lo}$ is also necessarily *orientable* and *balanceable* via its proprioceptive capacitie. To the extent that I am aware of myself as *here*, I am

[50] See, e.g., Campbell, *Reference and Consciousness*.

also aware of myself as facing left, facing right, right-side up, recumbent, upside-down, or tipped sideways. I feel the difference between my right side and my left side, between the upper and lower bounds of my body, and between my front and my back. And furthermore, I always place myself, as relatively balanced or poised, in some orientation or another.

The commonplace subjective experiences of disorientation or of loss of balance are not counterexamples to these claims. For me to feel dizzy or lost is not for me to be aware of myself *non-orientably,* as if I were somehow taking a walk along the surface of a Möbius strip, or *without any sense of balance whatsoever*, as if I were somehow no longer a prisoner of gravity. The subjective experiences of disorientation and unbalance are merely limiting cases—or, as the Scholastics might say, "privations"—of the intrinsic neurophenomenological structures of orientation and balance in proprioception, and not their denials.

(4) *Temporality: spontaneity, motility, and kinaesthesia in proprioception.* Many of the same points go, mutatis mutandis, for the necessary temporality of consciousness$_{lo}$. In having an essentially embodied consciousness, I necessarily experience my conscious states as occurring *now*, even if I happen to be self-consciously or self-reflectively confused or mistaken about what the time or date actually is. Insofar as I am conscious in the sense of sensorimotor subjectivity, I thereby necessarily uniquely locate myself in time, just as I necessarily also uniquely locate myself in space.

The subjective experience of temporality also carries within itself a subordinate set of intrinsic neurophenomenological structures. Necessarily I embed myself in time in direct relation to earlier time and later time (the "B series" described by the Pythonesquely-named Cambridge philosopher John McTaggart Ellis McTaggart), and also to past, present, and future time (McTaggart's "A series"). This includes

(i) my pre-reflectively conscious, sensorimotor-subjective, non-conceptual experience of what *just happened* in the immediate past, or *short-term memory*,

in relation to

(ii) my pre-reflectively conscious sensorimotor-subjective, non-conceptual experience of what is *now just happening* in the present, or

> *first impressions*, like the pointed prow of a moving ship leaving a continuous wake of immediately past experiences trailing behind it in a recursive sequence of acts of short-term memory,

and also in relation to

(iii) my pre-reflectively conscious, sensorimotor-subjective, non-conceptual experience of what is *just about to happen* in the immediate future, or *anticipation*.

This tripartite neurophenomenological structure, which can be quickly sketched as follows—

Short-term memory	←	*First impressions*	→	*Anticipation*
of		*of*		*of*
Immediate past	←	**Present moment**	→	**Immediate future**

—closely corresponds to what Husserl called "retention," "primal impression," and "protention."[51] Furthermore, as Shawn Gallagher has correctly noted, my subjective experience of the just-happened in short-term memory, or retention, is particularly closely associated with my conscious, intentional sense of having *ownership* of my past subjective experiences together with the past states of my living animal body, and thus with my personal identity over time. By contrast, my subjective experience of the just-about-to-happen in anticipation, or protention, is particularly closely associated with my conscious intentional sense of *being ready to choose or do things myself or have them merely happen to me*, and thus with my sense of *agency*.[52]

Furthermore, the basic structure of temporal subjective experience bears a striking analogy to my spatial subjective experiences of *left, egocentric center*, and *right*. What we mean is this: My retentive subjective experiences of the just-happened and my protentive subjective experiences of the just-about-to-happen both necessarily require my egocentrically-centered immanently reflexive primal impression of the now-just-happening in order to disambiguate possible or actual qualitatively identical counterpart events—e.g., two successive ringings of the same bell. Analogously, my two (virtually, and for all ordinary intents and purposes, indiscriminably) qualitatively

[51] Husserl, *Phenomenology of Internal Time Consciousness*.
[52] Gallagher, *How the Body Shapes the Mind*, ch. 8.

identical hands are such that all their parts and properties correspond one-to-one, but cannot be made to coincide by a rigid translation within the same spatiotemporal framework, so need to be disambiguated by my egocentrically-centered immanently reflexive embedded standpoint in a globally orientable space.

Consider now the strange phenomenology of Billy Pilgrim in Kurt Vonnegut's blackly comic 1969 sci-fi novel *Slaughterhouse Five*, who felt himself to be coming "unstuck in time"; or the equally strange experiences of Jimmy the sailor—the man without long-term memory—as fascinatingly described by Oliver Sacks in *Awakenings*; or the equally strange phenomenology of the unfortunate amnesiac protagonist in the thoughtful and disturbing 2000 Christopher Nolan film, *Memento*; or similar pathologies associated with schizophrenia. *None* of these is, in fact, a counterexample to our thesis that conscious intentionality has an inherent temporal structure. To feel as if you were coming unstuck in time, or to feel as if you literally had no past or no future—while of course highly disturbing and highly disruptive of your consciousness of your own diachronic personal identity (and thereby of your sense of ownership) and also of your consciousness of your own intentional agency (and thereby of your sense of agency)—is not thereby to be *atemporally* conscious.

On the contrary, atemporal consciousness, while logically possible, is what we might call *a non-logical or strong a priori metaphysical impossibility or oxymoron*. Even in the strange or pathological cases mentioned above, there is some brute, immediate sense of temporal passage, even if it is highly bizarre. So consciousness$_{lo}$ is essentially temporal. If there is a God, and if God has a consciousness outside of time and space, then like the Emotional Zero or Hollow Man, this seems barely conceptually or logically possible in itself, but non-logically or strongly metaphysically a priori impossible for a consciousness$_{lo}$. So at the very least, God's consciousness would not be a consciousness on which we could ever get *any* sort of neurophenomenological handle. Interestingly enough, this thesis coincides with the traditional claim that the divine mind is "ineffable."

But here is where the important analogies between the necessary spatiality and the necessary temporality of consciousness$_{lo}$ stop; for the necessary temporality of essentially embodied consciousness$_{lo}$ seems also to include two entirely unique factors. The first is its inherent *spontaneity*, or the

subject's immediate sense of each new experiential moment within its own continuing minded animal life as being

(a) unprecedented (i.e., something that never happened in just this way before),
(b) underdetermined by what preceded it (i.e., having no nomologically sufficient antecedent cause),
(c) recursive (i.e., infinitely successively cognitively constructible by repeated applications of the same form of momentary present experience),
(d) creative (i.e., it would not have happened in just this way if the subject herself had not been directly involved in its etiology),

and finally

(e) self-guided (i.e., subjectively controlled and purposive).[53]

And the second is its inherent *kinaesthesia* and *motility*, or the essentially embodied subject's immediate sense of moving her limbs or changing her body position on her own through her intentional agency (whereby the bodily movements or changes are experienced as something that is *up to her*), or at least of being *able* to do so, or of being moved or changed by something else (whereby the bodily movements or changes are experienced as something that merely *happens to her*). Or more generally, essentially embodied consciousness$_{lo}$, by way of its temporality, subjectively experiences itself as inherently *causal-dynamic*.

It should be noted that the intrinsic spontaneity of a consciousness$_{lo}$ holds even if the experiential contents of the newly arriving moments are not terribly *exciting*—e.g., while waiting in an airport for your delayed flight to begin boarding, which for some strange reason is actually more intensely dull than watching paint dry. In its simplest form, the spontaneity of a consciousness$_{lo}$ is nothing more and nothing less than the immediate sense of time's asymmetric continuous forward flow, or the immediate sense of the intentionality of *temporal passage*.[54]

It should also be noted that the intrinsic proprioceptive kinaesthesia and motility of a consciousness$_{lo}$ holds even if you are conscious while

[53] See Hanna and Thompson, "Neurophenomenology and the Spontaneity of Consciousness."
[54] This is similar to what Bergson called *la durée* or "duration." See Bergson, *An Introduction to Metaphysics*.

your whole body is paralyzed, or stuck immovably in place like an insect in amber. The terrifying neuropathological phenomenon of *locked-in syndrome*—that is, consciousness$_{lo}$ together with temporary virtually complete bodily paralysis—is presumably an example of this. (Temporary bodily paralysis, even if it is a virtually complete paralysis, is not in and of itself either death or the permanent disruption of our vital organs, systems, or processes, and so is not a counterexample to essential embodiment.) Surely what makes shut-in syndrome, or even the very thought of it, so terrifying is precisely the fact that consciousness$_{lo}$ is inherently *kinaesthetic* and *motile*, and thereby either moving or on the verge of intentional body movement. You intensely desire to move your own body and be like everyone and everything else in this dynamic natural world, getting on with your life, but you simply *cannot*—everyone and everything is passing you by, and there is simply nothing you can do about it. Only a necessarily kinaesthetic and motile consciousness, i.e., a consciousness$_{lo}$, would *ever* care about this. So in its simplest form, the necessary kinaesthesia and motility of consciousness$_{lo}$ is nothing more and nothing less than the vivid sense of *things happening* and *being on the move* both inside you and outside you—just like those March of Time newsreels from the 1930s and 40s neatly parodied by Orson Welles in the opening frames of *Citizen Kane*.

2.4 Embodiment, Intentionality$_{lo}$, Focus, and Intensity

(5) *Embodiment: the immediate sense of a unique continuing essential embodiment.* It has often been said, or anyhow implied—e.g., by Kant, Nagel, and Searle—that consciousness$_{lo}$ is necessarily a *unity*, or *unified*, in the sense that a subject with a consciousness$_{lo}$ necessarily brings (or at least necessarily possesses the ability to bring) all of its current phenomenal or representational contents into a single phenomenal field. But this seems to be false. The well-attested empirical phenomena of divided attention, peripheral attention, subliminal attention, dissociated information processing in cases of neo-commissurotomy, and cognitive priming by means of masking, all seem to show just the opposite. These are all cases in which a subject with a consciousness$_{lo}$ possesses an unbroken continuity of a single conscious life, or a diachronic psychological personal identity, in the phenomenologically

robust sense that she feels herself to have, and to be actually living, a fully coherent *life of her own from its very beginning up to just now*, yet she precisely *does not* or even *cannot* bring all of her current phenomenal or representational contents of consciousness into a single phenomenal field.

So what accounts for the (normally[55]) unbroken, phenomenologically robust diachronic continuity of a single conscious, intentional personal life in a minded animal, even in cases of divided attention, peripheral attention, and so on? Our view is that it is the minded animal's immediate sense, given in and through her primitive bodily awareness, of her *unique continuing essential embodiment*. In other words, a minded animal, as a pre-reflectively conscious sensorimotor subject, experiences herself as a coherent and individual essentially embodied mind if and only if she immediately feels herself to be standing in a direct, causally efficacious, and sensibly intimate connection to all the actual and potential movements, vital systems, vital organs, vital processes, and overall condition of her own living animal body. If this neurophenomenological thesis is correct, then the sense of one's unique continuing essential embodiment in pre-reflectively conscious sensorimotor subjectivity is not only more fundamental than the unity of consciousness$_{lo}$ in the high-powered Kantian sense described in the Transcendental Analytic section of the *Critique of Pure Reason*, but also a necessary structure of consciousness$_{lo}$ itself.

This structure of consciousness$_{lo}$ is obviously closely related to *the body schema* that is the essentially non-conceptual intentional content of the primitive bodily awareness which originally expresses our pre-reflectively conscious sensorimotor subjectivity. In every moment of the mental life of a minded animal, the animal is immediately, non-conceptually, and pre-reflectively aware, with lesser or greater accuracy, of the shape, position, boundaries, extent, actual movement, potential movability, and balanceability of her own body. For example, normally I do not have to tell myself *where* my two hands are, or *whether* they are the same as or different from one another, *whether* they are the same as or different from the table sitting in front of me on which my hands are resting, or *how* to lift my hands in the air, or *when* to begin to form an intentional grasping movement with my right hand as I reach forward to pick up a pint glass of beer. Finegrained and hyper-finegrained sensorimotor activity normally operates along with,

[55] With exceptions for temporary periods of unconsciousness or coma. See note 14 above.

but can also operate independently of, both bodily self-consciousness or self-reflection and the body image. But the presence of primitive bodily awareness and body schemata are necessary and sufficient conditions of the conscious essentially embodied self, the pre-reflectively conscious sensorimotor subject.

Body schemata are not static structures. Sometimes they can be temporarily disrupted, as when my hand "goes to sleep." Sometimes they can be temporarily distorted, as in the well-known "phantom limb" illusions, in which body schema and body image are confused by a self-conscious or self-reflective amputee. Sometimes, catastrophically, body schemata can be permanently disrupted, as in Ian Waterman's partial proprioceptive apraxia. And sometimes they can be temporarily causal-dynamically extended to bodily prosthetics, as is shown by the experiences of amputees and by V. S. Ramachandran's fascinating experiments with rubber arms,[56] not to mention the more commonplace experience of skilled stickhandling hockey players, who can feel the puck literally *at* the tip of their stick, hence *at* their dynamically extended fingertips.

Nevertheless the disruptions, distortions, and extensions of specific body schemata in particular, and of the sense of a unique continuing essential embodiment more generally, are not counterexamples to their existence—rather they are only the more extreme adventures in the unique continuing first-personal life of an essentially embodied mind. It seems highly intuitive that the step-by-step, more or less gradual, and finally total loss of *every* sense of a unique continuing essential embodiment would also entail the extinguishing of a consciousness$_{lo}$. And presumably, that is what-it-is-like for a creature with a consciousness$_{lo}$ to die a natural death.

(6) *Intentionality*$_{lo}$. As we have said, intentionality$_{lo}$ is either the capacity of a consciousness$_{lo}$ to direct itself at or towards objects, actions, locations, events, other conscious creatures or itself (i.e., intentional targets), or else the fact that a conscious creature like us has mental states that are "about" something or another, by virtue of their content. These two types of intentionality$_{lo}$ are also called "referential intentionality" and "content intentionality" respectively.[57]

[56] See Blakesee and Ramachandran, *Phantoms in the Brain*.
[57] See, e.g., Kim, *Philosophy of Mind*, ch. 1.

Sometimes intentionality$_{lo}$ has been identified with the fact that minded animals can have mental states with *conceptual* content or *propositional* content. But this identification seems false. If, e.g., sense perception can have *essentially non-conceptual content*—that is, representational content whose structure and function are essentially different from the structure and function of conceptual content[58]—and all conceptual content and propositional content presuppose essentially non-conceptual content, then it is wrong simply to identify intentionality with conceptual content or propositional content. As we mentioned above, it seems reasonable to identify conceptual contents with *descriptive* representations, whose minimally necessary function it is to categorize, classify, discriminate, and identify things, and provide allocentric or third-personal indirect objective and linguistically communicable information about them, without our necessarily having to be egocentrically directly acquainted with those things. Then it also seems to be the case that only the essentially non-conceptual spatiotemporal content of perception, as given in the body schemata of primitive bodily awareness, can adequately discriminate *incongruent qualitative material counterparts* in orientable space, like the right and left hands, or *incongruent earlier and later occurrences of the qualitatively same event*, like the sound of a bell ringing, in thermodynamically irreversible time.[59]

If this is correct, then it entails that the structure of body-schematic essentially non-conceptual content necessarily involves egocentrically-centered representations of the intrinsic relational topological and temporal properties of creatures and things embedded in a global orientable space and thermodynamically irreversible time. The special function of such content is to guide or mediate the finegrained and hyper-finegrained pre-reflectively conscious sensorimotor control of the living body in cognition and intentional action. Such bodily control can be seen, e.g., in our causally efficacious, fluid, highly responsive pre-reflective awareness of how to turn ourselves when we want to look at, hear, or smell something, and of where to place our feet when we walk so as to avoid perceived obstacles. Thus intentionality$_{lo}$ *can* involve states with conceptual content or propositional content, but it *need not either exclusively or necessarily* do so.

[58] See note 18 above. [59] See note 24 above.

Furthermore, it is arguable that the contents of some conscious, intentional states in minded animals are altogether concept-less and propositional-less. This is sometimes called "the autonomy thesis."[60] For example, just as, by virtue of primitive bodily awareness and body schemata, and thereby just by virtue of representing myself essentially non-conceptually, I do not have to tell myself where my hand is or whether it is the same as or different from other things, I also do not have to *think* about my hand in order to be able to use it skillfully. Indeed, the whole process of learning to play a musical instrument (say, a piano) seems to be based on the presupposition that the movements and positioning of one's hands, fingers, and the rest of one's body can occur in such a way that those positionings are *not* also self-consciously or self-reflectively represented by means of concepts or propositions, and that they can occur altogether independently of any of the relevant concepts and propositions. This is precisely because the generation and presence of these representations would slow down or even interfere with the causally efficacious, fluid, highly responsive bodily performance itself.[61] Very similar points go for athletics, dancing, and riding motorcycles—not to mention something as commonplace as an infant's learning how to walk (although this in fact turns out to be a highly complex dynamic cognitive and volitional process[62]), or an adult's learning *again* how to walk, after a serious leg injury.[63]

Is consciousness$_{lo}$ necessarily intentional? Brentano[64] and other philosophers of mind have thought that it is. But it has been argued, e.g., by Searle, that there are both actual and possible cases in which a subject is occurrently conscious but not occurrently attentively directed to any determinate objects.[65] This seems to be true of certain very good moods (e.g., free-floating happiness, or joy) or very bad moods (e,g., free-floating *Angst*, or depression), Buddhist meditative consciousness, and the everyday experience of spacing or zoning out. And if normal dreamless sleep has a special phenomenal character—and as we have said, we think it does, analogously to the way that the experience of white noise has a special

[60] See Gunther (ed.), *Essays on Non-Conceptual Content*, part IV.
[61] See, e.g., Sudnow, *Ways of the Hand*.
[62] See Thelen and Smith, *A Dynamic Systems Approach to the Development of Cognition and Action*.
[63] See Sacks, *A Leg to Stand On*.
[64] See Brentano, *Psychology from an Empirical Standpoint*, p. 88.
[65] See Searle, *Rediscovery of the Mind*, 130; and Searle, *Mind*, 139.

auditory phenomenal character—then that would also count as a case of occurrent consciousness like ours without occurrent single-focused, vivid intentionality of objects. But, in any case, it seems clearly to be possible to be in an occurrent state of consciousness like ours, which is always a consciousness-*with* and *in-and-through* the living body, without also at the same time being occurrently conscious-*of* any object in particular. Therefore occurrent consciousness like ours does not seem to entail occurrent single-focused, vivid intentionality of objects.

A similar point goes, conversely, for the connection between occurrent intentionality$_{lo}$ and occurrent consciousness$_{lo}$: occurrent intentionality$_{lo}$ does not seem to entail an occurrent single-focused, vivid consciousness$_{lo}$. It seems clearly possible for me to be in a relatively non-conscious state but still be directed to objects, actions, locations, events, or myself. For example, I can look and point at things while sleepwalking, or while suffering an absence automatism,[66] and also I can be peripherally or subliminally aware of objects, actions, events, locations, or myself—e.g., of the sound of an air-conditioner, or the hum of the lights, or Muzak, or the air temperature, or the humidity, or the feel of my clothing against my body, etc.—while entirely focusing my attention on something else.

But if the Deep Consciousness Thesis we asserted and defended in Section 1.2 is correct, then all of these "marginal" forms of intentionality must nevertheless still include pre-reflectively conscious sensorimotor subjectivity and primitive bodily awareness, even if this occurrent consciousness is neither single-focused, nor vivid, nor self-conscious, nor self-reflective, nor directed to objects. And that point correspondingly raises a deeper and more difficult question: Is it possible for a creature like us to be occurrently conscious while not manifesting *some* sort of occurrent intentionality, even if it is neither single-focused, nor vivid, nor self-conscious nor self-reflective, nor directed to objects?

The answer to this deeper and more difficult question, it seems to us, is: *No, it is not possible, hence consciousness$_{lo}$ necessarily also always manifests occurrent intentionality$_{lo}$.* If we are correct that to have a consciousness$_{lo}$ it is necessarily to be a situated, forward-flowing living organism of a suitable level of neurobiological complexity, and with a capacity for desire-based

[66] See, e.g., Damasio, *The Feeling of What Happens,* 6–7, 95–101, and 122.

emotion or emotion$_d$, then since all desire is directed, and since the *targets* of intentional directedness can certainly be other than just *objects*, it follows that all occurrent consciousness$_{lo}$ necessarily includes some sort of occurrent intentionality$_{lo}$. Furthermore, we cannot think of any actual or possible counterexamples to this thesis. The supposed actual and possible cases of occurrent consciousness$_{lo}$ without occurrent intentionality$_{lo}$ are invariably cases in which the subject is able temporarily to slip into, or able to put herself temporarily into, a non-self-conscious, pre-reflectively conscious state that lacks a single focus or any relatively high degree of vividness, or whose intentional target is simply not any determinate object but instead an action, location, event, or oneself.

Now occurrent conscious intentionality$_{lo}$, even non-self-conscious and pre-reflective intentionality$_{lo}$, can be multi-focused, non-focused, or relatively unvivid, as the everyday examples of divided attention and moving around while feeling very sleepy show. And intentionality$_{lo}$ can be directed to many intentional targets (e.g., events, locations, actions, and oneself) that are sharply distinct from mere objects. So it seems to us highly intuitive that the very idea of an essentially embodied mind or minded animal with an occurrent consciousness$_{lo}$ whose inner life consisted entirely and intrinsically of *total whiteout*—somewhat like the visual effect of cutting a ping-pong ball in half, filling the two half-spheres with gauze, then putting them over your eyes, but now completely generalized over all of a subject's experiences—is again barely conceivable or logically possible, but non-logically or strongly a priori metaphysically impossible. Indeed, the *actual* result of the ping-pong ball experiment on creatures like us is *the Ganzfeld Effect*. The Ganzfeld Effect is the fact that our visual systems soon simply *shut down* due to lack of input, and we become temporarily blind—as, e.g., in snowblindness. So it seems reasonable to conclude that the result of introducing total whiteout into a consciousness$_{lo}$ would be the *total shut-down* of that consciousness$_{lo}$. Thus the conscious life of a desiring creature, or a minded animal, clearly and distinctly seems to be necessarily a life in which *things are always happening*, even if those things are not always terribly exciting or noisy. A totally whiteouted, non-intentional sort of conscious existence could not be any sort of *real* inner life like ours at all. Some of these points will also carry over directly to our discussions of the seventh and eighth structures of consciousness$_{lo}$.

(7) *Focus: single-focus/multi-focused/non-focused*. The phenomena of attention and inattention are obviously central facts about consciousness$_{lo}$. So too Searle's distinction between the center and the periphery of conscious states obviously picks out something centrally important about consciousness$_{lo}$. So too the foreground/background structure described by Gestalt psychology is similarly centrally important. But we think that neither attention/inattention, nor center/periphery, nor foreground/background is *in itself* sufficiently deep or sufficiently general to capture an arguably strictly universal and necessary relational feature of consciousness$_{lo}$.

One problem is that all of these distinctions are *binary*. But it seems to us that a more basic relational feature at play in consciousness$_{lo}$ would have to be *triadic*. There are also three good neurophenomenological reasons for this. First, a conscious subject's having

(i) *divided attention* (e.g., as between reading the newspaper and simultaneously listening to the radio),

is clearly structurally distinct from both

(ii) *singular attention* (e.g., just reading the newspaper)

and also

(iii) *inattention* (e.g., trying to read the newspaper and failing to do so, while thinking about something else).

Therefore divided conscious attention is triadic and not binary. Second, a conscious subject's having

(i) *multiple centers* of conscious content (e.g., stereophonic perception of music via the two earbuds of your iPod),

is clearly structurally distinct both from having

(ii) *a single center* of conscious content (e.g., monaural music-perception via just one of the earbuds of your iPod)

and also from having

(iii) *non-focused* conscious experience—where this could be either
(iiia) *surrounded* conscious experience (e.g., listening to music via a 1970s style quadrophonic speaker system)

or

(iiib) *peripheral* conscious experience (e.g., monitoring the sound of Muzak while grocery shopping).

Therefore the centering of conscious attention is also triadic and not binary. Third and finally, a conscious subject's having

(i) *multiple foregrounds* within conscious content (e.g., the simultaneous auditory experience of the sound of a trumpet alongside the visual experience of the trumpet player in a spotlight)

is clearly structurally distinct both from having

(ii) *a single foreground* within conscious content (e.g., just the sound of a trumpet alone in a darkened room),

and also from having

(iii) *a non-foregrounded* conscious experience—where this could be either
(iiia) *an all-foregrounded* conscious experience (e.g., the experience of standing at the end of a runway as a Boeing 747 takes off directly overhead),

or

(iiib) *an all-backgrounded* conscious experience (e.g., the experience of one's surroundings receding as one falls asleep).

Therefore the foregrounding and backgrounding of conscious experience is also triadic and not binary.

A second problem, at least with the candidacies of center/periphery and foreground/background for arguably achieving necessity and universality with respect to consciousness$_{lo}$, is that both are obviously *spatial* structures. But while it seems that all consciousness$_{lo}$ has intrinsic spatiality, including every conscious *vehicle* of content—e.g., linguistic mental imagery—it also seems that not every vehicle or content of consciousness$_{lo}$ is intrinsically structured by spatiality *alone*. Indeed, on the contrary it seems that all vehicles of content and most contents of consciousness$_{lo}$ are intrinsically structured by space and time taken *together*. Moreover, an intrinsically spatiotemporal consciousness$_{lo}$, if rational, can think of a *non*-spatiotemporal intentional object—e.g., a classical logical truth.[67] Thus a more basic

[67] See, e.g., Hanna, *Rationality and Logic*, ch. 6.

consciousness-structure of the same general sort as center-periphery and foreground/background should, at the very least, be *neutral* as between spatial structure and temporal structure, and if possible, neutral as between a spatiotemporal and non-spatiotemporal—e.g., logical—structure in the *intentional targets* of consciousness$_{lo}$, even if (as we hold) consciousness$_{lo}$ *itself* is necessarily spatiotemporal.

A third and final problem, this time with the attention/inattention dyad, is that it is heavily weighted towards the subjective aspect of consciousness$_{lo}$ and not towards its experiential aspect. What we mean is that attention or inattention is something that the egocentrically-centered *subject* does. By contrast, the *content* of conscious experience is itself neither attentive nor inattentive. Nevertheless, it seems that a properly universal and necessary neurophenomenological structure should be neutral as between the two basic subjective (egocentric) and experiential (contentful) aspects of consciousness$_{lo}$, in the sense that it applies without special bias, or *equally*, to both.

So our overall critical conclusion from these considerations is that consciousness$_{lo}$ inherently contains a *triadic, spatiotemporal or non-spatiotemporal*, and *equally subjective and experiential* structure that captures all the phenomena of the attention/inattention, center/periphery, and foreground/background distinctions, but is not *restricted* to any of these. For this reason, we propose that every consciousness$_{lo}$ is necessarily such as to implement a triadic structure such that our subjective experiences are either

(i) single focus,
(ii) multi-focused,

or

(iii) non-focused.

This triadic structure is abstract enough to hold for all the spatiotemporal or non-spatiotemporal phenomena of a consciousness$_{lo}$, and it also applies equally to the subjective and experiential aspects of our consciousness.

(8) *Intensity: degrees of experience*. In the *Treatise* and first *Enquiry* Hume claims that every sensory experience and every passion has some degree of "force and vivacity." Building on that Humean idea, in the Anticipations of Perception section of the *Critique of Pure Reason*, Kant claims that perceptual

content is continuously divisible into definite degrees of intensity, within definite kinds. This Humean–Kantian point seems to generalize smoothly to the mental life of any consciousness$_{lo}$. For it seems to us that necessarily the mental life of a consciousness$_{lo}$ is always more or less *vivid* to some degree, no matter how we determine and measure that vividness. More precisely then, we are saying along with Hume and Kant that necessarily every phenomenal character of a consciousness$_{lo}$ has some or another definite degree of intensity within some or another definite kind of phenomenal character. If we are correct about this, then the subjective experience of a color, e.g., always would be delivered to us at some definite levels of brightness and hue within a color-kind, and the subjective experience of sound always would be delivered to us at some definite levels of volume and pitch within a definite sound-kind—and so-on, mutatis mutandis, for all the external senses, modes of primitive bodily awareness, and desire-based emotions.

Correspondingly, temporal experience also includes a constant and regular variation in degrees of intensity as natural time goes by. For example, suppose I suddenly clap my hands. My subjective experience of that clapping includes a set of fairly vivid tactile, proprioceptive, auditory, and visual phenomenal characters.There is a maximal intensity of that total complex experiential content in the now-just-happening present moment as experienced in first impressions, or Husserlian primal impression, but even as that clapping experience occurs, it also inevitably gradually fades in intensity as it is experienced in short-term memory, or Husserlian retention, as it proceeds into the just-happened past, like the trailing wake of a ship.

Precisely how we draw the divisions, and how we self-consciously or self-reflectively determine and measure the determinate, measurable degree-units of intensity within the divisions, is doubtless constrained by human neurobiology, and also to some extent relativized to human interests and needs, both personal and social. But while there is bound to be a certain amount of contextual variation and epistemic vagueness, this is perfectly consistent with the existence of precise boundaries in reality. In any case, the necessary presence of degrees of intensity within specific kinds of phenomenal character, whether definite or vague, and whether in external sense perception, temporal experience, or primitive bodily awareness, seems neurophenomenologically very obvious.

Indeed, we even hold that the thesis of the necessary presence of degrees of intensity within kinds of phenomenal character is equally true of *dreamless sleep*, which seems very phenomenologically different from dreamful sleep, being passed out, or being otherwise unconscious. Otherwise, as we mentioned before, what would be the point—beyond mere politeness or social convention and the obvious intention of finding out how someone feels upon waking—of asking someone how he or she slept last night? Sleeping well, surely, is not *just* the fact that you do not wake up many times and do not have nightmares, but also that the *salient phenomenal character* of your dreamless sleep is *pleasant and serene*. Drunken sleep, by sharp contrast, even though it is quite deep, often without dreams, and usually without nightmares, has an unpleasant and tense salient phenomenal character. Obviously this is largely due to the fact that it typically involves strained breathing, dehydration, and other neurobiological anomalies due to the residual presence of alcohol in one's vital organs, systems, and processes. So one's primitive bodily awareness in drunken sleep directly conveys the minor disruption of one's vital processes as an uncomfortable extended subjective experience.

If we are correct, then there is a *neurophenomenology of sleep*, whether dreaming sleep or dreamless sleep. On the other hand, however, there is no neurophenomenology of *being* passed out or *being* unconscious, but rather only a neurophenomenology of *passing* out or *becoming* unconscious. As one passes out or becomes unconscious—say, during a fainting fit or seizure—the vividness of consciousness$_{lo}$ is continuously reduced through a series of degrees of intensity within some or another kind of phenomenal character to a vanishing point (like one of those old TV sets from the 50s), and then altogether extinguished as it reaches the lower bound of the series of degrees within that kind. If what we said above was correct, then the experience of falling asleep is just a transition from one kind of phenomenal character to another—from waking consciousness$_{lo}$ to sleeping consciousness$_{lo}$. But the continuous reduction of the vividness of waking consciousness$_{lo}$ through a series of degrees of intensity, until the threshold of the transition into sleeping consciousness$_{lo}$ has been passed, is surely *structurally analogous* to the neurophenomenology of passing out or becoming unconscious.

Similarly the neurophenomenology of passing out or becoming unconscious is, no doubt, structurally analogous to the neurophenomenology of

natural death in creatures with consciousness$_{lo}$. Shakespeare made this point about the structural analogy between falling asleep and our experience of natural death very movingly:

> To die—to sleep—no more; and by a sleep to say we end the heartache, and the thousand natural shocks that flesh is heir to. 'Tis a consummation devoutly to be wished. To die—to sleep—perchance to dream: ay, there's the rub! For in that sleep of death what dreams may come when we have shuttled off this mortal coil, must give us pause.[68]

On the other hand, and with what Searle wrily calls "his usual gift for catchy phrases," Kant dubbed this same phenomenon "elanguescence."[69] So Kant, for all his intellectual and literary virtues, was no Shakespeare. But whatever we call it, it does seem to be a genuine intrinsic structure of every consciousness$_{lo}$.

In any case, if we are correct about all of this, then, leaving aside the epistemic issue of precise division, determination, and measurement, then necessarily every consciousness$_{lo}$ has a graduated intensity within some or another kind of phenomenal character. Yet this is rarely noticed by other contemporary philosophers of mind, even by self-described "consciousness freaks," and one wonders why.

One reason for this, it seems, is the needlessly narrow focus of many consciousness freaks on *external sensory* experience as opposed to *desire-based emotive* experience. (In turn, this avoidance of the conatively affective and emotive domain may also reflect a deeper philosophical bias—see Chapter 5). So one reason why we are especially able to recognize this neurophenomenological structure has to do with our two-part claim that consciousness$_{lo}$ is fundamentally manifest as desire-based emotion, which in turn is originally given in primitive bodily awareness. For whether or not one agrees with our claim about the omnipresence of degrees of intensity within kinds of phenomenal character in consciousness$_{lo}$, it surely is self-evident, at least, that your own desire-based emotions necessarily always have some or another degree of intensity within some or another kind of desire-based emotion, and that this intensity is inherently connected with your necessary and complete neurobiological embodiment.

[68] Shakespeare, *Hamlet*, act III, scene I, lines 60–8.
[69] See Searle, *Rationality in Action*, 77; and Kant, *Critique of Pure Reason*, 449, B414.

Indeed, in order to recognize this fact clearly and distinctly, you need only ask yourself the following question: "On a scale of 1–10, how happy (sad, lively, tired, bored, interested, hungry, thirsty, anxious, relaxed, etc.) do I feel right now?" Everyone capable of understanding that question will be able to answer it somehow at any time of the day or night, and also be able to compare her answers at those different times. In fact, this is a commonplace of the widespread contemporary practice of *emotional counselling*. So it seems that our thesis that necessarily all consciousness$_{lo}$ has graduated intensity within some or another kind of phenomenal character is at least *conditionally* true, on the hypothesis that every essentially embodied mind, or consciousness$_{lo}$, intrinsically involves conative affect or emotion$_{d}$ and primitive bodily awareness.

We began Chapter 1 with the primitive fact of a consciousness$_{lo}$. In turn, our neurophenomenological analysis of this primitive fact in Chapters 1 and 2 has revealed the basic kinds and basic contours of our essentially embodied conscious and intentional lives, as minded animals living together with other minded animals in the natural world. In the next six chapters, we will put this neurophenomenology fully into *action*.

3

Essentially Embodied Agency I: Actions, Causes, and Reasons

> Let us not forget this: When I "raise my arm," my arm goes up. And the problem arises: What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?
>
> Ludwig Wittgenstein[1]

> The act of raising the arm is a complex event, constituted out of a causally linked pair, the trying and the arm rising, which are . . . "made for one another."
>
> Brian O'Shaughnessy[2]

3.0 Introduction

For healthy ordinary people in ordinary situations, arm-raising seems like the simplest thing in the world. But philosophers are not ordinary people. And philosophers of *action*, in particular, have correctly observed that there is an obvious and categorical difference between

(i) my deliberately raising my arm to wave to a friend,

and

(ii) my arm's uncontrollably rising in a Dr Strangelove-like spasm,

even if the overt body movements are indiscriminable. Indeed, as we mentioned in the Introduction, *the problem of action* is how to give an adequate account of the categorical difference between the things we intentionally do, or *intentional actions*, and the things that just happen to us, or *mere bodily events*.

[1] Wittgenstein, *Philosophical Investigations*, 161e, §621.
[2] O'Shaughnessy, "Trying (as the Mental 'Pineal Gland')," 70.

It turns out that it is not at all easy to say what the categorical difference between an intentional *arm-raising* and a mere *arm-rising* really consists in. Now the *we* and *us* in action are conscious, intentional, motile, egocentrically-centered and spatially oriented, thermodynamically irreversible, suitably neurobiologically complex living organisms—minded animals. So our goal in this chapter and the next is to say as precisely as possible what, for minded animals, the categorical difference between intentional actions and mere bodily events really is.

More specifically, this chapter explores the neurophenomenological, conceptual, and metaphysical connections between intentional actions, causes, and reasons. In Section 2.1, we spell out and criticize classical causal theories of action in a general way. But as we point out, non-causal theories of action are also unacceptable, since they implausibly substitute teleological reasons-explanations for the basic causal facts that actually bring about intentional actions. Our response to this dilemma is to develop a *non*-classical but still *causal* theory of action—what we call the *Essentially Embodied Agency Theory*. In Sections 3.2 to 3.5 we motivate this theory of action by focusing specifically on Davidson's classical causal theory and then developing four fundamental worries about it.

Here is a quick Coming Attractions preview of the Essentially Embodied Agency Theory of action. As we see it, every classical causal theory inserts a vitiating metaphysical or temporal gap between antecedent mental causes and consequent body movements. Now for our purposes, a "body movement" in a creature minded like us is an integrated series of dynamic endogenous events involving both "covert" neurobiological processes as well as "overt" behavioral processes normally arising from these processes. Or more precisely put, body movements are of two importantly different but closely related kinds:

(1) *covert* body movements are internal *neurobiological* processes that occur between the vital organs and the muscle tissue/skin interface, and that normally begin prior to overt body movements,

and

(2) *overt* body movements are external *behavioral* processes, normally arising from and accompanying neurobiological processes, that begin at the muscle tissue/skin interface and engage with the outer world.

On our view, what distinguishes intentional actions (e.g., arm-raisings) from mere bodily happenings (e.g., arm-risings) are the causally efficacious operations of a certain mental activity of an essentially embodied intentional agent throughout the entire time in which some of that agent's body movements *covertly* arise neurobiologically and then display themselves *overtly* and behaviorally. *All* of these body movements are the agent's *intentional* body movements. So *if* we assume that the Essential Embodiment Thesis is true, then *since* the causally efficacious mental activity of the agent is necessarily and completely neurobiologically embodied, and *since* that mental activity is also synchronous with the complete two-part process that encompasses both the relevant covert neurobiological process as well as the relevant overt behavioral process, *it follows necessarily that* there are no vitiating metaphysical or temporal gaps whatsoever between the mental activity of the agent and her intentional body movements. Or more briefly and imagistically put, the conscious intentionality$_{lo}$ of the agent and her intentional body movements fit together as seamlessly as W.B. Yeats's *dancer* and her *dance*:

> O body swayed to music, O brightening glance,
> How can we know the dancer from the dance?[3]

3.1 Classical Causal Theories of Action, and Beyond

Classical causal theories of action say that what distinguishes intentional actions from mere bodily events is essentially a difference in the causal origin of those bodily events. The bodily events are the same in both cases, and intentional actions are supposed to be the events brought about by antecedent causes in some categorically different manner. There are three classical causal theories of action.

On the *agent-causal* view, the causal antecedence is metaphysical but not temporal. A pure mental substance, or "agent-cause," which exists outside the series of natural events, and thereby is naturally *un*determined by those events, is supposed to bring about the relevant bodily event in an incompatibilistically free way.[4]

[3] Yeats, "Among School Children," verse viii, 245.

[4] See, e.g., Chisholm, "Human Freedom and the Self"; Clarke, "Agent Causation and Event Causation in the Production of Free Action"; and O'Connor, *Persons and Causes*.

On the *volitional-causal* view, by contrast, the causal antecedence is temporal. A mental event of conscious willing at one time is supposed to cause a later bodily event.[5]

By another contrast, on the *Davidsonian* causalist view, which has dominated the philosophy of action since the 1970s, the causal antecedence is both metaphysical *and* temporal. *Reasons* are supposed to be *causes*, and an action is a physical event e_a that is caused by a "primary reason." A primary reason, in turn, is a psychological pair consisting of a belief and a desire that together "rationalize" or teleologically explain e_a.[6] The mental properties of this psychological pair are strongly supervenient in an "anomalous" way—i.e., in accordance with Davidson's *Principle of the Anomalism of the Mental*, which says that there are no strict deterministic *psycho*-physical laws—on fundamental physical facts about some earlier physical event e_c, that in turn naturally causes e_a under strict deterministic *physical* laws. And the psychological pair, considered as a single *mental event* of self-conscious deliberative intention—call it "*Me*"—is numerically identical with e_c.[7]

Following Arthur Danto, we accept the classical distinction between

(i) basic acts,

and

(ii) non-basic acts.[8]

Basic acts occur whenever an intentional agent performs a particular sequence of intentional body movements and no other acts are performed, and non-basic acts are acts that involve some basic acts but are not identical to those basic acts. Thus, e.g., someone waves to a friend (non-basic act) by raising her arm (basic act). Our analysis of action will focus primarily on basic acts.

It seems clear to us that all classical causal theories of action—whether agent-causal, volitional-causal, or Davidsonian—ultimately alienate the conscious intentionality$_{lo}$ of the agent from the intentional body movements

[5] See, e.g., O'Shaughnessy, "Trying (as the Mental 'Pineal Gland')." In Ch. 4 we will argue that a volitional, trying-based account of action coheres perfectly with a *non*-classical causal theory of action.

[6] See Davidson, "Actions, Reasons, and Causes," 27.

[7] See Davidson, "Mental Events"; and Davidson, "Thinking Causes."

[8] See Danto *Analytic Philosophy of Action* 31.

that are supposed to be the immediate effects of conscious intentionality$_{lo}$ in basic acts. This is because all such accounts imply that whenever a basic act occurs, there is *some* sort of vitiating gap—whether metaphysical, temporal, or both—between conscious intentionality$_{lo}$ and its immediate bodily effects.

In agent-causal theories, the alienation is the result of a vitiating *substance-dualist gap* into which a mysterious Cartesian *causal interaction* must be inserted between transcendent mental substances and fundamentally physical events.

In volitional-causal theories, the alienation is the result of a vitiating *temporal gap* into which *deviant causal chains* can always be inserted between earlier mental events and later physical events involving body movements, thereby making those later body movements unintentional. For example, someone tries to raise her paralyzed arm and fails, but her simultaneous frustrated desire to move her arm accidentally triggers a nearby brain scanner, which accidentally triggers someone else's Blackberry, which accidentally connects with the digital control system of a tractor-beam ray gun on Mars, which accidentally zaps her arm perfectly into place above her head. This is what you might call causal deviance *with an altitude*.

And in Davidson's theory, in addition to the same vitiating temporal gap that is always open to deviant causal chains, the alienation is also the result of a vitiating *property-dualist-without-substance-dualist gap* into which *upwards determination relations* must be inserted between the causally efficacious properties of the physical event e_c and the strongly supervenient mental properties of the conscious intentional mental event *Me* that is numerically identical with e_c. Because the mental properties of *Me* are strictly upwardly determined by the causally efficacious properties of e_c, and because *Me* is numerically identical with e_c, the mental properties of *Me* do not have any causal efficacy apart from the properties of its underlying physical base, and must be causally inert or *epiphenomenal*. This is very unfortunate for *Me*. Otherwise put, the agent is causally superfluous.

But whatever the origins of the classical causalist gap, the result is always the same. The *causal autonomy* of the agent's conscious intentionality$_{lo}$ undermines the *causal efficacy* of her conscious intentionality$_{lo}$, and thus intentional action is not adequately explained.

Some action theorists, finding serious problems with the Davidsonian account, have advanced non-causal, teleological theories of action.[9] Carl Ginet, for example, suggests that for a teleological reasons-explanation to be true, the action must be accompanied by an intention with the right sort of content: The subject intended of that action that by it she would *A*.[10] His view suggests that the mere presence in the agent of an intention about her *A*-ing is sufficient for that intention's being explanatory of her action (i.e., her *A*-ing). However, it seems clear that our body movements might merely accidentally coincide with our desires or intentions, without those movements' being explained by these desires or intentions. If, to use the example described above, an intentional agent's arm-movements were actually and accidentally caused by a tractor-beam ray gun on Mars, it would be highly implausible to say that the agent's frustrated desire to move her arm *explained* those movements.[11] Indeed, unless desires and intentions play a direct causal role in the production of body movements, it seems that the conscious, intentional animal does not act as a genuine agent, but is instead under the control of outside forces. Also, without appealing to causal facts, it is very difficult to make sense of what makes it true that an agent acted in pursuit of one goal rather than another, or for some reason rather than another. The natural answer to the question of what makes it true that an agent acted in pursuit of one goal rather than another or for some reason rather than another is that the mental event or process that explains the particular action in question is the one that "figures suitably in the etiology of the action or of [the subject's] completing that action."[12] Therefore a causal theory makes much better prima facie sense of what it means to perform body movements and act for the sake of some goal, than any non-causalist view does. But the $64, 000.00 question is: Can one be a causalist about action without also being a *classical* causalist?

Our answer is: *Yes*, but only if the classical causalist gaps have been closed up tight from the start. So our response to the dilemma that both classical causal theories and non-causal theories are manifestly inadequate is to present a *non*-classical but still *causal* theory of action that is designed to rule out the various vitiating gaps associated with classical causal theories

[9] For some non-causal theories of action, see Anscombe, *Intention*; Sehon, "An Argument Against the Causal Theory of Action Explanation,"; and Sehon, "Connectionism and the Causal Theory of Action Explanation."

[10] Ginet, *On Action*. [11] Mele, *Motivation and Agency*, 46. [12] Ibid., 40.

and also to account for the full range of action-types that ordinary agents perform. This is the Essentially Embodied Agency Theory of action. The two basic features of the Essentially Embodied Agency Theory are

(i) that it explicates intentional body movement in terms of synchronous trying and its active guidance of the agent's motile living body,

and

(ii) that it explicates trying and its active guidance in terms of desire-based emotions or emotions$_d$. (N.B. we are using "trying and its active guidance" as a *singular term*.)

The present chapter and Chapter 4 concentrate on the first element. For us, trying and its active guidance, like all mental states, events, or processes of a consciousness$_{lo}$, is necessarily and completely neurobiologically embodied. This enables us to solve the problem of action by holding that intentional actions are not *mere bodily events* of *any* kind, but instead are *essentially embodied events* of a *certain* kind, namely those inherently involving a synchronous trying and its active guidance of the agent's own motile, living animal body. For when we combine the Essential Embodiment Thesis with the thesis that intentional actions are brought about by synchronous trying and its active guidance, it directly follows that the conscious intentionality$_{lo}$ of the agent and her intentional body movements in basic acts are not only temporally in sync, but also are metaphysically connected as closely as possible, *short of strict identity*, by virtue of their *intrinsically reciprocal*, or inherently two-way, non-logical or strong metaphysical a priori necessitation.

These points about non-identity and intrinsic reciprocity require particular emphasis. The relation of strict identity, whether construed as either

(i) *type*-identity (identity of properties or universals),
(ii) *token*-identity (identity of individuals or particulars, a.k.a. "numerical identity"),

or

(iii) *essential* identity (identity of kinds),

can always be construed as a *reductive* relation, whenever it is represented by means of two terms that differently name or describe the same thing.

For given the right theoretical backdrop (say, a scientific essentialist theory of physical microstructure), then one term t_1 (say, 'water') can then be used to name or describe something that is "nothing but" or "nothing over and above" what is named or described by the other term (say, 'H_2O'), by virtue of the strict identity relation between them (in this case, essential microphysical identity). But even if a *non*-identity between the referents of the terms can be somehow demonstrated, then a certain kind of fairly robust reduction is still possible if one or both of the properties, individuals, or kinds is *logically strongly supervenient* on the other. This is because the supervening individual or properties can then be held to be "fully determined" or "fully fixed" by the properties of its corresponding physical supervenience base, whether by logical or analytic necessity alone (a priori physicalism or reductionism), or by logical or analytic necessity together with causal laws (a posteriori physicalism or reductionism). Supervenience, in turn, can be either

(i) one-way (*asymmetric*), as, e.g., in the "upwards" determination of temperature properties on the mean molecular motion of the particles comprising the material bodies, gases, or liquids that have temperature,

or else

(ii) two-way (*bilateral*), as, e.g., in the "back-and-forth" mutual determination of force and the product of mass and acceleration according to the classical Newtonian equation $F = ma$.

Furthermore it seems that even if there are some *other* ways in which reduction is possible (e.g., finegrained logically necessary equivalence of properties, or perhaps logically necessary co-extension), all of these will also entail either strict identity or at least logical strong supervenience. So the disjunction consisting of *either strict identity, or finegrained logically necessary equivalence, or logically necessary coextension, or one-way logical strong supervenience, or bilateral logical strong supervenience* would seem to be necessary and sufficient for physicalist reduction. If this is correct, then since the relation we are positing between the conscious intentionality of the agent is at once a relation of *non-logically necessary co-extension* (and thus is neither finegrained logically necessary equivalence nor logically necessary co-extension), *non-identity* (and thus is neither type-identity, token-identity, nor essential

identity), and *reciprocal intrinsicness* (and thus is neither non-asymmetric nor only extrinsically symmetric, as in one-way or two-way logical strong supervenience), it is therefore *non*-reductive, no matter how modally airtight that relation may otherwise be.

More precisely then, on our view fundamental mental properties (involving conscious intentionality$_{lo}$) on the one hand, and certain corresponding fundamental physical properties on the other, are

(i) non-logically or strongly metaphysically a priori necessarily co-extensive in all and only living organisms of a suitable level of neurobiological complexity,
(ii) non-identical,

and

(iii) reciprocally intrinsic properties of those very organisms.

Or in other words, those organisms are *essentially mental-and-physical*. As we will see in Chapters 6, 7, and 8, when looked at from a metaphysical standpoint, this non-logically or strongly metaphysically a priori necessary co-extension together with reciprocal intrinsicness, but without identity, means that the corresponding fundamental mental properties and fundamental physical properties in animals of a suitable level of neurobiological complexity are at once *mutually irreducible* and yet also *fused*.

And that thesis, in turn, allows us to answer Wittgenstein's deep question, "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?," correctly and directly. The correct, direct answer is: That would be like trying to subtract the dance from the dancer—but you *cannot* know the dancer from the dance!

More precisely, we think that it is non-logically or strongly metaphysically a priori *impossible* to subtract the fact that my arm goes up from the fact that I raise my arm. In my intentional action of raising my arm, the two facts become non-logically or strongly metaphysically a priori *non-detachable*. So, given the essential embodiment of minded animals, there simply cannot be and thus never is a metaphysical or temporal gap between our conscious intentionality$_{lo}$ and our intentional body movements.

When looked at from a philosophy-of-action standpoint, however, the very same non-detachable but non-reductive relation of mental-physical property fusion is also interpreted by us as a synchronous causal relation

of *effective desiring, willing*, or *trying-and-its-active-guidance*. As we said in Chapter 1, the essential embodiment of consciousness$_{lo}$ entails the Essentially Embodied *Cogito: I desire, therefore I am*. In Chapters 3–4, we will argue that effective desiring is the foundation of willing, and also that willing is the same as trying and its active guidance of body movements. If so, then the Essentially Embodied *Cogito* is also in effect a causal-intentional *Cogito: I effectively desire, therefore I am simultaneously intentionally moving my body*. Under favorable endogenous and exogenous conditions, just by effectively desiring to move my living body, I thereby also simultaneously will my own intentional body movements, which is the same as the fact that my trying and its active guidance simultaneously *self-determine* my intentional body movements. In this way, both the causal autonomy *and* the causal efficacy of the conscious intentionality$_{lo}$ of the agent are jointly secured.

Much contemporary philosophy of action begins with the more or less explicit assumption that it is always possible to construct a broadly Humean, belief-desire based, instrumental reasons-explanation for action.[13] For example, Jane's desire for a doughnut and her belief that a doughnut is in the cupboard explain why she walks over to the cupboard. This standard account of action entails that intentional actions are done because the agent has a certain belief-desire pair that explains the action by giving an instrumental teleological rationalization of it.

A version of this broadly Humean story has been very influentially defended by Davidson, whose central claim, as we have said, is that a primary reason for an action is its cause. As we have also said, a primary reason consists of a belief-desire pair that instrumentally teleologically rationalizes a certain physical event e_a, the action. So whenever someone does something for a reason, he has some sort of pro-attitude toward actions of a certain kind and believes that his action is of that kind. To know a primary reason for action is to know the intention with which the action was done, and also to know how that action is coherent with certain traits, both long and short termed, of a rational agent. Davidson points out that a person can have a reason for action, perform the act in question, and yet not act on the basis of this reason. If a reason is to explain an action,

[13] See, e.g., Hume, *Treatise of Human Nature*, book II, part III; and Blackburn, *Ruling Passions: A Theory of Practical Reasoning*.

the agent must perform the action *because* he had that reason. Davidson regards such instrumental teleological rationalizations of behavior as a type of ordinary causal explanation, so that the 'because' in "Mary went to the fridge because she wanted a beer" is a causal 'because'.

We do not wish to dispute in any way Davidson's claim that actions are suited for causal explanation. But we do also hold that an adequate causal explanation of an action cannot be secured just by placing an action in the context of an instrumental teleological rationalizing reason that strongly supervenes in an anomalous way on a causally efficacious physical event, precisely because this does not show how the conscious intentionality$_{lo}$ of the agent can have *both* causal autonomy and causal efficacy. Instead, in order to secure the fusion of causal autonomy and causal efficacy in agency, we must tell a metaphysically plausible story about the causality of the action *from the inside*—where 'from the inside' means both *from the first-person standpoint* of the agent as a conscious, intentional (and in some cases, also a rational human) animal and equally also *from the endogenous standpoint* of the agent as a motile, situated, forward-flowing, complex living organism dynamically embedded in and dynamically engaged with the natural world. Indeed, this metaphysically plausible "complete insider's story" about the causality of the action is precisely what we need if we are ultimately to combine our theory of action in Chapters 3, 4, and 5 with our metaphysics of agency in Chapters 6, 7, and 8.

In order to motivate our Essentially Embodied Agency Theory of action, we will work through four serious worries about Davidson's theory of action. Here are capsulized versions of the worries.

First, Davidson's theory of mental events, together with his action-theory, jointly entail that the instrumental teleological rationalizations to which he appeals do not truly refer to *mental causes* at all. By his own admission, the reasons that explain action are causally efficacious only by virtue of their token-identity with physical events.

Second, it appears that the possession of a reason is not in and of itself sufficient for action: some further mental effort or exertion is required on the part of the agent if an intentional body movement is to take place.

Third, we think that Davidson's theory cannot adequately account for the full range of actions carried out by ordinary intentional agents, and that it errs by narrowly concentrating on actions associated with instrumental rationality. One obvious version of this worry is the objection that many

non-human animals and young human children, none of whom can be plausibly taken to be self-conscious and deliberative agents capable of forming or recognizing instrumental reasons for their actions, are nevertheless minded animals who can act intentionally. But even when we focus exclusively on agents who *are* capable of self-conscious deliberative action via instrumental reasons—e.g., rational human animals, or real persons—it seems clear that there are several types of action that fall below Davidson's radar and which do not actually require self-conscious deliberative actions via instrumental reasons. Here we will consider

(i) pre-reflective or *spontaneous* actions,
(ii) akrasia, or so-called "weakness of the will," which we will appropriately re-name *impulsiveness of the will*,

and

(iii) so-called "desire-independent," or *non-instrumental*, reasons for action.

Fourth and finally, we will show how the well-known worry about deviant causal chains poses a fundamental problem not just for Davidson's theory, but for *all* classical causal theories of action.

3.2 Against Davidson 1: Reasons are Epiphenomenal

The first serious problem with Davidson's theory of action emerges in conjunction with his well-known solution to the mind–body problem, *Anomalous Monism*.[14] In accordance with the Principle of the Anomalism of the Mental, Davidson denies the existence of strict deterministic psychophysical laws, but also claims that psychological events are token-identical with certain physical events. He also adopts the Principle of the Nomological Character of Causality, according to which there exists a closed and deterministic system of strict laws into which all causally related events, when appropriately described, fit.[15] It follows that if psychological events are to cause physical events, then there must be strict deterministic physical laws that govern these causal relations. However, in that case mental events

[14] See Davidson, "Thinking Causes."

[15] Mele, "Introduction to *The Philosophy of Action*," 5.

cannot cause physical behavior by virtue of their *mental* properties—which are, after all, only *extrinsic* or *accidental, external* properties of the physical events with which those mental events are token-identical, even if those physical events can be shown to have correct, instrumentally and teleologically illuminating intentional descriptions—but instead only by virtue of the fundamental *physical* properties of the physical events with which mental events are identical.

In this way, because reasons as psychological event-tokens have to be identical to physical events in order to cause our actions, it cannot be the case that they are causally efficacious *as* mental. Insofar as all the real causal work goes on at the underlying physical level, the instrumental teleological rationalizations of which Davidson speaks turn out to be merely ways of informatively and usefully re-describing action. Or otherwise put, reasons can have causal *relevance* because they provide illuminating descriptions of the bodily physical events—descriptions that are perhaps accessible in no other way than through certain teleological and instrumentally rational concepts whose content, due to semantic holism, is irreducible to mechanistic physical concepts—but they have no causal *efficacy*. A type does not have causal efficacy just because one or more of its tokens has causal efficacy. A type has causal efficacy if and only if at least one of its tokens has causal efficacy and the type is an *inherent* or *intrinsic* property of that token. But on Davidson's account the mental properties of physical events are at best extrinsic or accidental, external properties of those events, and all of their intrinsic properties are fundamentally physical. Thus Davidson has not shown us that reasons can actually *do* anything. For *X* to be able to *do* something, presumably, requires that (or at the very least, has as a sufficient condition that) either

(i) *X* is a simple singular event[16] that is a nomologically sufficient condition of a physical event,

or

(ii) *X* is a simple singular event that belongs inherently or intrinsically (i.e., as a necessary proper part) to a complex singular event that is a nomologically sufficient condition of a physical event.

[16] For definitions of the notions of *simple event, complex event, singular event*, and *compound event*, see Section 6.1 below.

And Davidson has not shown us that reasons *as* reasons can satisfy either of these conditions.

Moreover, according to Jaegwon Kim's Explanatory Exclusion Principle or EEP, "two or more complete and independent [causal] explanations of the same event or phenomenon cannot exist."[17] Complete explanations are self-contained and require no other concepts or principles in order to apply to the relevant event or phenomenon. Or in other words, complete explanations are *self-sufficient*. Independent explanations are complete and also rule out other logically distinct concepts or principles from applying to the relevant event or phenomenon at the same time and in the same respects. Or in other words, independent explanations are both self-sufficient and *unique*. So given Kim's EEP, the obtaining of a complete and independent physical causal explanation *excludes* any complete and independent mentalistic causal explanation. Furthermore, the actual existence of the physical causal event confers *epiphenomenality*, or causal inertness, on any corresponding mental event whose properties are strongly supervenient on the fundamental physical properies of the underlying physical event. Therefore, if, as Davidson's theory entails, all the real efficacious causal work is being done by fundamental physical properties and events, then it follows directly from Kim's principle that the instrumental teleological rationalizations appealed to by Davidson are "causes" only because the mental events that constitute reasons are identical with physical events. *In the Davidsonian world, the psychological and rational facts, as psychological and rational, have no causal efficacy whatsoever, even if they do have explanatory causal relevance*.[18]

For Davidson, ultimately, we need to appeal to reasons and mental events rather than merely to physical events, only because *reasons-talk* has epistemic and pragmatic force. For example, if we say that Mary got off the couch because she wanted a beer and believed that a beer was in the fridge, we are *informatively and usefully re-describing* her body movements in terms of her

[17] See Kim, "The Myth of Nonreductive Materialism," 268.

[18] The very same problem applies to the sophisticated Davidsonian account of mental causation offered by MacDonald and Macdonald in "The Metaphysics of Mental Causation." At most their account shows that the *causal relevance* of reasons is not ruled out by Kim's exclusion worries. But this does not show that reasons are themselves *causally efficacious*.

primary reason. But these body movements already have a nomologically sufficient physical cause, so they already have a complete and independent causal explanation. Therefore, our illuminating (i.e., informative, useful) re-description of it cannot have any substantive implications for action-causation. Davidson's theory of action, in effect, falsely substitutes the epistemology and pragmatics of causal explanation for the *metaphysics* of action-causation.

In this connection, then, it is very important to distinguish carefully between

(i) explaining *why* some action happened, i.e., describing the agent's reasons and other motivations,

and

(ii) explaining *how* some action happened, i.e., describing the causal process that actually brought about the action.

When we ask why someone acted as he did, we want to have an illuminating interpretation, "a new description of what he did which fits it into a familiar picture"[19] so that we can make sense of his behavior. Thus, if we want to explain *why* Mary got up from the couch and walked towards the fridge, we are likely to provide a narrative and cite her desire for a cold beer. But on the other hand, if we ask *how* her desire for a beer got her up from the couch and walking towards the fridge, the story is going to be quite different. The teller of the *how*-story must describe an efficacious causal link between Mary's beliefs, desires, intentions, and her intentional body movements. For us, the natural and obvious place to look for this efficacious causal link is in the dynamic neurobiological processes and overt body movements of essentially embodied conscious, intentional minds_{lo}. But merely to place Mary's action in a "wider social, economic, linguistic, or evaluative context,"[20] by supplying a reason for her action, seems to offer little or nothing whatsoever towards a proper characterization of the agent-centered (first-personal, endogenous) causally efficacious process that actually *produced* her action.

[19] Davidson, "Actions, Reasons, and Causes," 33. [20] Ibid.

3.3 Against Davidson 2: Reasons are Insufficient for Actions

Much more must be said about the mental causes of intentional action than is offered by Davidson's theory. As we have just seen, the Davidsonian theory invokes desires, beliefs, and self-conscious deliberative intentions—in short, *reasons*—as causes of action. But in fact it is evident that these factors are not in and of themselves sufficient for action. Mary could want a cold beer, believe that a cold beer is in the fridge, and intend to get one. But obviously it does not necessarily follow that she will actually get up from her comfortable seat on the couch to do just that. She could, consistently with the possession of a complete Davidsonian reason, and without any mental breakdown or pathology of volition whatsoever, *simply continue to be a couch potato*. Those of us who are fond of cold beer and couches alike know this to be all too obviously true. Sometimes the couch just wins out.

So as Searle has pointed out, intentional causation is radically unlike billiard-ball causation in several crucial respects. Even if desires and intentions are present, they are not yet sufficient to compel the agent to act. Necessarily there is what Searle calls (somewhat misleadingly, as we will argue in a moment) a "Gap" between intentions and action, such that the intentional agent is able either to choose or not choose the object of her intention.[21] Or as T. S. Eliot more darkly and poetically puts it:

> Between the idea
> And the reality
> Between the motion
> And the act
> Falls the Shadow.[22]

But one person's Gap and another person's Shadow can also be a third person's *Time to Dance*. What we mean is that if beliefs, desires, and self-conscious deliberative intentions are not themselves sufficient for action, then this is precisely the point at which it seems natural and plausible to say that the volitional phenomenon of *trying* enters in. No matter what her desires, beliefs, and intentions, Mary will never intentionally get up off the couch to go to the fridge for a cold beer if she does not even *try* to get up.

[21] Searle, *Rationality in Action*, 231.

[22] Eliot, "The Hollow Men," verse V, line 31.

And whenever she does in fact intentionally get up off the couch to go for a beer, it is always fundamentally *because* of her trying to do so, together with whatever desire-based reasons she might also have for acting.

In any case, we think it would be a big mistake to think of Searle's Gap or Eliot's Shadow as a nomological fissure in which the deterministic or statistical laws of nature somehow fail to hold, and into which we must insert an agent-cause or some other radically indeterministic source of libertarian free will. Instead, we think that the so-called Gap or Shadow between intentions and acts is far more adequately understood in terms of dynamic systems theory (DST),[23] as merely a *far-from-equilibrium*, or *unstable*, phase in the natural processes of the essentially embodied life of a minded animal.

Now by the notion of a "far-from-equilibrium, or unstable, phase" we mean an ongoing situation in natural processes such that very small changes in the initial conditions of events can lead to unpredictably large effects. Examples would include the Big Bang, black holes, the straw that broke the camel's back, the shout that triggered the avalanche, the flat tire that caused traffic gridlock all over Manhattan, the large effects of small changes in the environment on the weather, the large effects of small environmental and external stimuli on the internal states of living organisms, the large effects of small environmental and external stimuli on the neurobiological and overt intentional movements of minded animals, and so on. Far-from-equilibrium or unstable phases are thus dynamic periods in which the nomological causal architecture of nature—the complete set of general and more specific natural causal laws, all the way down to actual events—is still inherently in process of formation as regards its *finegrained* or *hyper-finegrained* structure.

Here we need to pause briefly to define our terms. In this connection the notion of "roughgrainedness" means that a given concept, property, proposition, or law is identical with any other concept, property, proposition, or law that correctly applies to all the same actual and possible objects or states of affairs. When roughgrainedness holds specifically for causal laws of nature, this means that their application is *invariant with respect to reversals in the direction of time*, or in other words that they presuppose *symmetry*.[24] So roughgrained causal laws are *symmetry-based* causal laws.

[23] Here we are going to deploy, in an anticipatory and intuitive way, some concepts of DST. See Section 7.3 for more details.

[24] See, e.g., Van Fraassen, *Laws and Symmetry*.

The notion of "finegrainedness," by contrast, means that even two concepts, properties, propositions, or laws that correctly apply to all the same actual and possible objects or states of affairs can still significantly differ in their internal structures. When finegrainedness holds specifically for causal laws of nature, this means that their application is *variable with respect to reversals in the direction of time*, or in other words, that they presuppose *thermodynamic asymmetry*.[25] So finegrained causal laws are *asymmetry-based* causal laws.

And the notion of "hyper-finegrainedness" means that even two concepts, properties, propositions, or laws that correctly apply to all the same actual and possible objects or states of affairs, and *also* share the same internal structure, can still significantly differ in how they are presented to or evaluated by a living organism—e.g., a plant, or a non-human animal, or (most interestingly), a conscious, desiring, willing, rational living human organism. When hyper-finegrainedness holds specifically for causal laws of nature, this means that their application is *highly variable with respect to reversals in the direction of time, non-equilibrium, non-linear, dissipatively structured, and naturally purposive*, or in other words that they presuppose *thermodynamic self-organizing complexity*.[26] So hyper-finegrained laws are not only *asymmetry-based* causal laws, they are also *highly context-sensitive* causal laws.

Finally, both finegrained and hyper-finegrained causal laws of nature are *ceteris paribus* laws, which is to say that they are modally robust causal rules of nature that specify necessary connection only against a determinate backdrop of special actual conditions or constraints whether world wide or contextual.[27]

One very simple actual example of a natural process constrained and governed by a hyper-finegrained, *asymmetry-based* causal law in a far-from-equilibrium or unstable phase of nature would be the highly idiosyncratic and fairly silly-looking 2.5-fingered search-and-smash typing movements by which I just quickly hammered out the last sentence of the just-previous paragraph on my Dell Latitude D810 laptop's keyboard. I violated no prevailing roughgrained or *symmetry-based* general laws of physics, whether deterministic or statistical, in order to do this—but at the same time

[25] See, e.g., Prigogine, *Being and Becoming: Time and Complexity in the Physical Sciences*.

[26] See, e.g., Nicolis and Prigogine, *Self-Organization in Nonequilibrium Systems*.

[27] See, e.g., Rupert, "*Ceteris Paribus* Laws, Component Forces, and the Nature of Special Science Properties."

my body movements were *neither* necessitated by the roughgrained or *symmetry-based* general laws of physics together with the settled facts about the past *nor* were they merely the more or less random result of some complex natural statistical sequence of actually past chance events. Those body movements did not just *happen* to me. Silly-looking as they were, they were *uniquely mine*. They were *intentional*. I *performed* them. For better or worse, they were really and truly *up to me*.

It is crucial to recognize that far-from-equilibrium or unstable phases are perfectly *consistent* with whatever roughgrained or *symmetry-based* general deterministic or statistical laws of nature there already are. No roughgrained general deterministic or statistical law is ever violated by an unstable phase and what happens in it. But, in particular, consistency with roughgrained general deterministic laws is not the same as *entailment* by roughgrained general deterministic laws. Because of their inherently "in-process" character, such phases are *not* deterministic in the logical or causal sense of classical LaPlacean determinism. That is, the finegrained or hyper-finegrained *asymmetry-based* natural causal architecture of what occurs during unstable phases is *not* logically or causally entailed by facts about the past together with roughgrained general deterministic laws of nature. Something with new natural causal powers dynamically emerges and makes a novel determinate contribution.

At the same time, however, precisely because uniquely new determinate finegrained or hyper-finegrained natural causal architectures are actually being formed during such phases, such phases and what happens during them are also not strictly *in*deterministic, or the mere logical or causal result of accumulated antecedent random facts, or chance, together with roughgrained general statistical laws. Hence an event or process can be *non*-deterministic, in the sense of being not completely deterministic, without also being *in*deterministic, in the sense of being completely indeterministic. This is because even though such events or processes could not have been predicted *in advance*, at least in principle, they *could* be predicted *from the inside, as the event or process is actually unfolding*, which is to say that, at least in principle, they could be predicted from the standpoint of the uniquely new determinate finegrained or hyper-finegrained causal architecture that is occurrently in process of self-production, and which ultimately stabilizes the natural disequilibrium or instability. Another way of putting this is to say that the event or process is *naturally purposive* or naturally teleological.

Adopting this insider's in-process standpoint—the standpoint, in effect, of *what-it-is-like-to-become-something*, the internal standpoint of a naturally purposive dynamic system—is of course not practically feasible in the case of the Big Bang, black holes, traffic jams, the weather, and living organisms like plants. No matter how much a cosmological physicist tries to imagine her way into "what-it-is-like-to-become-the-Big-Bang," no matter how much a meteorologist tries to imagine her way into "what-it-is-like-to-become-a-thunderstorm," and no matter how much a botanist tries to imagine her way into "what-it-is-like-to-become-an-oak-tree," she is always restricted to an external, or outsider's, explanatory standpoint. But in the case of intentional body movements, the essentially embodied agent *can* predict her own body movements from the insider's in-process standpoint, precisely because *she herself* occupies the standpoint of her own dynamic consciousness$_{lo}$ and intentionality$_{lo}$, and precisely because *she herself* is synchronously willing those body movements. Intentional causation is simply *what-it-is-like-to-become-an-intentional-body-movement*. In the intentional act of dancing, the dancer *consciously becomes* her dance. Wittgenstein puts this crucial point perfectly: "when people talk about the possibility of foreknowledge of the future they always forget the fact of the prediction of one's own voluntary movements."[28]

In short, then, on our view there are three categorically different kinds of events or processes:

(1) *completely deterministic* events or processes, i.e., events or processes that are logically or causally entailed by roughgrained general deterministic laws plus antecedent facts,
(2) *completely indeterministic* events or processes, i.e., events or processes that are logically or causally entailed by roughgrained general statistical laws plus antecedent facts,

and

(3) *natural causal singularities*, which are events or processes that are *neither* logically or causally entailed by roughgrained general deterministic laws plus antecedent facts *nor* logically or causally entailed by roughgrained general statistical laws plus antecedent facts, and also exert

[28] Wittgenstein, *Philosophical Investigations*, 162^{e}, § 629.

> their finegrained and hyper-finegrained causal powers consistently with whatever roughgrained general deterministic or statistical laws there are.

We hold that some events and processes in the natural world are completely deterministic, e.g., the acceleration of falling bodies due to gravity, and systematic changes in the perihelion position of Mercury. To that extent, we are *semi-determinists*. And we also hold that some events and processes are, at least arguably, completely indeterministic, e.g., quantum mechanical phenomena in the microphysical world, and coin-flipping sequences in the macrophysical world. To that extent, we are *semi-indeterminists*. But we also hold that some events and processes in the natural world are natural causal singularities, e.g., the Big Bang, black holes, traffic jams, the weather, and the biological processes and endogenously produced overt movements of living organisms. Amongst the natural causal singularities are what we call *self-determining* events and processes—i.e., intentional body movements. So to that extent, and most importantly, we are also *self-determinists*.

Sometimes Wittgenstein's brilliant remark is read as support for compatibilism—the thesis that free will and determinism are consistent (weak compatibilism), and that both exist (strong compatibilism, a.k.a. "soft determinism"). But as we have just seen, it is arguable that Wittgenstein was driving at something else altogether. An event or process's self-determining intentional predictability *from the inside, as the event is actually unfolding* does not entail complete determinism, even though it also rules out complete indeterminism. So because they are based on dynamic instability, self-determining events or processes are *not completely deterministic*; but because they are also predictable from the inside, self-determining events or processes are equally *not completely indeterministic*. For us then, "causal self-determination" means that an event or process has neither a *closed future* (as in universal natural determinism) nor an *open future* (as in universal natural indeterminism) precisely because, consistently with all the roughgrained general deterministic or statistical laws there are, *it spontaneously, consciously, and intentionally naturally creates its own future* by what it actually does in the ongoing, forward-flowing present.

The larger metaphysical story about the mind–body relation and mental causation that is making up the backdrop to our remarks here will be spelled out in detail in Chapters 6 to 8. But just to present the essentials

of our metaphysical story here as a preliminary soundbite, we will argue in those chapters that the correct description of such semi-deterministic, semi-indeterministic, and self-determining natural events is that they are

(i) *natural causal singularities*, which is to say that they are uniquely new nomologically sufficient causes of physical events, that are also consistent with whatever prevailing roughgrained general deterministic or statistical laws there are,

(ii) *dynamically emergent*, which is to say that they generate causally efficacious truly global or inherently dominating intrinsic structural properties of dynamic systems that are neither reducible to the intrinsic non-relational properties of their parts nor strongly supervenient on the intrinsic non-relational properties of their parts together with all their extrinsic relational properties,

and

(iii) *intentional body movements* of essentially embodied intentional agents.

In this way, far-from-equilibrium or unstable natural dynamic phases in the life of an essentially embodied intentional agent are "Times to Dance," and thereby spatiotemporal sites for the manifestation of *natural creativity*, of which intentional agency is only one special kind—although, obviously, the natural creativity of intentional agency is of great importance for minded animals, whether rational or non-rational, and whether human or non-human.

In any case, our main point here is that we should think of a Davidsonian self-conscious deliberative intention as a normatively *em*powering but not causally *over*powering state in a minded animal that cannot cause an action all by itself, and that instead requires some *additional* mentalistic factor to solidify the initial conditions of events and help constitute a new regime of natural stability. This new regime of natural stability, in turn, is nothing more and nothing less than a natural causal singularity in the essentially embodied agent's neurobiological processes and overt body movements that, along with this additional mentalistic factor, jointly constitute the intentional action. The crucial additional mentalistic factor for constituting intentional action, we shall argue at some length in Sections 4.1 and 4.2, is *trying and its active or initiating guidance of the agent's own living animal body*. But for the moment the crucial point is that if we are to tell an adequate

causal story about action, we cannot treat an action *merely* as a body movement that stands in an illuminating instrumentally and teleologically rationalizing relation to certain belief-desire pairs. We have to say just *how* the action itself happens, *from the inside, as that event is actually unfolding in a naturally purposive way*, by appealing to some mentalistic causal factor *beyond* Davidsonian reasons.

As Searle correctly points out, choosing and deciding are themselves partially grounded in the presupposition that Davidsonian reasons are not in and of themselves causally sufficient for our actions.[29] If they *were* sufficient, then obviously we would always act once the appropriate beliefs and desires were in place, and nothing further would be required. But even given the appropriate beliefs and desires, it is a plain fact that we do *not* always act intentionally, and this in turn is not in every case just because we are prevented from acting by unpropitious circumstances, or compelled to act by coercion or overwhelming outer forces. Sometimes we just shy away from choice: here we either *neglect to try* or *refuse to try*. So there is in Searle's or Eliot's terminology as we have said, a Gap or Shadow—or as we think it should be more accurately described, a far-from-equilibrium or unstable phase, a Time to Dance, or a spatiotemporal site for the manifestation of natural creativity—in the dynamics of intentional action.

But there is not only *one* kind of far-from-equilibrium or unstable phase in the dynamics of intentional action. Indeed, it appears that there are at least *three* different kinds:[30]

(1) There is an instability between desiring to do something and deciding to do it. To want ice cream and believe that it is in the fridge is not yet to decide to go get some. I can just put off deciding.
(2) There is another instability between deciding to do something and actually trying to do it. For example, I may decide to get out of bed many minutes before I actually make the effort to do so, or decide to get out of bed and then not make any effort whatsoever to follow up on this decision.
(3) And there is also another instability between beginning to try to do something and carrying out that task to its completion.To see this, note that trying to run a marathon does not occur in one fell swoop.

[29] Searle, *Rationality in Action*, 71. [30] Ibid.

> To accomplish this task, even professional athletes typically need to try for over two hours. And, as all long distance runners intimately know, it is possible to give up at any point before the end of the race.

Some further factor is needed to stabilize each of these kinds of action-based instability. But it is not going to be *reasons* that do it. To be sure, Davidson admits that in order to be causally effective, reasons must be appropriately connected to resulting actions, and there must be no deviant causal chains at work. However, Davidson does not say enough about the nature of this appropriate connection, nor indeed does he ever acknowledge that there is an inherently action-based instability that needs to be stabilized—much less, three different kinds of instability in need of stabilization.

If it is *trying* that is additionally required in order to cause intentional actions, then we will need a detailed account of the nature and causal role of trying. As a methodological bridge to such an account, Searle's distinction between "prior intentions" and "intentions-in-action" is instructive. While prior intentions are the plans that one often has before undertaking some action, the intention-in-action is the intention one has while actually performing the action.[31] In many cases, we first form a prior intention, and then perform the whole action, which consists of the intention-in-action together with the overt intentional body movement. But in cases of actions that are not premeditated, there is no prior intention, but only an intention-in-action. Therefore some actions are done intentionally even though the actor has formed no prior intention and no prior plan or self-reflection is involved. Such a view seems to reflect a *synchronous* causation model: Searle seems to be saying that intentions-in-action are how an agent is in touch with her body *during* and *throughout* basic action.

But whether or not Searle himself would want to frame it this way, this is the account we will adopt. Our idea is that the mental causes of action—the conscious, intentional mental states, events, and processes of trying and its active guidance—are synchronous with the neurobiological processes and overt body movements that essentially embody those mental activities, and, indissolubly together with those mental activities, jointly constitute intentional actions by virtue of the diachronic instantiation of truly global or inherently dominating, intrinsic structural, irreducibly mental properties

[31] Searle, *Rationality in Action*, 44.

of the agent's living animal body. In the intentional act of dancing, then, the conscious mind of the dancer *simultaneously structures* her dance.

It is true that there is a purely *superficial*, although real enough, time-lag between the time at which the relevant neurobiological process starts in the brain—as indicated, e.g., by neural imaging devices—and the time at which the relevant overt body movement starts at the muscle tissue/skin interface. But this time-lag is entirely due to the fact that it takes time for a neurobiological process to be propagated through all the vital organs and vital systems and become fully displayed as an overt body movement. And thus this time-lag is *not* a temporal difference between an antecedent mental cause that expires before (or just as) its effect begins, and a later bodily event that is its effect, as in classical volitional-causal theories of action. On the contrary, on our view the mental cause is right there at the very beginning of the intentional action in the form of a conscious state, event, or process of trying that is essentially embodied in neurobiological processes; it is right there throughout the development of the neurobiological process in the form of trying's active or initiating guidance as it controls the overt intentional body movements that arise from that neurobiological process and accompany it; and it is *still* right there at the end of the relevant sequence of neurobiological and overt movements, as contained in trying's active guidance of the agent's own body through her successful completion of the entire action. Therefore, the mental cause is synchronous with *all* of the intentional movement's constituent phases. To return again to the example from Yeats, in the intentional act of dancing the dancer's conscious mind simultaneously structures the *whole* dynamic natural event of her dance. As we have said, we will recapitulate and argue for this doctrine at some length in Sections 4.1 and 4.2, and then give a metaphysical analysis of it in Chapters 6, 7, and 8.

In this connection, Mele argues that intentional body movements are causally initiated by *proximal* intentions, that is, "intentions to *A* straight-away,"[32] and that these proximal intentions also play a persisting role in causally sustaining the relevant bodily motions. Sensorimotor feedback indicates whether one's body is moving according to plan or whether things are veering off course and require correction of bodily motions. A plan embedded in the agent's persisting proximal intention is what provides the

[32] Mele, *Motivation and Agency*, 54.

instructions for such corrections, so that intention can play a central role in the causal guidance of ongoing bodily motions. We think that Mele's thesis that a persisting proximal Davidsonian intention can sometimes causally sustain action by providing an empowering plan is quite plausible, and that it improves Davidson's theory. Nevertheless, we also need to be careful *not* to construe intentional body movement in an overly intellectualist way, and also *not* to think that persisting proximal Davidsonian intentions are *universally* necessary for intentional body movement.

3.4 Against Davidson 3: Actions without Reasons

Davidson's theory holds that intentional action is always and essentially a self-conscious or self-reflective and deliberative affair, carried out by means of causally responsible instrumental reasons. But one obvious objection to this view is that many non-human animals and all neurobiologically normal human infants—none of whom can be plausibly taken to be self-conscious or self-reflective, deliberative agents capable of forming or recognizing instrumental reasons—are nevertheless also minded animals who can act intentionally. Of course, we who are members in good standing of the Universal Community of Rational Animals or Real Persons can cognitively *project* instrumental rationality onto their body movements, and admit them as more or less permanent associate members. But as Dennett has pointed out, the same cognitive projection of instrumental rationality could in principle be extended to all sorts of machines and other non-animals.[33] So even if this cognitive projection is both informative and useful to us, it does not at all follow that minded non-human animals and human infants *actually* operate by means of instrumental rationality, unless we have independent reason to believe that they possess the psychological capacities that would support this sort of activity.

Notoriously, Davidson holds that only "talking animals"—i.e., linguistically competent animals—are capable of thought and intentional action.[34] Moreover, he argues as if the conditions for our *ascribing* instrumental rationality to talking animals are the same as the conditions for their actually *having* instrumental rationality. But short of an independent and sound argument

[33] Dennett, *The Intentional Stance*. [34] See Davidson, "Thought and Talk."

for anti-realism—the thesis that the *truth*-conditions of our judgments depend essentially on the conditions under which those judgments can be *asserted* by us—Davidson is not entitled to this identification. And the available evidence very strongly suggests that while many non-human animals and all normal human infants are indeed conscious, intentional agents, nevertheless they are not capable of practical *reasoning*, since this also requires logical reasoning capacities and conceptual capacities, which are intrinsically bound up with linguistic competence,[35] which of course they do not possess. So contrary to Davidson's theory of action, it seems far more plausible to hold that many non-human animals and all normal human infants operate as conscious, intentional agents fundamentally on the basis of *desire-based emotions*, together with whatever else it takes to cause actions.

Moreover, even when we focus exclusively on agents who are *capable* of self-conscious or self-reflective, deliberative action via instrumental reasons—rational animals or real persons, whether human or non-human—there are at least three important types of action that Davidson fails to acknowledge, none of which actually requires self-conscious or self-reflective, deliberatively formed or recognized instrumental reasons for action. Let us now look at these three types in turn.

(3.4.1) *Pre-Reflective or Spontaneous Actions*

Having said what we just said, we clearly need to explore further the role that *intentions* typically are believed to play in action. Davidson claims that a person does not perform an action unless his movement occurs as a result of his relevant beliefs and desires, and that an individual is the agent of an act if and only if what he does can be described under an aspect that makes it intentional.[36] But what does it mean to say that all action is intentional under some description? We cannot suppose that *whenever* an agent acts intentionally, he necessarily goes through a process of self-conscious or self-reflective, deliberative instrumental practical reasoning.[37] Davidson notes that an individual may intend to do X without having decided to do X, deliberated about X, reasoned about X, or formed an intention to do X. What Davidson calls "pure intending" is an unconditional judgment that

[35] See Hanna, *Rationality and Logic*, chs. 4–6.
[36] Davidson, "Agency," 46.
[37] Davidson, "Intending," 85.

an action of a certain sort is desirable, and he claims that such judgment can occur without practical reasoning, action, or consequences.[38] He also holds, however, that if someone acts with an intention, he must have attitudes and beliefs from which, had he been aware of them and had sufficient time, he could have reasoned that his action was desirable. Note that, according to Davidson, while one does not act *intentionally* when one makes a mistake, one nonetheless acts. This is because in cases where I make a mistake and fail to achieve my goal, I nevertheless intentionally do *something*. It is simply that I have done something with the intention of achieving a result that is not forthcoming. Here the question of whether an act was intentional or not quickly becomes very muddled. We believe that some clearer distinctions concerning intention and intentional doings ought to be made.

First, Harry Frankfurt introduces an illuminating, although terminologically somewhat clumsy, distinction between "intentional action" and "intentional body movement."[39] According to Frankfurt, *intentional action* is self-conscious or self-reflective, deliberative action resulting from instrumental reasoning based on desires; but the occurrence of *intentional body movement* need not be intended by the agent by way of either self-conscious or self-reflective forethought or reflective assent.[40] This distinction captures Davidson's recognition that even when I fail to achieve what I intend in a self-conscious or self-reflective, deliberative way, I may nevertheless intentionally do something or other. Unlike Frankfurt, we use the term 'intentional action' more broadly so that it includes all cases of intentional body movement, as well as all cases of self-conscious or self-reflective, deliberative action. Moreover, as we shall argue below, not all self-conscious or self-reflective, deliberative actions motivated by desires are based on *instrumental* reasons. But at the same time we fully endorse the basic upshot of Frankfurt's distinction between the two types of intentional action: it is possible for an agent to do something intentionally, without doing it as a result of self-conscious or self-reflective, deliberative intentions.

Along these same lines, Michael Bratman points out that there is a distinction to be made between "intentionally *A*-ing" and "intending to *A*."[41] To illustrate this, he asks us to imagine a marathon runner who

[38] Davidson, "Intending," 101. [39] Frankfurt, "The Problem of Action," 79. [40] Ibid., 74
[41] Bratman, "Two Faces of Intention," 179.

gradually wears down her shoes over the course of a race. He suggests that while she runs the marathon, she can intentionally wear down her sneakers even if she did not intend to wear them down.[42] This is because what one intends to do is a matter of one's future-directed conduct and what one plans to do, while intentionally moving one's body need not be a matter of plans and explicit purposes.

More generally, it seems quite obvious that minded animal intentional agents do many things intentionally without in any way self-consciously or self-reflectively and deliberatively intending to do them. When one walks to work, e.g., although each of one's steps along the way is an intentional movement, it seems clear that each step does not require its own distinct, self-conscious or self-reflective, deliberative prior intention, plan, or goal.[43] If our efforts always required this degree of self-consciousness or self-reflection, then the fluidity of behavior that we so often see would surely be mostly absent. On our view, then, the paradigmatic and universal sort of intentional action is *intentional body movement*, and not self-conscious or self-reflective, deliberative intentional action. Whenever we engage in intentional body movements, we intentionally act by moving our bodies—but we need not intend to move them in the sense that we self-consciously or self-reflectively *plan* to move them.

Second, it seems that we need not have any Davidsonian reason *whatsoever* in order to move our bodies. To see this, consider the class of "intrinsically motivated actions" that are performed only for their own sake and not for the sake of some further goal or purpose.[44] For example, a person who absent-mindedly hums or drums her fingers on the table normally does so intentionally, even though she has no further goal or "end-directed intention": the reason for action is simply an intrinsic desire, fundamentally based on our primitive bodily awareness, to hum or drum. Otherwise put, the agent hums or drums just because she suddenly *just feels like doing it* at that very moment. While Mele admits that some intentional movements are not done for reasons that involve a belief component of the sort required by Davidson,[45] he believes that this is merely a technical problem for Davidson's account that can be remedied by granting that intrinsic desires to *A* are themselves reasons for *A*-ing. But is intrinsically motivated action really intentional in the Davidsonian sense? That is, even supposing

[42] Ibid., 199. [43] Mele, *Motivation and Agency*, 205. [44] Ibid., 71. [45] Ibid., 73.

counterfactually that the agent had been self-consciously and reflectively aware of what he was doing, would he necessarily have judged that his action was desirable?

Our answer is: *No*. It seems clear to us that intrinsically motivated actions are done for no instrumental reason at all, and that this is sufficiently shown by the dual fact that

(i) the agent acts pre-reflectively or spontaneously without a self-conscious or self-reflective, deliberative intention, or because the agent suddenly just feels like doing it at that very moment,

and

(ii) the action would not necessarily have been judged desirable either at the moment of action or upon reflection by the agent performing it.

For obvious reasons, let us call all such actions *pre-reflective or spontaneous actions*. Here are some examples: humming the faintly annoying Speedy Muffler jingle that you heard on radio and television commercials circa 1976; drumming your fingers on the desk while talking on the telephone; idly wiggling your toes while reading a book; suddenly frowning when the sun goes behind the clouds; biting your fingernails while working on your tax returns; throwing your cell phone across the room in a fit of anger; jumping up and down as an expression of excitement; throwing your hands in the air while freestyle hip-hop dancing; and (making an abrupt Nietzschean shift from the dionysian to the tragic) rolling in the clothes of one's dead wife as an expression of grief.

Rosalind Hursthouse calls these "arational actions," and very plausibly claims that they are all intentional actions explained by their intrinsically resulting from occurrent emotions—or as we would put it, by occurrent desire-based emotions.[46] Hursthouse's term 'arational' is arguably a misnomer, however, in that such actions are done *intentionally* by agents who are fully *sapient* and *sane*. To be sure, there is a sense in which such actions, considered as act-tokens, could be judged to be less than optimally rational from the normative standpoint of instrumental rationality. In that context, the act-token would not necessarily be judged to be in the agent's best self-interest. But unless the normative standard of optimal act-token

[46] Hursthouse, "Arational Actions," 58.

instrumental rationality must be privileged over every other sort of rationality (and it seems plausible to us that it need not be[47]), then because in all *other* respects these acts are rational, they are *authentically* rational.

The everyday phenomenon of losing one's temper is a perfect example. (We are not of course denying that it is possible to have "an anger management problem"—i.e., to have a pathological tendency to experience uncontrollable fits of temper in inappropriate circumstances. We are here talking about *normal* anger.) To run a minor variation on a famous observation by Hume,[48] 'tis not *contrary* to reason to prefer the non-negligible loss of utility consequent upon smashing my expensive cell phone to bits, to my not flinging it across the room. In such cases, we will be strongly inclined to say that while the action was intentional, and while the agent was fully sapient and sane, nevertheless she did not do it for any further goal or purpose. There is no need to ascribe a suitable self-conscious or self-reflective belief or desire, precisely because the very fact that the agent was in the grip of some strong desire-based emotion adequately explains the action. She is a rational animal who acts that way just because she *suddenly just feels like doing it at that very moment*. End of story.

Here someone might object that the agent desires to have and to express her desire-based emotion, and also believes that her action will express it. However, it seems very unlikely that in every or even most such cases the agent has a distinct self-conscious or self-reflective desire that is separate from the desire just to, e.g., fling a cell phone across the room. Rather than acting *in order* to express an emotion, it seems more reasonable to suppose that the agent's action *just is* the expression of a certain essentially embodied, pre-reflective desire-based emotion: in this case, being *completely pissed off*. No doubt it is true that a desire-based emotion itself can count always as a reason for some action or another, and that someone in a desire-based emotional state will tend to have the further self-conscious or self-reflective beliefs and desires typically associated with that state. But it seems obvious that in many cases such beliefs and desires do not count as any sort of instrumental means-ends rationalization for an agent's actual desire-based emotional behavior. So in desire-based, emotionally-driven, pre-reflective or spontaneous action, there need not be some associated

[47] See, e.g., Hanna, *Rationality and Logic;* and Searle, *Rationality in Action*.
[48] See Hume, *Treatise of Human Nature*, book II, part III, section iii, 416.

occurrent or even dispositional belief that acting in such a way will serve some goal, or even bring one pleasure.[49]

Consider again being completely pissed off and then throwing your cell phone across the room. Surely you do not do so with the belief that you will advance some goal. In fact, given the high probability that you will destroy your expensive and highly stylish little communicator, you clearly have much more reason *not* to throw it. And in fact if you reflectively backed off and considered your own desire-based emotion dispassionately, you probably even would desire *not* to throw your cell phone across the room. Similarly, consider the individual who tears out pages of his mathematics textbook in frustration. In carrying out actions such as these, agents do not have any instrumental purpose or aim at some goal. It is not the case that they desire *A* and believe that performing *B* is the best way to achieve *A*. Moreover, it is not the case that their purposes and aims are bizarre, or that their reasons are bad, but rather simply that any instrumental reasons in Davidson's sense for doing that act simply either do not exist, or at least are psychologically suspended. After all, what instrumental purpose could an agent possibly have in tearing out the pages from his mathematics textbook? After doing so, and in a different mood, the agent in all likelihood would judge that his action was not desirable and that it was unlikely to help him do any better at maths. In other words, there is nothing "to be said for" the action "from the agent's point of view"[50]—except, and this will make all the difference, that it is uniquely the agent's *own* desire-based emotion that is being expressed.

This is particularly clear in the case of pre-reflective or spontaneous acts of self-expression. Consider again freestyle hip-hop dancing, and someone's suddenly throwing her hands in the air. She acts that way *precisely because she wants to give bodily expression to the desire to throw her hands in the air.* In effect, she is trying on this desire-based emotion for size, so that the act is then a *bodily picture* of that very desire-based emotion. This nicely supports Wittgenstein's famous remark in the *Philosophical Investigations*, often misinterpreted as behaviorism: "The human body is the best picture of the human soul."[51] In his philosophical terminology, carried over from the *Tractatus*, a "picture"—as opposed to an "image"—captures the internal

[49] Hursthouse, "Arational Actions," 63.
[50] Davidson, "Actions, Reasons, and Causes," 32.
[51] Wittgenstein, *Philosophical Investigations*, 178e.

structure of a real fact.[52] So if the body is the best picture of the soul, then the intentional agent's body in motion is internally structured by the agent's effective desire to move her body in just that way. The freestyle hip-hop dancer gives immediate bodily expression to her desire to dance, and thereby *shows* that pre-reflective or spontaneous intentional action is possible.

So while Davidson believes that an agent who wants to do *X* also believes that he ought to do *X* and that it is desirable to do *X*, the examples we have described in the last few paragraphs clearly show that his depiction of desire as instrumentally evaluative in nature puts too much emphasis on the role of self-conscious deliberation and instrumental reasoning in producing intentional movement. On Hursthouse's view, which we share except for the terminological quibble we mentioned earlier, the cases described above form a single class of actions that are neither intentional actions in Davidson's preferred full-blown sense nor unintentional body movements—they are neither planned nor accidental. This, again, is the class of pre-reflective or spontaneous actions, which inherently involve intentional body movements in the sense that the agent controls her own body. The body movements are neither random nor the result of inner or outer compulsion. The agent is fully sane and her rational capacities are all online. As David Velleman aptly puts it, the action is seemingly "effortless"[53]—although as we will argue in Chapter 4, this effortlessness is perfectly consistent with its also being the result of a certain kind of *trying*. It is true that, as act-tokens, such pre-reflective or spontaneous actions would not necessarily be judged to be perfectly rational by the normative standards of instrumental rationality. But they are still authentically rational intentional acts. It is just that this sort of rational intentional act is the intrinsic result of a desire-based emotion.

Our conclusion is that for rational minded animals, pre-reflective or spontaneous actions are both rationally and intentionally done, and therefore are intentional actions. The Davidsonian theory of intentional action is too narrowly committed to self-conscious or self-reflective, deliberative instrumental rationality. The intentional movements of an essentially embodied rational agent can be authentically intentional without her having any further purpose or primary reason in Davidson's sense.

[52] See Wittgenstein, *Tractatus,* 39–43, props. 2.1 to 2.225; Wittgenstein, *Philosophical Investigations*, 101e, §301.

[53] Velleman, "The Way of the Wanton."

(3.4.2) *Akrasia and Impulsiveness of the Will*

Cases of akrasia, or so-called "weakness of the will," likewise show that the causes of action need not involve reasons in the way that Davidson's theory implies. Davidson holds that if an agent wants to do *X* more than he wants to do *Y* and is free to do either, then he will intentionally do *X* if he does either *X* or *Y* intentionally.[54] He also holds that if an agent judges that it would be better to do *X* than do *Y*, then he wants to do *X* more than he wants to do *Y*. However, he notes that these two theses are inconsistent with the existence of akratic actions. In order to address this difficulty, Davidson suggests that while the akratic person regards one course of action as better (for a reason), he nevertheless does something else (for a reason).[55] That is, while the akratic person judges that all things considered, it would be better to do *X* than *Y*, he can nevertheless unconditionally judge that doing *Y* is better than doing *X*. And while it is only unconditional judgments that can actually bring about intentional action, irrational unconditional judgments may sometimes lead to akratic action. Davidson asserts that "there is no paradox in supposing a person sometimes holds that all that he believes and values supports a certain course of action, when at the same time those same beliefs and values cause him to reject that course of action."[56]

But by reflecting on our own actions and by looking around at everyday life, it seems to us that akratic actions are happening *all the time*. So we find it strange to suppose that agents' all-things-considered-judgments conflict with their unconditional judgments on a regular basis, for this would suggest that a serious form of cognitive and practical dissonance, or irrationality, is an everyday occurrence. On the contrary, everyday life seems to be awash with cases of non-irrational akratic action in which fully sapient, sane agents unconditionally judge that it would be better to do *A* than *B*, are free to do *A*, and yet spontaneously intentionally do *B*.[57] For example, Theresa can fully believe that it would be better not to smoke, be entirely free from compulsion, be perfectly rational and sane, yet just haul off and smoke anyway. Furthermore, the fact that her smoking is an intentional action does not entail that she *intends* to act akratically, for Theresa may intend

[54] Davidson, "How is Weakness of the Will Possible?," 23. [55] Ibid., 34. [56] Ibid., 41.
[57] Mele, *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*, 43.

and plan not to smoke, but then just smoke anyway. Similarly, while an unwilling drug addict, also perfectly rational and sane, but struggling against his addiction, may neither want to desire heroin nor self-consciously and deliberatively intend to take heroin, and only do so as a result of his addiction, nonetheless his movements as he sticks the syringe into his arm are intentional movements, spontaneously (although in this case of course unfortunately) chosen by him. These acts are purely hedonic, but neither egoistic nor self-interested. He *knows* that his heroin addiction is a truly terrible thing for him personally.

So we are saying that actions can be perfectly sane and rational, intentional, akratic, and *neither* egoistic, *nor* in what the conventional wisdom takes to be the agent's best interest, *nor* even in what the agent *himself* takes to be his own best self-interest. This again highlights the fact that self-conscious or self-reflective, deliberative plans and purposes that result from the desires rational agents deem best, are not in any way necessary for intentional movement. We can act pre-reflectively and spontaneously. But it also discloses the deeper and perhaps surprising fact that rational agency and so-called weakness of the will are intrinsically connected. We also can act impulsively, against either what is, or appears to be, our own best interest. And thank god for that, since it is obvious that acting in the service of either what is or appears to be our own best interest is *not* always the best way to act. In this way, to take akrasia seriously is to *open up* our conception of rational agency.

More precisely, cases of akrasia reveal that an agent's current Davidsonian or self-conscious or self-reflective, deliberative intentions can be easily defeated or deflected by occurrent opposing wants and desires, and that making an unconditional judgment is not a causally sufficient condition for performing a particular action. Because one's self-conscious deliberative evaluation of a particular desire need not match the motivational force of that desire, it is obvious that instrumental reasons and self-conscious or self-reflective, deliberative intentions do not tell the full story.[58] For example, while Janet may think it best to satisfy her desire to write a philosophy paper this afternoon, her desire to go out drinking with her friends may simply turn out to have much more motivational force, in which case she will carry out the impulsive act of going out drinking with

[58] See also ibid., 84.

her friends. Similarly, people sometimes impulsively fall passionately and truly in love with one another entirely against their own better judgment. Such cases also show that the desires we do judge, or would judge, to be *worse* often exhibit much greater motivational force than the desires that are in line with our self-conscious deliberative instrumental judgments, prior intentions, decisions, and plans. So this is in direct conflict with Davidson's claim that necessarily if one judges it would be better to do *X* than *Y*, then one wants to do *X* more than one wants to do *Y*.[59]

In this way, whereas Davidson holds that desire and self-conscious or self-reflective, deliberative instrumental evaluation are intrinsically linked, we believe on the contrary that the everyday, widespread fact of akrasia makes it very likely that a narrowly instrumental rationalist view of action is simply *incorrect*. But this is not an attack on the very idea of rational agency. Given the widespread fact of akrasia, it is not the case that we are really *less* rational than Davidson takes us to be, nor is it the case that we are somehow fundamentally *irrational*, nor even that we are fundamentally *arational*. It is rather that our rationality, like our consciousness, is essentially embodied and thereby grounded on desire-based emotion, and so intrinsically *open* to akrasia. In this way, our rationality is intrinsically dynamic and *impulsive*: that is, we are rational animals whose nature is such that, when push comes to shove, we can just haul off and ignore, override, or reject our selfish or best or all-things-considered self-conscious or self-reflective, deliberative instrumental reasons.

This is perhaps most obvious and plausible in cases when adopting such instrumental reasons would lead to morally *impermissible* acts, as in the standard Kantian counterexamples against ethical egoism, where the impulse to ignore, override, or reject our selfish or self-interested instrumental reasons is a necessary condition of moral autonomy.[60] But the same point holds in certain consequence-independent cases in which both we ourselves and everybody else would in fact be much better off if we adopted an instrumental reason and acted that way. Consider, e.g., (to use a minor variant on Bernard Williams's famous example of "George the chemist") an unemployed chemist, call him George*, who impulsively refuses to take a job doing chemical and biological warfare research, even though

[59] Davidson, "How is Weakness of the Will Possible?" 23.
[60] See Kant, *Groundwork of the Metaphysics of Morals*, sections I–II.

his family badly needs the money, and even though he will actually then be in a position to make the research go more slowly than other keener candidates.[61] In such cases, adopting the instrumental or consequentialist reason as our own would violate our fundamental projects and our personal integrity, and so we feel deeply in our guts and hearts that we must ignore or reject it, and so act on an impulse contrary to or without instrumental reasons, and independently of the consequences. In other words, we non-instrumentally *plump* for that action. This sort of non-instrumental, non-consequentialist impulsiveness or "plumping" is *not* irrationality or arationality, precisely because it preserves our psychological authenticity (the internal coherence of our desires, beliefs, volitions, and actions), which in turn is a necessary condition of our rational autonomy.

Akrasia is thus a partial expression of our deepest *freedom* of the will. If this is correct, then it is not accurately described as any sort of volitional *failure*. So akrasia should *not* be thought of as "weakness of the will," which seems to reflect a much more modern, sanctimonious, and sternly moralistic notion than the ancient Greeks actually had in mind, but instead as *impulsiveness of the will*. To be sure, free will itself is not the *same* as impulsiveness of the will. The complete fact of free will is much more than that, arguably including negative freedom (freedom from preventative checks or overriding compulsion), positive freedom (freedom to choose or do what I want), psychological freedom (the subjective experience of being negatively and positively free), deep freedom (transcendental freedom or original "up-to-me-ness"), moral responsibility, moral authenticity or integrity and rational autonomy (self-legislation).[62] But at the same time, free will nevertheless includes impulsiveness of the will as a necessary condition.

(3.4.3) *Desire-Overriding Reasons*

In a closely-related way, what seems to us to be the obvious fact of *non-instrumental* reasons for action also poses serious problems for Davidson's account of action. Following a broadly Kantian line, Searle, e.g., defends the possibility of non-instrumental rationality and maintains that there are motivations that lead agents to do things that they would honestly say they

[61] See Smart and Williams, *Utilitarianism: For and Against*, 97–8. See also Sartre, "Existentialism is a Humanism"; and Williams, *Moral Luck*.

[62] See Kane, *A Contemporary Introduction to Free Will*.

do not *want* to do.[63] One can eat carrot sticks, brush one's teeth, or behave morally even when one seemingly (or at least initially) has no desire to do so. Searle argues there are things we find ourselves obligated to do whether we want to at that moment or not, and that in virtue of certain speech acts, we create "desire-independent" reasons. When a speaker makes an assertion, he creates a non-instrumental reason for accepting the logical consequences of what he has just said.[64] Likewise, when one makes a promise, one commits oneself to carrying out what one has promised, so that the making of various commitments is built right into the structure of speech acts. Searle believes that insofar as such commitments create non-instrumental reasons for action, obligations create a species of action in which occurrent desires do not play a central causal role. Correspondingly, it may seem that some actions are fully explained by evaluative beliefs about what one should do. Louise may invite Marty to the party simply because she believes it is the right or appropriate thing to do, and not because she wants to invite him or would enjoy seeing him there. These sorts of non-instrumental reasons appear to pose a challenge to the standard Humean–Davidsonian account, which sees all actions as explicable by pairs of desires and means-ends beliefs.

While we fully agree that the Humean–Davidsonian account requires significant revision, we also believe that Searle's characterization of non-instrumental reasons as "desire-independent" is significantly misleading. In this context, it seems that the notion of "the desire-independence of a reason" can mean either

(1) a reason that is exclusive of any and all desires (strong desire-independence),
(2) a reason that is underdetermined by any and all desires even if it is always associated with desires (moderate desire-independence),

or

(3) a reason that is underdetermined by certain first-order desires, even if it is always associated with desires (weak desire-independence).[65]

[63] See Searle, *Rationality in Action*, ch. 6; Kant, *Groundwork of the Metaphysics of Morals*; and Kant, *Critique of Practical Reason*.

[64] Searle, *Rationality in Action*, 173.

[65] The notion of underdetermination in this connection is most easily explicated as *non-supervenience*, such that X underdetermines Y if and only if Y does not supervene on X. See Section 1.1.

The strong desire independence, or desire-exclusion, of a reason means that the reason bears no relation whatsoever to any and all desires. For example, a divine or angelic being might have a reason for acting that is strongly desire-independent. The moderate desire-independence, or universal desire-underdetermination, of a reason means that the reason bears no *intrinsic* or *necessary* relation to any desires whatsoever, although it could also still bear extrinsic or contingent relations to some or even all of the agent's desires. For example, an indeterministic agent cause or a radical existential voluntarist, who is somehow capable of making choices or decisions while standing apart from all his own desires, might have a reason for acting that is moderately desire-independent. But the weak desire-independence, or specific-desire-underdetermination, of a reason as we will understand it, means only that the reason is *not necessarily related to certain first-order desires*.

This leaves open the possibility that acts are still always justified and motivated by facts about our desires, even though we are not always justified or motivated to act by a special class of first-order desires. For example, we could sometimes be justified and motivated to act by a special class of *higher*-order desires (say, the desire to be moved at time *t* by a non-selfish, non-egoistic, non-self-interested, non-hedonistic, and consequence-independent concern for others) that are directly in opposition to certain *first*-order desires (say, any selfish, egoistic or self-interested, hedonistic, or consequence-driven first-order desire at *t*), and so adopt a reason that is provided for an agent by facts about higher-order desires, but not provided for an agent by facts about all first-order desires.

So while desire-independence in the first sense entails desire-independence in the second and third senses, the converse is not the case. More generally, it seems that Searle is implicitly equivocating between these three logically distinct kinds of desire-independence, and also that most of what he says is in fact consistent with *weak* desire-independence, and thus consistent with a hierarchical desire-based approach to practical reasons. But in order to sort things out properly, we will first have to say something about our own general approach to the nature of reasons.

It seems plausible to hold that reasons are (or are provided for agents by) facts that normatively support (and thus justify or motivate) beliefs and intentional aims or actions, and do not merely cause or mechanically trigger those beliefs, aims, or actions. It remains possible, however, for reasons to be

at least proper parts of *mental causes* of those beliefs, aims, or actions—e.g., in self-conscious deliberative intentional actions. Reasons for beliefs are *epistemic* reasons, and reasons for intentional aims or actions are *practical* reasons. Now there may be deep connections between epistemic reasons and practical reasons. For example, Searle has argued that epistemic reasons are a sub-class of practical reasons—and if so, then epistemic reasons are nothing but practical reasons for undertaking the intentional act of believing. For the rest of this sub-section, however, we will concentrate exclusively on practical reasons.

Internalism about reasons says that reasons both justify and motivate our intentional aims or actions. So all practical reasons are internal reasons. Internalism normally entails a desire-based theory about the nature of justifying reasons. According to that theory, justifying reasons are (or are provided for agents by) facts about the desires of persons. By contrast, *Externalism* about reasons says that while all reasons justify our intentional aims or actions, nevertheless at least some and perhaps all reasons fail to motivate our aims or actions. So some or all practical reasons are external reasons. Externalism normally entails an objective-value-based theory of the nature of justifying reasons, according to which justifying reasons are (or are provided for agents by) facts about the ends or objective values recognized by persons, and not by facts about their desires.

The primary philosophical virtue of Internalism about reasons is that it offers a very plausible account of action. But its primary philosophical vice is that it cannot account for *desire-overriding* non-instrumental or consequence-independent justifying reasons, since it is normally assumed that all desire-overriding non-instrumental, consequence-independent justifying reasons must also be *desire-independent*. Correspondingly, the primary virtue of Externalism about reasons is that it *can* account for desire-overriding non-instrumental, consequence-independent reasons. But its primary philosophical vice is that it *cannot* plausibly account for how justifying reasons can cause actions if they are not based on desires.

The view of reasons we want to defend is a non-standard form of Internalism about reasons that *also* fully accounts for desire-overriding non-instrumental, consequence-independent justifying reasons. It does this by holding that all reasons are (or are provided for agents by) facts about the desires of persons, including not only all instrumental, consequence-dependent reasons, but also some *desire-overriding non-instrumental,*

consequence-independent reasons. And this, in turn, is because at least *some* reasons express some essentially non-selfish, non-egoistic or non-self-interested, non-hedonistic, and consequence-independent desires that can themselves override our selfish, egoistic or self-interested, hedonistic, and consequence-dependent first-order desires. So we call our view about practical reasons *Desire-Overriding Internalism*. More explicitly, according to Desire-Overriding Internalism about reasons,

> **Thesis A:** while all reasons are both justifying and motivating, and all reasons are (or are provided for agents by) facts about the desires of persons,

nevertheless it is also the case that

> **Thesis B:** while many or even most reasons are instrumental and consequence-dependent, at least some reasons are desire-overriding non-instrumental, consequence-independent reasons, precisely because they are (or are provided for by) facts about some essentially non-selfish, non-egoistic or non-self-interested, non-hedonistic, and consequence-independent desires of persons.

But how can Desire-Overriding Internalism be true? How can some of our desires ground desire-overriding non-instrumental, consequence-independent reasons? The answer is that if one adopts, as we do, a version of the *hierarchical desire* model of the will and personhood first developed and defended by Frankfurt,[66] then Desire-Overriding Internalism about reasons is perfectly coherent and possible.

But before we unpack this idea, we will need to pause very briefly to make some distinctions and acquire some terminology. A desire is a felt need for something, or a preference for something, or a wish for something. Desires and wants can be taken to be essentially equivalent. To desire X is to want X; to desire to X is to want to X; and conversely. Now according to Frankfurt (and we fully agree), some conscious animals have not only *first-order desires* but also *effective first-order desires*. Effective first-order desires are desires that move (or will move, or would move) the conscious animal all the way to action. An effective first-order desire, that is, is the same as a conscious animal's *will*. For example, I have an effective first-order

[66] Frankfurt, "Freedom of the Will and the Concept of a Person."

desire for a cold beer that gets me off the couch and all the way to the refrigerator, the bottle opener, and finally to guzzling that beer, and so that is what I have willed, or is my will, on that occasion. First-order desires may or may not also be accompanied by *second-order desires*. In having a second-order desire, I want (not) to want *X*, or to want (not) to want to *X*. For example, I may get a beer from the refrigerator and drink it because I effectively desire one, but also at the same time I might want *not* to want that beer, because I know that drinking it (and its friends) will make me podgy and sleepy, when I should on the contrary be skinny and alert. Now suppose, counterfactually, that my wanting not to want that beer actually stops me from getting up off the couch. In this way, at least some of my second-order desires can be directed to the determination of precisely which first-order desire is the effective first-order desire, or my will, and such desires are *second-order volitions*.

This in turn leads to an account of personhood framed primarily in terms of desires and the will, and only secondarily in terms of rationality: real persons are all and only the essentially embodied conscious, intentional creatures—i.e., minded animals—capable of having second-order volitions. The capacity for having second-order volitions also automatically confers a capacity for *instrumental* rationality on any creature that has that capacity. But it does not thereby confer a capacity for self-conscious or self-reflective, autonomous, non-instrumental rationality. According to our account, then, necessarily all real persons are essentially embodied rational animals, but not all real persons are self-conscious or self-reflective, autonomous, non-instrumentally rational animals. For example, normal small children between the ages of two and five (a.k.a. "toddlers") are obviously not self-conscious or self-reflective, autonomous, non-instrumentally rational animals, although they are real human persons.

Here we differ substantively from Frankfurt. He holds that some rational animals are not persons because they are "wantons" and (temporarily or permanently) incapable of having second-order volitions. This seems wrong. On our view, essentially embodied rationality is inherently related to intentional agency, and entails real personhood. Frankfurt also holds that small children and all non-human animals are wantons. This too seems wrong. Toddlers, e.g., are obviously not self-conscious or self-reflective, autonomous, non-instrumentally rational animals. But they are also certainly *meta*-conscious (i.e., they are conscious of their own consciousness),

instrumentally rational (i.e., they very often know exactly what they want, and intentionally pursue their goals with great intensity and stubbornness), and highly *willful* (i.e., their wills are highly impulsive). Hence toddlers are capable of second-order volitions, and are real persons. And the same seems true of at least some non-human animals, e.g., Great apes.

In any case, the crucial point is that if we adopt the hierarchical desire model of the structure of the will, then we can hold that it is *effective first-order desires* that always move us to action, so that **Thesis A** above—which says that all reasons are both justifying and motivating, and grounded on desires—will be obviously true with respect to our effective first-order desires.

On the hierarchical desire model developed by Frankfurt and also adopted by our Desire-Overriding Internalism, some special second order-desires, namely second-order volitions, are directed to the determination of precisely which first-order desires are to be effective. But as Desire-Overriding Internalists, we *also* adopt the following doctrine, which we call *the Desire-Overriding Desires Thesis*:

> The hierarchical volitional constitution of every self-conscious or self-reflective, autonomous rational animal is such that some second-order volitions are not only able, under successful volitional conditions (which Frankfurt calls "freedom of the will"), to determine just which first-order desire is the one that moves us on that occasion, but also can *override* an occurrent first-order desire that would otherwise have motivated the agent to action on that occasion, either by
>
> (1) *impulsively generating a new first-order desire* in order to substitute it for the first-order desire that would otherwise have effectively moved the agent to action on that occasion,
>
> or else
>
> (2) *impulsively super-charging another relatively motivationally weak occurrent first-order desire* in order to select it to be the effective first-order desire instead of the would-be effective first-order desire.

This impulsive desire-overriding activity would occur, e.g., in a case in which a self-conscious, autonomously rational agent has a motivationally forceful occurrent first-order desire to embezzle some money without

fear of detection and thereby gain some significant personal benefit (say, paying for his law school education), and also significantly benefit others (he also intends to give a large donation to CARE). But then he recognizes a non-instrumental, consequence-independent reason for *not* embezzling the money, despite the risk-free and significantly beneficial personal and social consequences of embezzlement, which then evokes a successfully overriding second-order volition to be moved by a first-order desire to be an honest person. And this in turn happens either

(i) by impulsively generating a new first-order desire to be an honest person

or

(ii) by impulsively super-charging an already-existing but less motivationally forceful desire to be an honest person.

But more precisely, just *how* can the self-conscious or self-reflective, autonomously rational animal actually either impulsively generate new first-order desires or impulsively super-charge existing ones, in this desire-overriding way?

It seems clear enough that *instrumental, consequence-dependent* first-order desires can be either impulsively generated or impulsively super-charged merely by saliently presenting or re-presenting their intentional contents to a rational animal. This principle is of course an axiom of consumer advertising. In this way, e.g., while driving along the highway you may not begin to impulsively want a chocolate milkshake until you see a billboard with pictures of them. Or while sitting on your comfortable couch in front of the television set you may not begin to impulsively want a brand-new BMW until you watch a television commercial about them. Alternatively, seeing the billboard or watching the TV commercial may simply impulsively reinforce and strengthen a pre-existing desire to drink a chocolate shake or buy a new BMW.

Now we need only make this process reflexive or self-applying, and hierarchical. Since every second-order desire (say, my wanting to want a chocolate shake) includes the intentional content of a first-order desire (in this case, my wanting a chocolate shake) within its own content, second-order desires can either impulsively generate new first-order desires or impulsively super-charge pre-existing first-order desires in essentially

the same way as the billboard or TV commercial cases, that is, just by making the content of those first-order desires *salient to oneself*. So, under the right conditions, I can either make myself impulsively want a chocolate shake or impulsively super-charge a pre-existing desire to want a chocolate shake, by simply *wanting to want a chocolate shake*, and thereby getting an impulsive first-order desire to want a chocolate shake to be the effective one. This self-prompting of impulsive effective first-order desires by means of second-order volitions—surely—happens all the time.

But how will this work in the case of the first-order desire to embezzle, which is then overridden by a non-instrumental, consequence-independent second-order volition? Here one might ask oneself,

> Do I really want to want to embezzle that money? Won't I be just a *dishonest* [insert here your most personally meaningful profane expletive for characterizing a bad person] if I do this? And isn't my intention to give some of the money to CARE just a pathetic attempt to hide my disgusting dishonesty from myself?,

and this might either impulsively generate a new first-order desire to be an honest person or impulsively super-charge a pre-existing desire to be a honest person, and thereby get the first-order desire to be an honest person to be the effective one. Of course, the same volitional effect can also be brought about through the practical imagination. You can, e.g., imagine yourself both resisting the awful temptation to embezzle and then later giving a large donation to CARE as a kind of penance, and then love that idealized image of yourself so much that it either impulsively generates a new first-order desire to be an honest person or impulsively super-charges a pre-existing one, and thereby get a first-order desire to be an honest person to be the effective one.

So let us assume that there are at least *some* conscientious, self-conscious, autonomously rational people who are *sometimes* tempted to do risk-free bad things, but *still* manage to fight off those temptations and impulsively motivate themselves to be morally *good* instead of morally bad. In the natural order of things, the volitional successes of such conscientious people rarely reach the local newspapers, radio news, or television news, since by hypothesis these volitionally successful—and in Frankfurt's terminology, "free"—agents never in fact *do* the bad things, and also rarely tell anyone else about their temptations and inner struggles, since that

would be intensely embarrassing. As they say, no good deed ever goes unpunished. In any case, if our assumption about the existence of some conscientious people is correct, then the self-prompting impulsive production of desire-overriding non-instrumental, consequence-independent effective first-order desires by means of second-order volitions therefore happens *sometimes* too.

None of this will make *any* sense unless we can explain how it is that someone has a desire-overriding non-instrumental, consequence-independent *second-order volition* to want to want to be honest despite the intense temptation to embezzle. On our view, this desire-overriding non-instrumental, consequence-independent second-order volition can be explained by postulating the existence of a universal and innate desire-based emotional disposition in self-conscious, autonomously rational animals *to be moved at least sometimes by non-selfish non-egoistic, non-self-interested, non-hedonistic, and consequence-independent effective first-order desires*.

For lack of a better name, let us call this universal and innate emotional disposition to be moved at least sometimes by non-selfish, non-egoistic, non-self-interested, non-hedonistic, and consequence-independent effective first-order desires, *the desire for self-transcendence*. The desire for self-transcendence, as we understand it, is a fundamental felt need for some form of self-abnegation, self-denial, self-discipline, and self-effacement in our lives, independently of the consequences. It demands that we at least sometimes delay, reject, or sublimate the satisfaction of occurrent first-order desires, and it also demands that we sometimes widen our outlook from the narrowly selfish, to the standpoints of significant others, and even to a synoptic standpoint encompassing all persons, and that in so doing we do not pay any sort of attention to the calculation of consequences. In this latter respect it is quite close to what Hume calls *sympathy*,[67] and so this important Humean moral emotion could be usefully regarded as a sub-species of the desire for self-transcendence. In any case, the ultimate aim of the desire for self-transcendence is to integrate and unify the self in a better and more complete way by sometimes rigorously disciplining one's current self, regardless of the consequences, and so its upshot is that to the extent that we can satisfy this desire, we *can impulsively overcome*

[67] See Hume, *Treatise of Human Nature*, book II, part I, section XI.

the incoherence, inauthenticity, narrowness, egoism, and calculating enslavement to consequences that constitutively characterize our current selves.

Self-transcendence, as we are understanding it, clearly seems to play an essential role in morality, religion, mysticism, cults, martyrdom, sainthood, artistic genius, scientific genius, philosophical genius, certain kinds of athleticism (e.g., long distance running, and many other high-performance sports), and also in seemingly bizarre personal projects such as that of Simeon Stylites, who lived for many years on top of a pillar.[68] It also includes the highly perverse kind of non-selfish non-egoistic, non-self-interested, non-hedonistic, and consequence-independent desire described famously by Hume, and that we alluded to earlier in this section, namely that it would not be contrary to reason for me to prefer the destruction of the world, including of course the total destruction of myself along with everyone else, to the scratching of my finger.[69] Moreover and more radically, it seems impossible to conceive of any self-conscious or self-reflective, autonomous rational animal that is completely *in*capable of feeling the desire for self-transcendence. This becomes much clearer when we note, as the perverse Humean desire shows, that the desire for self-transcendence does *not* necessarily imply self-transcending goals that are morally *good*, morally *permissible*, or even particularly *nice*.

Indeed it is a striking, surprisingly widespread, and occasionally tragic fact that self-transcending values can also be highly *immoral* or just highly *perverse*. Why else would it be true that many otherwise ordinary and decent-seeming people will often go well out of their way, usually in a completely self-destructive and apparently unself-satisfying manner, and more generally for no instrumentally good reason at all, just to be *horrid brutes* to other people? Anyone who has worked in an academic department at a university for a few years, or in a law office, or in a business corporation, or even just at a fast food place, knows this to be all too true. The recently popular British and then American TV program *The Office* brilliantly displays this striking fact about rational human nature in a highly humorous way.

[68] It is of course possible to have deep philosophical worries about the moral defensibility and value of the desire for self-transcendence. See, e.g., Nietzsche, *Beyond Good and Evil*; and Wolf, "Moral Saints."

[69] See note 67 above. There is another reading of Hume's remark, to the effect that he is saying that it would not be contrary to reason to want to avoid scratching my finger even though the *rest* of the world would be destroyed. But then Hume is just talking about mere titanic selfishness, which presumably no one would ever think of sharply contrasting to instrumental rationality.

More dramatically and sublimely, however, the novels of Fyodor Dostoevsky,[70] which are filled with characters who are what Lillian Hellman very aptly called "sinner-saints,"[71] also brilliantly reveal that even the most selfish, egoistic, hedonistic, calculating, and wicked people care very deeply about self-transcending values, and are likely to reveal this in times of greatest personal crisis and stress—such as being condemned to death by a firing squad and pardoned at the last moment, a truly terrifying event that Dostoevsky himself actually experienced, and which changed his life.[72]

If we are correct, then it is *also* a consequence of our view that if a minded animal *lacks* any capacity to have a desire for self-transcendence, or experiences a significant disruption or distortion of this capacity, then he is to that extent non-rational. Sociopaths, e.g., seem to be clear examples of human beings who are significantly damaged in that way.

In any case, the main point we are driving at is this. *If* we posit a universal and innate desire-based emotional disposition to, and thus a fundamental felt need for, self-transcendence in all self-conscious or self-reflective, autonomous rational animals, then not only is **Thesis A** above—which just re-states standard internalism about reasons—obviously true. It is *also* the case that **Thesis B** above—which says that while many or even most reasons are instrumental and consequence-dependent, at least some reasons are desire-overriding non-instrumental, desire-independent reasons, precisely because they are (or are provided for by) facts about some essentially non-selfish non-egoistic, non-self-interested, non-hedonistic, and consequence-independent desires of persons—is obviously true. This is because **Thesis B** applies directly to all the would-be selfish, egoistic or self-interested, hedonistic, and consequence-dependent effective first-order desires that are overridden by successful second-order volitions expressing the desire for self-transcendence.

If all of this is correct, or even just roughly correct, then the way is open to allowing for reasons that are independent of certain first-order desires that an agent has prior to his recognizing non-instrumental and consequence-independent reasons, while at the same time acknowledging that all reasons are (or are provided for agents by) facts about the desires of real persons.

[70] See, e.g., *The Brothers Karamazov, Crime and Punishment, The Devils*, and *The Idiot*.

[71] Hellman, "Introduction," to Hammett, *The Big Knockover*, viii. She is referring specifically to Hammett.

[72] See Dostoevsky, *The House of the Dead*.

According to Searle, to recognize a reason for acting is already to recognize a reason for *wanting* to accept it. So it seems to us that Searle's so-called "desire-independent" reasons in fact generate second-order volitions that inherently express the desire for self-transcendence in some way or another, and thereby motivate the self-conscious, autonomous rational agent to act by getting a desire-overriding first-order desire (whether new or super-charged) to be the effective one.[73]

For example, according to Kantian ethics, an individual's second-order volition to desire to keep a promise is derived from the fact that she recognizes she has made a promise, together with a universal, innate desire-based emotional disposition that is automatically triggered by that recognition. This universal, innate desire-based emotional disposition, which Kant calls *respect* (*Achtung*), generates the desire be moved to action by non-selfish, non-egoistic or non-self-interested, non-hedonistic, consequence-independent, and *morally correct* first-order desires.[74] Thus, Kantian respect, like Humean sympathy, and like Humean finger-scratching world-destroying perversity, is another important sub-species of the innate desire for self-transcendence. Of course, an individual might recognize her moral obligation to keep a promise and still not act on it. The second-order volition generated by her recognition of her obligation together with the higher-order moral emotion of respect may fail to determine the desire that is effective in producing action, and thus be volitionally unsuccessful (or in Frankfurt's terminology, "unfree"). For Kant, however, for a person to recognize her moral obligation is *thereby* also to recognize a desire-based and yet also desire-overriding justifying and motivating reason, co-grounded in the moral emotion of respect, for action.

What we are claiming, then, is that for self-conscious or self-reflective, autonomous rational animals (a class which would include all ordinary sane, sapient adult human beings), at least some non-instrumental, consequence-independent reasons, as recognized under the appropriate conditions, trigger the universal, innate dispositional desire for self-transcendence, and thereby give rise to corresponding second-order volitions to be moved by the appropriate first-order desires. So we are saying that by their very *nature*, self-conscious, autonomous rational animals have the innate dispositional

[73] Searle, *Rationality in Action*, 176.

[74] See, e.g., Kant, *Groundwork of the Metaphysics of Morals* and Kant, *Critique of Practical Reason*.

desire for self-transcendence built right into all their thought and action. This deep-seated pull towards self-transcendence is not only very different from other latent or standing desires that an intentional agent may have, but it also seems to be an essential part of what makes us *us*.[75]

No doubt this is a controversial claim. After all, consider the case of Eve, who suffers from clinical depression and at first glance appears to be utterly unmotivated to do what she believes she is morally required to do. Suppose, e.g., that Eve recognizes that she is morally required to help her uncle, and yet utterly lacks the effective motivation to do so. Mele asserts that such cases are examples of the "problem of listlessness," and that in such instances, agents completely lack motivation to fulfill their obligations.[76] Nevertheless, we find it difficult to imagine that Eve, even though she is clinically depressed, has no motivation *whatsoever* to help her uncle. To be sure, her motivation to stay at home is ultimately motivationally significantly stronger, because by hypothesis it is the effective one. Still, because also by hypothesis Eve recognizes her moral obligation to help her uncle and because she is a self-conscious or self-reflective, autonomous rational animal, this recognition necessarily yields a second-order desire to be moved by a first-order desire to help her uncle. Thus, she *wants* to be moved by a first-order desire to help him. But because of her clinical depression, some disruption or distortion in the essentially embodying neurobiological basis of her desire-based emotions contingently prevents or undermines the generation or super-charging of a first-order desire to help him. Thus, her desire-overriding non-instrumental, consequence-independent second-order volition is wholly unsuccessful.[77] Yet that does not imply its non-existence as a psychological capacity. Indeed, the abject failure of her desire for self-transcendence in this context precisely implies its *existence* as a psychological capacity and vividly points up the pathological fact of her clinical depression.

Similarly, evaluative beliefs can play a direct causal role in impulsively producing new first-order desires or in impulsively super-charging the motivational force of other occurrent but relatively motivationally weak

[75] See Korsgaard, *The Sources of Normativity*.

[76] Mele, *Motivation and Agency*, 111.

[77] In at least some of the cases in which agents are completely unmotivated to do that which they recognize they are morally obligated to do—although this is not Eve's problem—this can be understood as a severe deficiency in rationality, and as a kind of insanity. So we will be strongly inclined to regard such individuals as sociopaths or psychopaths.

first-order desires. Along these lines, Michael Smith proposes that all intentional action is mediated by the overall tendency of our psychology towards personal coherence, and by the presence of a generic desire for this sort of coherence.[78] Among ordinary rational agents, there is some sort of necessary connection between believing an act to be desirable and having at least some motivation to perform the act. In cases of maximal consistency, the agent's evaluative judgments will be in line with his or her emotional and motivational states. We think that Smith's notion of generic desire for personal coherence is plausible, and also that it is closely related to what we have called "authenticity," but would also want to locate this desire more fundamentally in the innate dispositional desire for self-transcendence. So if this is correct, then the Smithean desire for coherence would count as another important sub-species of it, along with Kantian respect, Humean sympathy, and Humean finger-scratching world-destroying perversity.

In this connection, and now to borrow and slightly modify an example of Searle's, suppose that an individual recognizes the validity of a logical proof. Because the proof demonstrates that one chapter of her dissertation is utterly confused and mistaken, she does not want to accept it. However, once she recognizes that the argument is valid and sound, she has a reason for accepting it and some sort of desire to accept it. So she will come to have this desire impulsively and in spite of herself, although of course she may repress or suppress it. But what makes this possible? It seems that if upon recognizing the validity and soundness of the proof she had no desire whatsoever to accept it, then she would be deeply irrational in some way.

Note that it is not that she has a desire to be logically consistent in her beliefs and desires in the same way that she has a desire to be intelligent or witty. Rather, it is part of her very psychological nature and her existence as a rational being that she feels the deep need for some sort of consistency between her beliefs and desires quite apart from her egoistic interests, self-interest, hedonic interests, or her attention to consequences. Similarly, once a person recognizes that she has made a promise to write a letter of recommendation for one of her students by a certain date, she must as a matter of consistency come to have some sort of desire, however motivationally weak, to keep that promise. If she does not, her beliefs and desires will be utterly and completely out of joint, and then presumably

[78] Smith, *The Moral Problem*, 32.

she already is, or will become, unable to make sense of herself. As we all know first-hand, however, self-conscious or self-reflective, autonomous rational animals are not *perfectly* rational or *perfectly* moral beings. In fact, we very frequently screw up in little, medium-sized, or colossal ways. This means that although an individual recognizes the proof as valid, she may also impulsively fail to accept it. Similarly, although a person's recognition that she has made a promise must issue in a second-order volition to be moved by a first-order desire to keep her promise, this first-order desire need not be effective in action. The agent in question may very well end up impulsively going to the movies, then to the pub, and finally home to fall exhaustedly into bed, instead of writing that letter of recommendation.

One last point should be made in this connection. The Desire-Overriding Internalism about reasons that we are advocating, for all its radicalness, is still in one absolutely crucial respect importantly weaker than the more standard internalistic and instrumentalist view of reasons defended by Davidson and many others. Davidson says that if an agent unconditionally judges that it would be better to do *X* than to do *Y*, then he wants to do *X* more than he wants to do *Y*.[79] So on his view, if someone sincerely believes he ought to do something, then his belief must show itself in his behavior, his inclination to act, and his desire. Such a view reflects Davidson's belief that the natural expression of a desire is instrumentally evaluative in form. Someone who wants to do *X* believes that he ought to do *X* or that it is desirable to do *X*. For example, an individual who honestly believes that it is desirable to stop smoking has some pro-attitude toward his stopping smoking; feels some inclination to stop smoking; and will do so provided that nothing stands in the way, he knows how, and he has no contrary values or desires.[80] Thus for Davidson, pro-attitudes express instrumental value judgments that are at least implicit.

But while we agree that an unconditional belief that one ought to do something does ordinarily generate some sort of first-order desire to do that thing, we do not follow Davidson in holding that this desire must produce a corresponding action. The everyday widespread existence of cases of akrasia or (as we think of it) *impulsiveness of the will* clearly show that evaluative judgments need not match up with the motivational force of desire-based emotions. And actions done for non-instrumental,

[79] Davidson, "How is Weakness of the Will Possible?," 23. [80] Davidson, "Intending," 86.

consequence-independent reasons based on the innate dispositional desire for self-transcendence equally clearly show that instrumental, consequence-dependent reasons significantly underdetermine intentional agency. Since actions done for non-instrumental, consequence-independent reasons, in addition to being impulsive, are often also pre-reflective or spontaneous actions, our Desire-Overriding Internalism about reasons and the fact of pre-reflective or spontaneous actions naturally go hand-in-hand.

3.5 Against Davidson 4: Deviant Causal Chains Again

The last and certainly most extensively discussed problem for the classical causal theories of action is the possibility of deviant or wayward causal chains. As we noted at the beginning of the chapter, and as Frankfurt has pointed out, classical causal theories of action say that

> a bodily movement is an action if and only if it results from antecedents of a certain kind. Different versions of the causal approach provide differing accounts of the sorts of events or states which must figure causally in the production of actions. The tenet they characteristically share is that it is both necessary and sufficient, in order to determine that an event is an action, to consider how it was brought about.[81]

But if a belief-together-with-desire, or a Davidsonian reason, produces action in the wrong way, or *accidentally*, then the body movements that result are a *mere* causal effect of them rather than an intentional act caused *in response* to them.[82] For example, there is the case of the unfortunate rock climber who, because it is so painful to hold the rope that supports her partner, naturally desires to rid herself of the weight of her partner and knows that loosening her grip on the rope would do the trick. But, tragically, this belief so unnerves her that she loosens her hold, and drops her partner. And there is another case of someone who intends to spill his drink at a party in order to signal to some accomplices to begin a robbery. Although he has not yet committed any crime as he stands there sipping his drink, the thought of doing so makes him so nervous that his arm trembles and he spills his drink, thereby initiating the robbery. In these cases, it is highly implausible

[81] Frankfurt, "The Problem of Action," 70.

[82] Audi, *Action, Intention, and Reason*, 17.

to hold that the agent's dropping her climbing partner or spilling his drink is an intentional act, precisely because she or he does not actually control the arm-movement that dropped or spilled. Hence the dropping and the spilling were *unintentional* body movements, not intentional movements.[83]

One might think, however, that these cases are quite different from "normal" cases of unintentional body movement—e.g., someone who trips over a curb in the dark or loses his balance when a train suddenly jerks forward. More specifically, it seems that both the unfortunate rock climber and the trembling robber bear some degree of *causal* responsibility for their unintentional movements, even if they are not strictly speaking *morally* responsible. This is because these body movements *do* result from their desires, but *not* in the proper way, that is, as a result of trying and its active guidance. It is true that one would no doubt hold the trembling robber morally responsible for setting up a state of affairs in which something like this could happen, and for having wicked intentions, even if his unintentional body movement is something for which he is not strictly speaking morally responsible. But when one trips over a curb in the dark or loses his balance when the train jerks forward, on the other hand, this is not in any way a result of desire, so it seems that there is not even any causal responsibility in such cases, much less moral responsibility.

This calls for a brief remark on the notion of unintentionality and the ascription of it in ordinary language. Often we will correctly say that someone does or did something unintentionally—say, tripping over a curb in the dark—but this should not be taken to mean that there is such a thing as *unintentional basic acts*, or "unintentional intentional action." Tripping over the curb in the dark is an unintentional body movement, but not strictly speaking an act of any sort. Moreover, sometimes even in cases *other* than those of unintentional body movements, we will also correctly say that someone does or did something unintentionally. But here it seems that we are really talking about unintended *effects* or *side-effects* of our basic actions and intentional body movements, whether foreseen or unforeseen, and not about our basic acts and intentional body movements themselves. For example, we might say that I acted unintentionally in shooting someone by accident or by mistake, even though the trigger is intentionally pulled by me in both cases. Or again we might say that I acted

[83] Frankfurt, "The Problem of Action," 70.

unintentionally by killing some innocent civilian bystanders with a missile when the target was a military one, even though, again, the missile-firing button is pressed intentionally by me. And so on. In such cases, we certainly do ascribe causal responsibility to the agents, and sometimes ascribe moral responsibility too. Nevertheless, the agent might *also* use the fact that the effects were unintended (say, in the case of shooting someone by accident) as part of an excuse in order to avoid accepting moral responsibility, blame, or punishment.[84]

In any case, examples involving deviant causal chains lead to a general worry about classical causal theories of action, as Mele observes, because

> whatever psychological causes are deemed both necessary and sufficient for a resultant action's being intentional, cases can be described where, owing to a deviant causal connection between the favoured psychological antecedents and a pertinent resultant action, that action [i.e., that body movement] is not intentional.[85]

It is very important to note that the standard examples of deviant causal chains hold whether, as in the agent-causal theory, we place the agent *outside* of time, or as in the volitional-causal theory and the Davidsonian theory, we place the agent *inside* the series of mental or physical events. For even if we assume that the very idea of timeless causal agency for minded animals actually makes sense, such a timeless agent could then also always produce the relevant action through mere nervousness. Supra-temporality, presumably, is in and of itself no protection against the subjective experience of anxiety, as in the unfortunate rock climber and trembling robber cases. This shows that every classical causal theory of action is incomplete, for it cannot make proper sense of the causal connection that must obtain between a mental antecedent and a body movement in order for action to count as intentional. As a minimal condition of philosophical adequacy, then, any causal theory of action must be fully equipped to explain precisely what goes wrong in cases of causal deviance.

We believe that our own non-classical causal theory, grounded in trying and its active guidance, smoothly explains these types of cases. The unfortunate climber never actually *tries* to loosen her hold, and the

[84] See, e.g., Austin, "A Plea for Excuses." Moral responsibility, blame, and punishment are clearly not the same. Person X can hold person Y morally responsible for doing something bad, while at the same time also forgiving Y, thereby either not blaming Y or at least ceasing to blame Y. And X can hold Y morally responsible and blame Y, while also reasonably refusing to punish Y.

[85] Mele, "Introduction to *The Philosophy of Action*," 6.

trembling robber never actually *tries* to signal to his accomplices. So any body movements resulting from their beliefs and desires—or indeed from any other causal antecedent—by a deviant causal chain will automatically be both uncaused by the agent and also unintentional, and thus not be a counterexample to our theory. Moreover, even if my frustrated trying to raise my paralyzed arm accidentally neurally triggers a strange causal mechanism that later brings about the rising of my arm in a very bizarre way—say, by triggering a signal to a ray-gun on Mars that sends a tractor-beam back to earth and levitates my arm—nevertheless we could not correctly say that my unintentional body movement is actually caused by me, since of course by hypothesis I was trying to *raise* my arm, and was not trying to bring about the *rising* of my arm. The rising of my arm merely *happens to me*. On our view, an arm-raising necessarily requires a synchronous trying and its active guidance, which in turn requires essentially embodied engagement. But during the time when my paralyzed arm rises by means of the tractor beam, I am not actually trying to raise it, nor am I actively guiding its movement. In fact, part of the causal chain leading to body movement is utterly detached from my body, and so it is obvious that this is something that merely *happens to me*. Therefore deviant causal chains can be easily accommodated by our non-classical causal theory of action.

In this connection, we think that it is very important to distinguish carefully between

(1) deviant causal chains

and

(2) non-standard causal mechanisms.

A non-standard causal mechanism is a causal process that produces a certain effect, but is not normally deployed for the production of that effect, or occurs relatively infrequently in the normal course of nature. Deviant causal chains accidentally bring about an effect by means of some or another non-standard causal mechanism. Yet a non-standard causal mechanism can also be used to bring about an *intentional* body movement. For example, if I discover that my left arm is paralyzed and want to raise it, I can move it into position using my right hand and right arm. The movement of my right hand and right arm is a basic intentional act, and the movement of

my left arm is a non-basic intentional act brought about by the intentional body movement of my right hand and right arm. I thereby raise my left arm by means of a non-standard causal mechanism. This is not, however, a deviant causal chain, since the effect is brought about *non-accidentally* even if in a non-standard way.

In this chapter we have developed a full-scale critique of Davidson's highly influential theory of action. We also sketched the outlines of a decisively *post*-Davidsonian and *non*-classical causal theory of action, the Essentially Embodied Agency Theory. More precisely, we argued that reasons in the Davidsonian sense are *never* mental causes of actions, even though every intentional action has a mental cause. Instead, the mental causes of basic actions are synchronous, essentially embodied, pre-reflectively conscious effective first-order desires, or *tryings that actively guide intentional body movements*. Furthermore, even though every intentional action is normatively supported by internal reasons, provided for agents by facts about their desires, many of these reasons are neither self-consciously nor self-reflectively recognized, and many of these reasons are *not* instrumental reasons. Finally, even when a reason for action is self-consciously or self-reflectively recognized, it is frequently a *non-instrumental* reason.

Of course, even beyond thoroughly criticizing Davidson's theory and handling the problem of deviant causal chains, our non-classical causal theory of action must also respond directly to the three other basic worries about the Davidsonian theory. Most importantly, our theory must provide independent positive grounds for claiming that trying and its active guidance explain intentional action. To begin the development of this positive account, in the next chapter we will look closely and critically at Frankfurt's guidance-based theory of action and O'Shaughnessy's theory of trying.

This page intentionally left blank

4

Essentially Embodied Agency II: Guidance and Trying

> Complexity of body movement suggests action only when it leads us to think that the body, during the course of its movement, is under the agent's guidance. The performance of an action is accordingly a complex event, which is comprised by a bodily movement and whatever state of affairs or activity constitutes the agent's guidance of it.
>
> Harry Frankfurt[1]

> Trying to move a limb is a unique mental event simply in being *standardly* a cause of physical change . . . But even more important is the fact that trying is *in essence* normally a cause of bodily change . . . Thus, it is a primitive constituent of animal consciousness, which yet constitutively cannot exist without bodily phenomena.
>
> Brian O'Shaughnessy[2]

4.0 Introduction

When I perform a basic intentional action, what am I actually doing? In this chapter we offer a new way of construing and combining two familiar answers to that very hard question. First, in Section 4.1 we look at Harry Frankfurt's notions of guidance and intentional movement, and argue that these should be understood in terms of essentially embodied synchronous mental causation. We also argue that Frankfurt's notion of guidance, when construed as *active* guidance, as opposed to what we call a merely *maintaining* guidance, allows us to distinguish between

[1] Frankfurt, "The Problem of Action," 73.
[2] O'Shaughnessy, "Trying (as the Mental 'Pineal Gland')," 66.

(i) self-conscious or self-reflective, deliberative intentional action (e.g., pre-planned arm-waving),
(ii) pre-reflective or spontaneous intentional action (e.g., impulsively throwing one's arm in the air while freestyle hip-hop dancing),

and

(iii) unintentional body movements (e.g., a Dr Strangelove-like arm-rising).

Second, in Section 4.2 we look at Brian O'Shaughnessy's conception of *trying* in order to help us make sense of the synchronous mental activity of active guidance that causes intentional action. But while O'Shaughnessy correlates successful tryings with corresponding overt intentional body movements that are temporally separated from trying events, by contrast we take trying to be a conscious, intentional process, grounded in desire-based emotion, that begins as incarnated by the neurobiological processes of the agent's living organismic body, occurs synchronously with those neurobiological processes, and also extends throughout the entire duration of the overt intentional body movements that arise from and accompany those trying-informed neurobiological processes. Finally, in Section 4.3, we respond to a serious challenge to our theory from contemporary cognitive science.

In effect, our new, non-classical causal theory of action combines a suitably-refined version of O'Shaughnessy's conception of trying with a suitably-refined Frankfurt-style guidance model of action. The metaphysical epoxy resin glue that binds them ineluctably and synergistically together, and fills all the classical causalist gaps, is the Essential Embodiment Thesis—that every consciousness like ours is necessarily and completely neurobiologically embodied. And this, of course, is why we call it the Essentially Embodied Agency Theory of action.

4.1 Towards a Non-Classical Causal Theory 1: Active Guidance

In "The Problem of Action" Frankfurt famously criticized classical causal theories for claiming that a body movement is an action if and only if it

causally results from antecedents of a certain kind. The nub of the difficulty here, according to Frankfurt, is that classical causal theories

> locate the distinctively essential features of action exclusively in states of affairs which may be past by the time the action is supposed to occur. This makes it impossible for them to give any account whatever of the most salient differentiating characteristic of action: during the time [an agent] is performing an action he is necessarily in touch with the movements of his body in a certain way.[3]

The problem isolated by Frankfurt is the now familiar failure of classical causal theories (whether of the volitional-causal variety or the Davidsonian variety) to close the *temporal* gap between an agent's intentionality and her intentional body movements (see Section 3.1). On Frankfurt's own view, however, to determine whether a set of body movements qualifies as an intentional action, we should ask "whether or not the movements as they occur are under the [agent's] guidance."[4] Thus rather than focusing on what was going on in the agent's mind *before* her body movements began, we should instead direct our attention to what is going on in the agent's mind *during* the time at which those body movements occur, and therefore to what is going on in the agent's mind *synchronously* with those body movements—namely, the agent's trying and its guidance of her body movements.

What is guidance? A characteristic feature of guided body movements is that they "cohere in creating a pattern which strikes us as meaningful."[5] While a pianist's hands moving over a keyboard create a meaningful pattern, the thrashing around of an epileptic does not. Frankfurt says that all intentional movement counts as *purposive* movement, and also that movements are purposive if and only if they occur "under the guidance of an independent causal mechanism, whose readiness to bring about compensatory adjustments tends to ensure that the behavior is accomplished."[6] But not all purposive movement is intentional movement. The dilation of one's pupils in response to light, for example, is purposive in that it is guided by an independent self-adjusting causal mechanism, but nevertheless it does not count as an intentional movement. So what is needed, over and above a movement's being guided by an independent

[3] Frankfurt, "The Problem of Action," 71. [4] Ibid., 72. [5] Ibid. [6] Ibid., 74.

self-adjusting causal mechanism, for it to be intentional? Here is what Frankfurt says:

> Complexity of body movement suggests action only when it leads us to think that the body, during the course of its movement, is under the agent's guidance. The performance of an action is accordingly a complex event, which is comprised by a bodily movement and whatever state of affairs or activity constitutes the agent's guidance of it.[7]

This, it seems, adds two further necessary conditions of guidance to the existence of an independent self-adjusting causal mechanism:

(1) the causal mechanism belongs to the living animal body of the intentional agent,

and

(2) the causal mechanism operates "not prior to but concurrent with the movements they guide,"[8] so that guidance of action necessarily occurs synchronously with body movements.

It seems obvious that Frankfurt is not objecting to causal theories of action per se, but rather only to the classical causal theories that frame action in terms of antecedent beliefs, desires, intentions, or reasons. For as we have just seen, Frankfurt himself explicitly postulates an independent self-adjusting causal mechanism, belonging to the agent's living body, that operates synchronously with action; and he also asserts that we cause purposive body movements precisely by deploying that self-adjusting mechanism to guide those very movements. To be sure, this entails the existence of *simultaneous* and *continuous causation*, whose possibility is sometimes denied. But there are also some quite compelling general metaphysical arguments in favor of simultaneous and continuous causation[9] that we will consider explicitly later, in Section 6.1.

Quite apart from that metaphysical issue, some theorists also have argued that there are direct counterexamples to Frankfurt's analysis of action. For example, George Wilson asks us to imagine the following case (which we have briefly discussed already in Section 3.5): An agent grasps her paralyzed left arm with her right hand and then uses her right arm to put

[7] Frankfurt, "The Problem of Action," 73. [8] Ibid., 75.
[9] See, e.g., Huemer and Kovitz, "Causation as Simultaneous and Continuous."

the left one into some desired position.[10] While Frankfurt holds that an intentional action is a body movement that is synchronously guided from start to finish by means of a purposive causal mechanism belonging to the agent's body, this case of a paralyzed left arm's manipulated and therefore, according to Wilson, "guided" movement does not seem to qualify as an intentional action, and thus it appears to pose problems for the sufficiency of the analysis. Anticipating a Frankfurtian response, Wilson also proposes the following revision: Intentional action is a body movement that is not only guided by the agent but also "performed" by the agent.[11] Cases in which an agent's right arm guides her left arm into position therefore do not exhibit the right sort of "performative" guidance. But what does it mean for an agent to "perform" the act? It is clear that more must be said about the intimate connection between the agent and the movements of her body during the time she acts.

We want to argue that this intimate connection can be adequately understood in terms of *trying*. According to our analysis of the paralyzed arm case, the right arm exhibits the intentional agent's trying and its active guidance, while the left arm remains paralyzed and passive, and is put into position by the intentional movements of the right hand and arm. The movement of the left arm as it goes into position, even though that arm is paralyzed, and even though this movement has been brought about by a non-standard causal mechanism, is not *un*intentional because the agent, by hypothesis, intends to get the left arm into position. But it is the intentional agent *quâ* her right-handedness and right-armedness, and not the intentional agent *quâ* her left-armedness, that exhibits trying. The right hand and right arm movement therefore is a case of normal intentional body movement, and a basic act. The positioning of the agent's left arm, by contrast, is a case of *slightly weird* intentional body movement brought about by successfully using a non-standard causal mechanism—the agent's unparalyzed right hand and right arm—and is a *non-basic* act. Trying is causally operative throughout, since although the agent is *intending* to get her left arm into position, she does so *by* trying to move her right hand and right arm.

Strictly speaking, however, even though the agent is intending to get her left arm into position by trying to move her right hand and right arm,

[10] Wilson, *The Intentionality of Human Action*, 48. [11] Ibid., 49.

she is not now *trying* to move her left arm. In fact she unsuccessfully tried to move her left arm *earlier*, found out that it was paralyzed, and *now* is bringing about the intended positioning of her left arm by a different and non-standard causal means. What she is *now* and successfully trying to do is just to operate this different and non-standard causal means, i.e., she is now successfully trying to move her right hand and arm in such a way as to get her left arm into the desired position. So, again, the agent's normal intentional movement of her right arm and right hand is her basic act, and the slightly weird intentional movement of her left arm is a non-basic act that is carried out by means of trying and its active guidance of the normal movement of her right hand and right arm.

Active Guidance vs. Maintaining Guidance

At this point it is important to note some ways in which Frankfurt's notion of guidance, as he presents it, is not wholly consistent with the Essentially Embodied Agency Theory. Frankfurt says that it is "not essential to the purposiveness of a movement that it actually be causally affected by the mechanism under whose guidance the movement proceeds."[12] As an illustration he gives the example of a driver who is allowing his car to coast downhill, and says that what makes this an instance of intentional action is that the driver is prepared to intervene actively if necessary. So according to Frankfurt, it is guidance in the sense of at least a *preparedness to intervene actively*, rather than antecedent causation in the sense of the classical causal theory, that is universally necessary for action.

But Frankfurt's point here is at best somewhat misleading, and at worst outright mistaken. The coasting driver example seems to us to be clearly a case of what we call *maintaining* guidance, and not *active* guidance. We hold that while maintaining guidance is involved in many or even most basic or non-basic intentional acts, nevertheless it is always parasitic on trying and its active guidance, and not primary. So if Frankfurt is intending to say that guidance, whether in basic or non-basic intentional acts, is *primarily* a matter of preparedness to intervene actively, and not primarily a matter of direct and active intervention in the neurobiological causation of intentional body movements and in the bringing about of the non-basic effects of intentional body movements, then we think he is mistaken.

[12] Frankfurt, "The Problem of Action," 75.

In this connection, it is important to distinguish between two subtly different senses of maintaining guidance:

(i) maintaining guidance in *basic* intentional acts, which requires that the guidance mechanism belongs to the living animal body of the intentional agent,

and

(ii) maintaining guidance in *non-basic* intentional acts, which does not require that the guidance mechanism belongs to the living animal body of the intentional agent.

In other words, the difference between the two senses of maintaining guidance is determined by whether the intentional agent's guidance mechanism is *infra-body* or *extra-body*. Since non-basic acts presuppose basic acts, something can be an extra-body guidance mechanism for a non-basic intentional act only if it operates via an infra-body guidance mechanism for a basic act. Thus no matter how skilled I am at operating automobiles or other simpler machines or tools (say, a hammer or a hockey stick), these items can function in non-basic intentional action as ''extensions of my body'' only if my living body is already directly involved in basic intentional action.

Consider again the coasting driver example. In sense (i) of maintaining guidance, the coasting driver's guidance mechanism is the set of vital systems of his body. Here he remains relatively relaxed and poised to make driving movements with his arms and feet, but is actually sitting still. By contrast, in sense (ii) of maintaining guidance, the coasting driver's guidance mechanism is the extra-body steering, accelerator, and braking system of the car. Here the car is not being actively driven, even though it is also neither driverless nor out of control. Now, considering *both* senses of maintaining guidance taken together, clearly the coasting driver could not currently be guiding either his own body or his car if he had not *actively intervened* earlier in order to make driving movements and to start driving the car. Also he could not currently be guiding either his own body or his car if he had not *actively intervened* again later to make the relevant driving movements and thereby start his car coasting downhill. Nor, finally, could he be currently guiding either his own body or his car if he were not *prepared to intervene actively again* in order to make the relevant driving movements and thereby prevent his car running out of control. But then

it must be the case that active intervention is the primary mental cause of intentional body movements, and also the primary fact of guidance. The mere preparedness to intervene actively is a derivative phenomenon. Actively intervening mental directedness and guidance, with respect to *basic* intentional acts, is the same as what we are calling "trying and its active guidance."

In light of these points, we want to say that the mental cause of basic intentional action is a desire-based emotive trying that controls body movements in two different but intimately related ways, and specifically such that the second way is parasitic on the first way. First, all intentional action starts in an actively intervening conscious mental directedness towards making an intentional body movement, or *trying*, and it can continue to be actively guiding throughout the duration of action—such as when an agent swims across a lake, or climbs a steep staircase. Second, once an overt intentional body movement is already well underway by means of trying and its active guidance, then trying can *also* play a merely maintaining role—in sense (i) above—by temporarily standing down from active guidance. Maintaining guidance in sense (i) is at work, e.g., when an agent casually reaches out across her desk for her cup of coffee, or comfortably swings her legs and arms while walking. But during such merely maintaining phases in the overall dynamic genesis of the intentional body movement, the agent remains vigilantly ready to intervene actively if body movements get off-track. Suppose, for example, that her hand trembles slightly as she reaches for her cup. Then she actively intervenes in order to slow down the reaching and grasping movement and steady her hand, thereby preventing a minor crash or disaster in the course of that body movement—say, knocking over the cup, or dropping it instead of picking it up—and thus keeping faith with the original trying.

It seems clear that *both* sorts of guidance—i.e., active guidance, and maintaining guidance in sense (i)—are involved in *most* cases of basic intentional action. Consider, e.g., a fieldgoal kicker as he kicks a football. The action starts when the ball is snapped and the kicker begins to kick the ball by engaging in an actively intervening trying that starts his body striding towards the place where the ball will be positioned by the holder. In the next stages of the kick, during his wind-up and including the highly graceful downward arc and torque of his leg and kicking foot as he drives his foot into the ball, the kicker's guidance is still

active because he is constantly renewing his actively intervening conscious control of the movements of his body, just as he has trained for years to do. But in the later stages of the action, after he has actually propelled the ball upwards from the ground and towards the goalposts, and as he follows through, he shifts from active guidance into a phase of merely maintaining guidance in sense (i), by allowing his leg and foot to carry on towards the goalposts and gradually relax as he completes the entire kicking movement. If, however, a would-be kick-blocker were suddenly to break through his protective cordon of linemen and throw himself at the ball, the kicker might also suddenly change his follow-through to avoid smashing his leg and foot into the flying body of the kick-blocker. If so, then this would involve another actively intervening trying, and thereby modulate the ongoing basic intentional act into a new phase of active guidance.

Similarly, in cases of everyday basic intentional acts, such as when an agent sits up and then remains sitting up, or starts to reach out for her coffee cup and then continues to move her arm casually towards the cup, maintaining guidance in sense (i) is all that is needed to keep the overall intended body movement on track. If, while sitting in his chair during a long and boring department meeting, an agent dozes off and begins to lose his balance and falls over towards the table, he is usually able to catch himself with a jerk and sit up again. If so, then he is able to re-engage himself in an actively intervening way by quickly trying to re-direct his ongoing body movement if it is necessary to prevent that movement (in this case, sitting up straight) from crashing. Or, if it suddenly occurs to an agent in mid-act that she no longer wants to reach for her coffee cup, then she can re-engage herself by gracefully aborting the cup-reaching movement, turn it into a reaching movement towards the mouse of her desktop computer instead, and then point and click.

Such cases demonstrate an important distinction between the early stages of action and the middle or later stages before the action is complete. While on our view mental directedness or trying and its active guidance is synchronous with every phase of the action, and thus lasts *throughout* the entire intentional body movement from its non-overt inception in neurobiological processes to the completion of the overt body movement, nevertheless a constant actively intervening *re-engagement* or *renewal* of trying (as, e.g., in swimming across a lake or climbing a steep staircase) often is

not needed during every single dynamic phase of the process. Rather, it is frequently or even normally the case that trying and its active guidance modulates temporarily into and out of maintaining guidance in sense (i). This is vividly evident, e.g., in the case of the fieldgoal kicker.

At this point one might wonder whether, on the assumption that we are correct that every basic intentional action is caused by trying and its active guidance—with an option to modulate temporarily into and out of maintaining guidance—it follows that an overt intentional body movement caused by such trying and its active guidance must *always* be accompanied by a corresponding actual shift in the location of the body or its limbs. Otherwise put, is it possible for an agent to perform an overt intentional body movement by keeping her body *motionless*?

Mele presents an illuminating case which indicates that the answer is a definite *yes*. Making a special effort to establish and sustain a certain position and orientation of one's body can *also* be an intentional body movement, even if it does not involve an actual change in the location of the body or limbs. Suppose that Ann, who wants to do her part in a passive resistance protest, makes herself into a deadweight by using special bodily self-control techniques she has learned in yoga classes.[13] Ann's remaining motionless while the police drag her away is then an overt intentional body movement, and one that can easily be made sense of within our trying-based and active-guidance-based approach, precisely because she actively intervenes in both her ongoing neurobiological processes and her overt body processes by *trying* to establish and sustain precisely that bodily position and orientation. After all, the condition of one's muscles in such circumstances is not in any way the same as lying on the beach in a state of relaxation. If Ann were passive with respect to her body movements, and lying comfortably on the beach, then her natural bodily reflex would be to squirm or jump up if someone touched her bare midriff with clammy fingers. So in order to make her body into a deadweight for the police—who, we can assume, have extremely clammy fingers—Ann must exercise special control over her body throughout the duration of that intentional action.

Now the case of Ann is in sharp contrast to the case of Al, who takes a pill that he knows will induce complete unconsciousness for ten

[13] Mele, *Motivation and Agency,* 147.

minutes.[14] When Al sees the police coming near, he swallows the pill and unconsciousness sets in, so that Al *also* is a deadweight for the police. However, because he fails to actively guide the position or orientation of his own body during the time these positionings and orientings occur, Al's being a deadweight is merely an intended consequence of his pill-taking action rather than an intentional action. Ann *actively makes herself into a deadweight* and thus is fully present as an intentional agent in the motionless position and orientation of her own body, whereas Al when he is unconscious is *nothing but a deadweight* and thereby temporarily absent as an agent. We believe the distinction between the cases of Ann and Al nicely highlights the fact that the infra-body neurobiological guidance mechanism from which overt intentional body movements normally arise, and whose normal modus operandi is a set of overt changes in the location of the agent's body or its limbs, can *also* be fully operative in motionless positions and orientations.

The fact that intentional movements can take the guise of motionless body positions and orientations points up another extremely important feature of the Essentially Embodied Agency Theory of action. Because conscious, intentional minds_{lo} are necessarily and completely embodied, there really are *no such things* as acts of "pure thinking," or purely mental-to-mental causation, precisely because all intentional acts are *also* intentional body movements. So on our view, even working out a logical, mathematical, or philosophical problem in one's head while sitting motionless, necessarily involves trying and its active guidance of a neurobiological process and an overt body movement. For even while a thinker is sitting motionless, her abstract thinking requires a certain amount of "brain power," or neural processing, and necessarily also a certain amount of vital activity in the body beyond the brain, and also a certain kind of motionless body position and orientation arising from these neurobiological processes and accompanying them.

In this way, as you engage in abstract thought you are also (say) sitting, and neither standing nor lying down; and normally if you are sitting, then you are also sitting upright, and neither sitting tipped sideways nor suspended upside down. Indeed, while it is certainly true that no *specific* types of body attitude are necessarily associated with corresponding *specific*

[14] Ibid., 148–9.

types of consciousness$_{lo}$ or intentionality$_{lo}$, as the behaviorists wrongly insisted,[15] nevertheless it seems that we do necessarily embody our consciousness and our abstract thinking in *some or another type* of intentional body movements, some of which do involve motionless positions and orientations.[16] Indeed, Rodin's iconic 1904 statue *The Thinker* shows that we even have a motionless bodily stereotype[17] for the essential embodiment of our abstract thinking. As everyone knows, *The Thinker* presents a naked man sitting with his head propped on his right hand and arm, and his right arm propped on his left knee. This specific motionless body position and orientation immediately betoken the cognitive attitude of contemplative thought. Could the anonymous contemplative man in *The Thinker* have been represented as swinging upside down from a trapeze? That would have been an absurd surrealist joke worthy of Duchamp or Magritte thirty years later. This all goes to show that although metaphysical Behaviorism—the reductive materialist thesis that mental properties and facts are nothing but second-order physical facts about dispositional input-output mappings within organismic or mechanical bodies—is certainly mistaken, there was something also deeply *right* about the philosophical impulse to Behaviorism, as the later Wittgenstein recognized. The truth in Behaviorism is captured adequately by our Essential Embodiment Thesis together with our Essentially Embodied Agency Theory of action.

Unintentional Body Movements vs. Pre-Reflective Intentional Body Movements

Now back to unintentional body movements. Certainly the crime-initiating man who spills his glass at the party by a deviant causal chain does not carry out an intentional body movement, precisely because he is not actually *trying* to spill his glass at that time, even though he otherwise wants to spill his glass and believes that by doing so he will start a crime that he also endorses. The body movement of spilling is caused by an accident of his psychology and his neurobiology. So the drink spiller's overt body movements are events that *just happen to him*, and are not his own. They

[15] See, e.g., Putnam, "Brains and Behavior." [16] See Kim, *Philosophy of Mind*, ch. 2, esp. p. 38.

[17] In contemporary theories of concepts, a "stereotype" is a mental representation, ancillary to a conceptual content, that captures some of the most typical features of instances of that concept in a shorthand format for purposes of easy recognition, but which does not uniquely determine the extension of that concept. Some contemporary cognitive psychologists claim that there are in fact no real concepts in the classical sense, but instead only stereotypes. See Margolis and Laurence (eds.), *Concepts: Core Readings*.

are not under his control, precisely because he does not try to make them, nor does he actively guide the neurobiological processes that give rise to his body movements. In this respect he is relevantly similar to someone whose overt body movements have been caused by his tripping over a curb in the dark, by a train suddenly starting up, by a Dr Strangelove-like spasm, or by a neuroscientist (say, Wilder Penfield) electrically stimulating his brain. On the other hand, as we noted earlier in Section 3.5, even though the drink spiller's body movements are unintentional, nevertheless moral responsibility for those body movements would still be correctly ascribed to him, since he did in fact intend to spill his drink in order to start a crime (although not at that very moment, but later), and since the neurobiological processes that caused his drink spilling did in fact occur inside his own body. This shows that conditions for the possibility of intentional action, and conditions for the possibility of *judgments* about moral responsibility, may sometimes come apart. It is possible to correctly judge someone to be morally responsible for a certain body movement (or for the consequences of that body movement) even if that body movement is strictly speaking unintentional and so strictly speaking not part of an intentional act. And that point, in turn, has an important bearing on current debates about the relation between free will and moral responsibility.[18]

In any case, it is very important to note that the mere fact that it is possible for me to *believe* mistakenly that I am *not* causing my own body movements—as in the schizophrenic delusion that I am a puppet of some evil alien or super-scientist—and also the mere fact that it is also possible for me to *believe* mistakenly that I *am* causing my own body movements—as in hallucinations of movement with paralyzed or phantom limbs, or when a scientist is covertly electrically stimulating my brain—are both orthogonal to the real agent-centered fact of intentional or unintentional movement. Intentional action is fundamentally manifest in pre-reflectively conscious, non-conceptual, effective first-order desires, or *willing*, rather than in belief-infused or concept-infused desires. So illusions that affect the agent's *beliefs* about himself do not necessarily determine the agent's will. We will come back to this crucial point in Sections 4.2 and 4.3, and again in Section 5.3.

[18] See, e.g., Frankfurt, "Alternate Possibilities and Moral Responsibility"; and Kane (ed.), *The Oxford Handbook of Free Will*, parts IV and V.

By sharp contrast to unintentional movements, mental causation is indeed at work in cases of pre-reflective or spontaneous actions. This includes both what we will now call the

(a) ''aimless'' pre-reflective or spontaneous actions mentioned in Section 3.4, motivated by desire-based emotions grounded in primitive bodily awareness (e.g., idly wiggling one's toes while reading, humming that faintly annoying 1970s commercial jingle, or suddenly frowning when the sun goes behind the clouds),

as well as the

(b) ''impulsive'' pre-reflective or spontaneous actions also mentioned in Section 3.4, motivated by desire-based emotions grounded in addiction, habit, passion, desire-overriding second-order volitions, or just the intrinsic desire to express one's own desires in body movements.

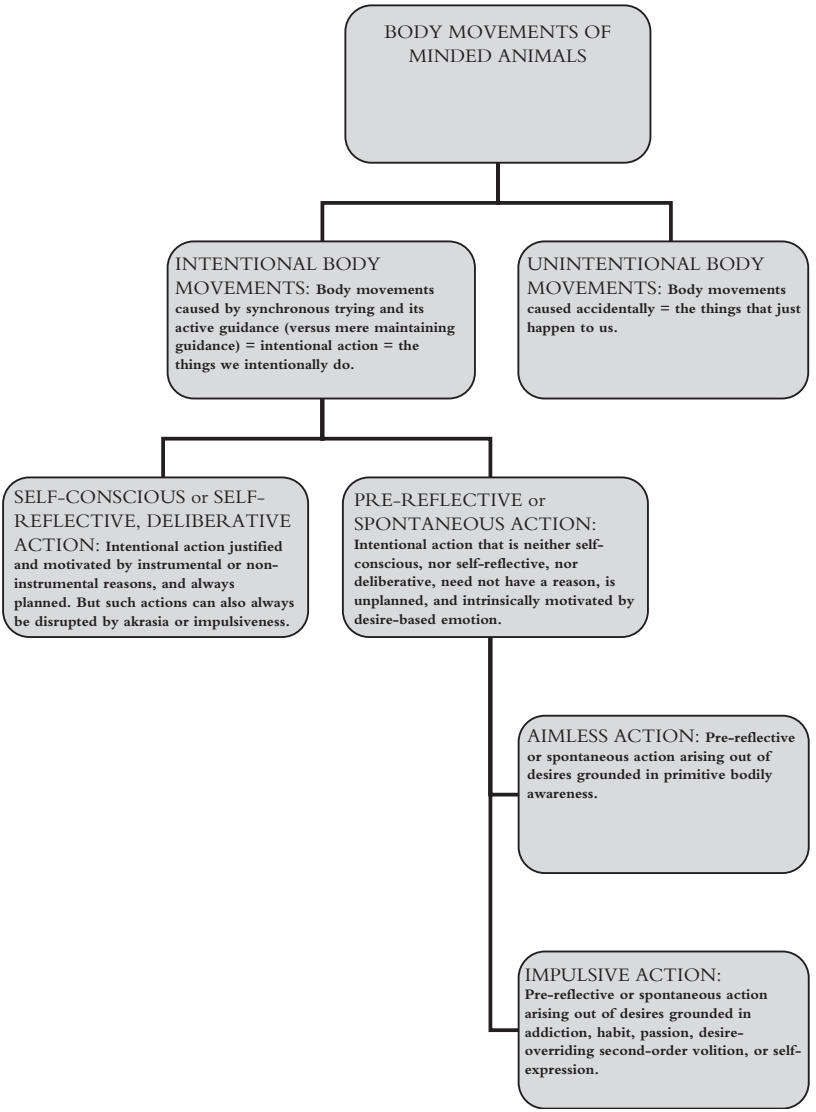
For example, consider Sally's thoughtless fingernail biting as she struggles to balance her checkbook. In this case Sally's behavior clearly does not result from reasoned desires, but does it exhibit trying and its active guidance? To understand the 'because' in ''Sally bit her fingernails because she wanted to,'' we will not want to refer to primary reasons or say that her behavior occurred as a direct result of something she self-consciously or self-reflectively and deliberatively desired and believed. Indeed, we can suppose that upon reflection Sally does not even *like* to bite her fingernails. Her desire to bite her fingernails just pre-reflectively or spontaneously thrusts itself forward by an impulse, in somewhat the way that an alcoholic's thirst just thrusts itself forward. Nevertheless, because Sally is clearly *trying* to bite her fingernails, this is not something that is *just happening to her*. Sally is not obsessive-compulsive—and if she were, then she would bite her fingernails (wash her hands, twist her hair around her fingers, scratch her arms, etc.) without even *trying* to do so, precisely by being *internally driven* and *forced* to do so, which of course would make her body movements unintentional. And it may well also be true that at least some of the characteristic body movements of at least some alchoholics or drug addicts are unintentional movements in this sense. But in the case as we have described it, we do want to say that these body movements are under Sally's control and that they cohere in creating a meaningful pattern. So too the characteristic body

movements of many alcoholics and drug addicts are under their control and cohere in creating a meaningful pattern. So as long as their body movements include a synchronous trying and its active guidance, then those movements will count as intentional acts.

Notice, however, that since the conditions of the possibility of intentional action are generally distinct from the conditions of the possibility of correct moral responsibility judgments, it does *not* automatically follow that an alcoholic's or drug addict's intentional acts are something we should hold them morally responsible for. Just as sometimes we correctly hold people morally responsible for bad things they did *not* actually intentionally do (e.g., the drink-spilling robber), so too sometimes we correctly release people from moral responsibility for things they *did* actually intentionally do, especially in cases of aimless or impulsive pre-reflective or spontaneous action.

For our purposes here, the crucial point is that pre-reflective or spontaneous aimless or impulsive actions are very different in both their basic structure and their action-theoretic implications from actions in which individuals do things as a result of deliberatively and self-consciously or self-reflectively intending or planning to do them, and also quite different from actions that can be rationalized in light of a Davidsonian primary reason. Indeed, by isolating the class of pre-reflective or spontaneous aimless or impulsive actions, we are able to pick out an essential feature of all intentional action. In the next chapter, Chapter 5, we explicitly extend the Essentially Embodied Agency Theory to include the thesis that it is essentially embodied *desire-based emotion*, rather than belief, prior intention, or judgment, which plays the primary causal role in bringing about an action.

We have seen that classical causal theories cannot account for the categorical difference between intentional actions and unintentional body movements (e.g., arm-raisings and mere arm-risings), because they cannot explain why body movements that are unintentional are not caused in the right way. Classical causal theories also overlook the important distinction between self-conscious or self-reflective, deliberative action and spontaneous or pre-reflective actions. Frankfurt's notion of guidance, when extended to active guidance, and when interpreted as synchronous mental causation, helps us to understand just how even spontaneous or pre-reflective intentional actions differ categorically from unintentional body movements and just what sort of fundamental connection obtains



Body Movements of Minded Animals: Six Varieties.

between an agent and her intentional body movements. But the crucial element in active guidance, as we have seen, is *trying*. So in the next section, we look more directly at trying in order to get even clearer about how agents synchronously cause their own intentional body movements in basic intentional acts. But before we move on, just to keep things orderly, we have also provided a chart on p. 174 that summarizes the basic claims and distinctions we have been making so far.

4.2 Towards a Non-Classical Causal Theory 2: Effortless Trying

In the case of unintentional body movements, what is lacking? We have argued that what makes the unfortunate climber's dropping his climbing partner or the robber's spilling his glass unintentional body movements (even if we *can* correctly ascribe moral responsibility for these movements) is the fact that those body movements are only *accidentally* caused. In this respect, the unfortunate climber and the drink-spilling robber are just like people who trip over curbs in the dark, lose their balance when a train starts moving, suffer Dr Strangelove-like spasmodic arm-risings, or make movements because their brains have been electrically stimulated by a neuroscientist. These are all things that just happen to those agents. We have also argued that what determines the difference between intentional movements and unintentional movements, and what would thereby guarantee that the former kind of movements are all caused in the *right* and *non-accidental* way, is the agent's synchronous trying and its active guidance.

Trying-Based Volitional Theories of Action and O'Shaughnessy's Account

An emphasis on trying and the will might initially point the action-theorist in the direction of a classical volitional-causal theory of action. And anyone who accepts a causal theory of action, even a non-classical causal theory like ours, is likely to be at least *somewhat* sympathetic to volitional-causal theories, for such theories provide a causal factor that "unifies actions in terms of a common kind of origin."[19] Volitionism in general is the

[19] Audi, *Action, Intention, and Reason*, 79.

doctrine that intentional actions, by their very nature, are constituted either partially or wholly by volitions or willings.[20] Such theories claim that volition or willing is a special kind of mental event that involves the essential property of being active. These events act as the immediate mental cause of the body movements involved in intentional action and serve as a special link between the agent's mental states and body movements, so that the volition's essential activeness is passed along to the resulting body movements.

One important sub-class of volitional-causal theories is the class of the trying-based theories, which identify psychologically and metaphysically independent mental trying events as the antecedent causes of action. What Timothy Cleveland calls "new-wave" volitional accounts focus on trying and also assert the following two claims:

1. Every physical action at least partially involves or is constituted out of a trying. Whenever one acts, one tries. (The Ubiquity Thesis)
2. Trying is the special kind of psychological event that gives rise to body movements.[21]

Body movements suitably brought about by temporally antecedent tryings are voluntary actions, while those that are not are mere happenings or mere reflex behavior. Insofar as tryings are mental efforts or exertions that insert energy into the action system,[22] the fact that an agent tried at some earlier moment causally explains why a bodily movement occurs here and now.

In his classic paper, "Trying (As the Mental 'Pineal Gland')," O'Shaughnessy quite accurately says that the mental event of trying

> consists in doing, intentionally and with just that purpose, whatever one takes to be needed if, the rest of the world suitably cooperating, one is to perform the action.[23]

O'Shaughnessy describes trying as an independent psychological event, that is, one that has both distinct causes and effects and is located at a distinct point in space and time. He also thinks that trying qualifies as an

[20] Cleveland, *Trying Without Willing*, 4.

[21] Cleveland classifies the theories of both Jennifer Hornsby and O'Shaughnessy as "new-wave" volitional theories.

[22] Audi, *Action, Intention, and Reason*, 83. Note that this is only Audi's depiction of volitional theories—he himself does not endorse such a view.

[23] O'Shaughnessy, "Trying (As the Mental 'Pineal Gland')," 56.

action, albeit an internal one, because it displays features typically associated with action. Such features include its being an event whose occurrence comes as no surprise to the subject, which happens because he chose and intended it should, and whose origin lies in his reasoned desires. According to O'Shaughnessy, cases of basic action or intentional body movement are the paradigmatic cases of trying. Trying to wiggle one's toes, e.g., does not consist in one's first performing some action or actions that are instrumental to toe-wiggling.[24] While someone can try to start the car by putting the key in the ignition, one cannot try to wiggle one's toes by doing something else, but rather does so directly, via some inner or conscious mental effort or exertion.

Again according to O'Shaughnessy, while it is true that the mental event of trying to raise one's arm and the physical event of arm-rising typically occur *almost* simultaneously, when trying occurs without its corresponding overt body movement, it is then clear that it is an independent mental event. This point seems to follow directly from cases of basic-act deception.[25] For example, suppose that upon being asked to raise his arm, a blindfolded patient believes he has succeeded, but later discovers that his arm is paralyzed. Though the agent has failed to raise his arm, he is nonetheless aware of having *tried* to do so. This mental exertion, had it been unimpeded, would have resulted in an overt body movement. In this way, on O'Shaughnessy's view, trying, as an independent mental event, causes a later body movement of arm rising, which is a distinct individual physical event. He claims that law-like psychophysical generalizations exist to cover this "causally linked pair of events," and characterizes trying to raise one's arm as "an *X* which in the state of psychophysical normality, world permitting, is sufficient to cause arm rising."[26]

Trying: Effortful and Effortless

In short, the intentional agent must always *try* in order to act intentionally, as O'Shaughnessy insists—but what precisely does this mean, and how can this claim be substantiated? Four problems arise immediately.

[24] Green, "Toe Wiggling and Starting Cars: A Re-examination of Trying," 173.
[25] Ginet, *On Action*, 28.
[26] O'Shaughnessy, "Trying (As the Mental 'Pineal Gland')," 52 and 65.

First, one might easily be tempted to reduce trying to an "immediate intention," or an intention to do something right now.[27] Searle's intentions-in-action, for example, might be characterized as tryings. But we have already seen that talk about "intentions" in the standard action theory literature can deeply obscure crucial differences between types of action. In particular, as we have seen, it is possible to do something *intentionally* yet not as the result of a *self-conscious or self-reflective, deliberative intention*. So, in a seeming paradox, it is possible to do something intentionally (in one sense of 'intentionally') but also not intentionally (in another sense of 'intentionally'). But we can avoid all such paradoxes, since on our view it is specifically trying and its active guidance of body movements, rather than an intention per se, that is the basic causal factor in bringing about action.

Second however, as Wittgenstein correctly notes, "When I raise my arm I do not usually *try* to raise it."[28] This is a perfectly correct observation, in that for many or even most intentional movements, e.g., raising my arm in an ordinary context, I do not have to make any *special* mental effort or exertion to try to raise it. Such a special mental effort or exertion might be needed if my arm were very stiff and sore, or if my arm were annoyingly caught up in my sweater, or if someone else were holding my arm down, or if I were trying to raise my arm in a hot and claustrophobic isolation cell whose ceiling is only a few inches above my head—as, e.g., in David Lean's epic 1957 war picture, *Bridge on the River Kwai*. Nevertheless, most arm-raisings are either fully pre-reflective or spontaneous, or at least relatively un-selfconscious, non-intense, and smooth, and we do not have to engage in a big-T "Trying" to do them in the sense of a specially effortful, intense, jerky trying. Nevertheless, since it is *always* at least possible for me to find that my arm is temporarily paralyzed—say, because I have been lying on it oddly while asleep—and since *whenever* I could encounter such paralysis it would be true of my act in that case that I had *tried* to raise my arm but failed, then it seems to follow necessarily that some sort of trying must *always* be present. We will come back to this line of argument later.

Right now, however, we need to distinguish between the *effortful* trying that Wittgenstein is isolating, and what we will call the *effortless* trying

[27] Audi, *Action, Intention, and Reason*, 76.

[28] Wittgenstein, *Philosophical Investigations*, 161, § 622.

that causes all intentional action. We borrow the important idea of an "effortless" mental directedness to basic action from David Velleman, who in turn draws upon the Daoist doctrine of *wu wei* and Mihaly Czikszentmihalyi's theory of "flow."[29]

The crucial point for our purposes, however, is that the phenomenology of *effortful* trying is that of an intense, jerky mental exertion. Effortful trying is *stressful*. Effortful trying is a *special*, non-commonplace kind of trying. By sharp contrast, the phenomenology of *effortless* trying is identical with the phenomenology of pre-reflectively conscious effective first-order desire, and so cannot be distinguished from any particular pre-reflectively conscious effective first-order desire for this or that, or to do this or that. Effortless trying is *not* a paradoxical "trying without really trying." On the contrary, effortless trying is *really* trying, but without necessarily including any *effortful* or *special* trying. So effortless trying is just *commonplace* trying. Like all forms of pre-reflectively conscious intentionality, a pre-reflectively conscious effective first-order desire may *also* be more or less self-conscious or self-reflective, more or less deliberative, more or less conceptually-determined, and more or less intense. But although there are as many different ways to engage in pre-reflectively conscious effortless trying as there are forms of pre-reflectively conscious effective first-order desire, effortless trying is particularly evident in pre-reflective or spontaneous action. In such cases, the special phenomenal character of pre-reflectively conscious effortless trying manifests itself as the subjective experience of *flowing forward* ***right into*** *intentional body movement*, as in Yeats's dancer *becoming* her dance. In this sense, pre-reflectively conscious effortless trying is *always* present in intentional action, *even* in cases of effortful trying, which is always just a complexification and an intensification of pre-reflectively conscious effortless trying. For example, as you effortfully try to raise your sore arm, you also effortlessly try to balance and orient the rest of your body. So pre-reflectively conscious effortless trying and its active guidance is the *default setting* and *normal* cause of intentional action, and effortful trying is relatively rare and special.

Third, even if, for the purposes of argument, we assume the general validity of the distinction we have made between pre-reflectively conscious effortless trying and effortful trying, must *every* basic intentional

[29] See Velleman, "The Way of the Wanton."

act and therefore *every* intentional body movement then involve either a pre-reflectively conscious effortless trying alone or also an effortful trying? Critics have argued that tryings *of any sort* simply are not needed in order to account for the origin of intentional actions.

One common view is that "suitably qualified intentions" are sufficient to bring about action.[30] An agent has this sort of self-fulfilling intention if she has made up her mind what to do, has not changed her mind about what to do, is neither confused nor forgetful, and is not prevented by lack of ability or other external circumstances. Because these full and present intentions are sufficient for action and require no further impetus, then acts of will or tryings are "otiose."[31] Indeed, for many action theorists, the claim that trying is present in all instances of bodily action is just plain counterintuitive. In part, this is because it is self-evident to them that one can act without the feeling that one has made any notable effort or exertion. Since many ordinary intentional body movements do not involve any notable difficulty, it may seem implausible to suppose that any trying has taken place.[32] For example, Robert Audi points to other possible causes of actions, such as perceptions, thoughts, decisions, resolutions, and changes in the balance of an agent's motivational forces. He thinks that while agents may try when they encounter resistance, there is no reason to suppose that trying is a basic element that serves as the foundation for every action.[33] After all, spontaneous or pre-reflective actions do not seem to *need* any exertion.

We think that this argument is unsound, for two reasons. First, we think that it depends precisely on *not* having made the distinction between effortless trying and effortful trying, and in mistakenly inferring that trying is not required for every intentional action just because *effortful* trying is not required for every intentional action.

Second, we also think that it depends on a fallacious inference from a linguistic fact. Attention to ordinary language use shows us that one can be normally correctly be said to "try" only if one has some *doubt* about whether one will succeed. For example, it would be odd and perhaps also misleading to say that Sally *tried* to move her finger if we had every reason to believe that she easily and effortlessly moved her

[30] Green, "Toe Wiggling and Starting Cars: A Re-examination of Trying," 178. [31] Ibid., 180.
[32] Cleveland, *Trying Without Willing*, 28. [33] Audi, *Action, Intention, and Reason*, 99.

finger. It then may seem that only *non-volitional* event causes can account for the execution of intentions and explain why movements occur at a particular time.

But this is a non sequitur. We agree completely that we would not ordinarily say, without deviating into linguistic oddness or being misleading to our interlocutors, that an agent whom we all believed just easily and effortlessly performed an act specifically *tried* to perform that act. Nevertheless, this is only a point about how we normally *talk* about trying and action in terms of trying, and correspondingly only a point about *conversational norms and conversational implicature*, and not a point about trying and action themselves. Normally, it is only if the fact of act-failure, or at least the possibility of act-failure, is salient (even if it is not actually expected, or feared) in some speech context that we specifically speak of *trying* to do *X*, as opposed to just doing *X*. Intentional action and especially intentional body movement is inherently success-oriented, so successful action is the norm, and thus deviations from it must be specially marked by some form of speech, which is *trying*-talk. But talk is one thing, and the concepts, properties, things, and facts expressed or described by talk are quite other things. Rules of talk are not rules of reality. So the pragmatics of trying-talk does not undermine our thesis that effortless trying is present in all cases of intentional action.

On our view, as we have said, necessarily *all* cases of intentional action involve effortless trying, even those cases that also involve effortful trying. This can be shown phenomenologically by the fact that even in cases of effortful trying, the phenomenological character of "flowing forward right into intentional body movement" is never *entirely* lacking. For example, suppose that you are walking on a sore right leg, and every step taken on that leg is the result of an effortful trying. You are concentrating on moving your sore leg, and it currently occupies the central focus of your intentional activity. Still, and necessarily, there are also going to be other elements of that intentional body movement that are effortlessly chosen and done by you, such as the swinging of your arms, the motion of your head, and the movement of your other leg. Effortless trying is just the constant background hum of the pre-reflective conscious workings of an intentional agent at the foundations of all her basic actions.

The ubiquity of trying, whether effortless or effortful, can also be shown by the following a priori argument, which we have already briefly

mentioned in passing at the beginning of this section in our discussion of the arm-paralysis case:

(1) Every intentional act either fails or succeeds.
(2) Suppose an intentional act fails. Then it is true of that act that the agent *tried* but failed. So the agent tried.
(3) Suppose an intentional act succeeds. Nevertheless, it still might have failed. For imperfect intentional agents like us in an imperfect world, whenever an intentional act begins, and even when agents acts effortlessly, there is always a logical, metaphysical, and nomological possibility, however minimal, that the act will not succeed. For the neurobiological facts, overt bodily facts, or the external world simply might *not* cooperate with the agent's effective first-order desire. If the act *had* failed, then the agent *would have* tried but failed. By hypothesis the act succeeds. So the agent tried *and* succeeded.
(4) Therefore every intentional action involves a trying, whether effortless or effortful.

Or in other words, even if the world inside us and outside us is in some sense necessitated or determined by logic or the laws of nature, that world is nevertheless clearly somewhat contingent and undetermined in relation to our effective first-order desires, and in relation to all our hopes, fears, and dreams, precisely because the world is obviously not fully necessitated or determined by our effective desires, hopes, fears, and dreams. The actuality and possibility of failure, frustration, pain, and suffering are all too obvious. Life is nasty, brutish, and short. The world is a vale of tears. Life's but a walking shadow, a poor player that struts and frets his hour upon the stage and then is heard no more. Life's a tale told by an idiot, full of sound and fury, signifying nothing. We are to the gods as flies to wanton boys. You can't always get what you want. Stuff happens. And so on. Let us call this the world's *desire-contingency*. Now because the world is desire-contingent, and because intentional action necessarily requires the *cooperation* of this desire-contingent world in our agency, then the possible *non*-cooperation of the desire-contingent world entails that we *must* try whenever we act. For it takes two to cooperate, and, sadly, sometimes, no matter how hard we try, the world is just going to thwart us. So our unique and ineliminable contribution to the necessarily cooperative relation between agency and world, whether we succeed or not, is effortless trying. Essentially embodied

desire-based effortless trying and its active guidance is nothing more and nothing less than *our* role in the necessarily cooperative relation between agency and world that constitutes intentional action.

Fourth and finally, as Wittgenstein also correctly notes, "I can always will only inasmuch as I can never try to will." And this in turn is because

> I can't will willing; that is, it makes no sense to speak of willing willing. "Willing" is not the name of an action; and so not the name of any voluntary action either.[34]

In other words, it is an important mistake to think of trying as an independent intentional act that somehow has to be brought into being by another prior mental act of trying, on pain of infinite vicious regress. On the contrary, on our account, effortless trying is just the *same* as willing, and willing is nothing more and nothing less than an essentially embodied pre-reflectively conscious effective first-order desire that actively guides intentional body movements. So where intentional action is concerned, there is simply nothing *behind* our essentially embodied effortless trying and willing. Effortless trying or willing is just the unprecedented *ground, origin, or source of basic intentional acts*, a pre-reflectively conscious actively intervening mental cause that is also a synchronous active guide of covert neurobiological processes and overt intentional body movements alike. Effortless trying is what is ultimately *up to me*. Wittgenstein beautifully captures this thought too:

> One imagines the willing subject here as . . . a motor which has no inertia in itself to overcome. And so it is only mover, not moved.[35]

Again, effortless trying is the *unmoved motor* of action, or what is ultimately up to me, but *not* because it exists as a noumenal person-substance outside of time, as in classical agent causation. Rather, as essentially embodied, effortless trying is the online efficacious *singular event-cause* of intentional body movements, and it occurs only *in-and-through* the neurobiological processes that fully embody it.

Essentially Embodied Trying

Obviously, the metaphysical mistake about trying and willing that leads to a vicious regress of distinct antecedent tryings-to-try or tryings-to-will is

[34] Wittgenstein, *Philosophical Investigations*, 161, §619. [35] Ibid., 160–1, §618.

closely connected with classical volitional-causal theories of action, which we have described above.

But as this chapter and Chapter 3 should now have shown, our non-classical trying-based volitional causal theory of action is both similar to and yet also crucially different from all classical causal-volitional theories of action, including of course trying-based classical causal-volitional theories. The similarity is the appeal to the phenomenology of trying and the mentalistic fact of the will. For us, trying or the will is an essentially embodied pre-reflectively conscious effective first-order desire (i.e., a pre-reflectively conscious desire that moves or would move or will move us all the way to action) occurring in a reflexive hierarchy of desires (i.e., a hierarchy of higher-order desires about lower-order desires). Needless to say, not every volitional theory identifies trying with pre-reflectively conscious effective first-order desire, nor does every volitional theory accept the hierarchical desire theory of the will. But the crucial difference between the Essentially Embodied Agency Theory and all other trying-based volitional-causal theories is our further pair of theses, both of which should be quite familiar by now, to the effect that

(1) conscious, intentional minds_{lo} are *essentially embodied*,

and

(2) trying (whether effortless or effortful) and its active guidance (which can also modulate in and out of merely maintaining guidance) is *synchronous* with the entire intentional action, including all the covert neurobiological body movements and overt body movements that necessarily and completely embody our conscious, intentional agency.

The conjunction of these claims avoids not only the universal problem for classical causal theories—deviant causal chains—but also avoids any sort of dualistic or supervenience-based metaphysical gap between the irreducible intentionality of the minded animal and its intentional body movements. The other basic problems of the specifically *Davidsonian* classical causal theory of action are avoided, first, by our Desire-Overriding Internalism about reasons (Section 3.4), and second, by our Emotive Causation Thesis, which says that trying and its active guidance is primarily a pre-reflective, desire-based emotive mental activity and only derivatively

a self-conscious or self-reflective, deliberative intellectual mental activity (see Chapter 5).

In this way, while we, like O'Shaughnessy, are "new wave volitionalists," our notion of trying nevertheless differs in certain important respects from O'Shaughnessy's brilliant and seminal analysis of trying. Sometimes O'Shaughnessy's theory oddly merges with classical trying-based volitional-causal theories, and characterizes trying as an independent mental event that expires before the physical event of body movement begins. By sharp contrast, we understand trying to be the irreducible mental *aspect* of an essentially embodied and therefore *essentially mental-and-physical* causal-dynamic living organismic process that therefore also has an irreducible physical aspect (see Section 7.1). So, for us, trying is an irreducibly mental event whose fundamental properties are *fused* with the fundamental physical properties of a physical event (see also Section 7.1) in the living organismic life of a certain animal, namely the intentional agent herself. On our view, an intentional action begins with an actively intervening trying that is synchronous with a neurobiological process that it actively guides—and, parasitically on that active guidance, can also temporarily maintainingly guide—until this neurobiological process completely manifests itself as an overt intentional body movement and the act is thereby successfully completed.

We think that two things, in particular, have gone wrong with O'Shaughnessy's otherwise excellent analysis.

The first error is that he fails to distinguish between, first, *neurobiological* body movements that are internal to the organism and occur in the dynamic region from the vital organs out to the muscle tissue/skin interface, and second, *overt* body movements that begin at the muscle tissue/skin interface and then extend outwards into the external world. While it is true that there is a real time-lag between the beginning of the neurobiological process that embodies trying, and the occurrence of the overt body movements that arise from this process and accompany it, this is *not* a time-lag during which and through which an earlier trying-event causes a later overt body movement. On the contrary, during this entire time trying and its active guidance is synchronous with a neurobiological process whose latter phases are also accompanied by overt body movements that arise from that very process. So in a successful intentional performance there is *never* a time during which trying and

actively guided intentional body movements are not *both* simultaneously occurring. In Sections 8.1 and 8.3, we will show how trying actively guides both neurobiological and overt body movements by means of *structuring causation*.

The second error in O'Shaughnessy's analysis is his interpretation of the case of the blindfolded man whose arm is paralyzed, and who mistakenly thinks he has raised his arm. According to the Embodied Agency Theory, what has happened in this case is that trying is essentially embodied in a neurobiological process that has been causal-dynamically disconnected by the paralysis (which could of course have different kinds of causes in different cases, e.g., brain trauma, stroke, poison, etc.) from the overt body movements that normally arise from it and accompany it. The beginning of the synchronous trying could be indicated by neural imaging, but then the normal causal-dynamic neurobiological connection between the vital organs and overt body movements at the muscle tissue/skin interface is disrupted by whatever is causing the paralysis. So the trying fails *not* because it is an independent mental event that fails to hook up causally in the right way with a later purely physical event, but instead because its essential embodiment in neurobiological movements has a causal-dynamic pattern that is in fact abnormally different from the one that would normally produce overt body movements. This is the case, even though the phenomenology of the paralytic embodied trying process is epistemically indiscriminable for the subject herself from the phenomenology of the embodied trying process in an actual arm-raising. Or in other words, the subject cannot *tell* the difference between the two cases, and thus can be deceived. But they are categorically causal-dynamically different cases nevertheless.

This element of first-person epistemic indiscriminability suggests an important parallel between O'Shaugnessy's use of the act-deception case, and classical indirect or non-relational (e.g., imagist, sense-datum, or intentionalist) theories of perception that postulate an intervening mental image, sense-datum, or mental content to explain the common factor across correct, veridical perception on the one hand and hallucinations on the other. For O'Shaughnessy, trying is the analogous common factor across real intentional acts and act-deceptions. But by sharp contrast, our approach to this issue closely resembles *direct realist disjunctivism* in the philosophy of perception, which says:

(i) that sense perceptions are all correct and veridical, and categorically different from imaginative or hallucinatory illusions,
(ii) that correct, veridical perception and illusion share only whatever is needed for the possibility of their first-person epistemic indiscriminability,

and

(iii) that correct, veridical perception is an unmediated, relational, sensorily conscious, intentional openness to the real objects of the external world and their properties.[36]

Correspondingly then, our direct realist disjunctivism in action theory says:

(i*) that intentional actions are all real, and categorically different from act-deceptions,
(ii*) that real intentional action and act-deception share only whatever is necessary for the possibility of their first-person epistemic indiscriminability,

and

(iii*) that real intentional action is a set of neurobiological processes and overt body movements that essentially embody synchronous trying and its active guidance.

In this way, although trying occurs in both intentional action and act-deception cases alike, just as sensory consciousness occurs both in sense perception and illusion, they are nevertheless categorically different types of trying, since their essential embodiment is categorically different. Successful trying is categorically different from failed trying, and both are again categorically different from mere non-performance (neglecting or refusing to act),[37] precisely because the phenomenological and causal-dynamic profiles of our essential embodiment are intrinsically different in each case. *Successful* trying is essentially embodied in a neurobiological process from which overt body movements arise that are the intentional targets of the conscious, intentional effective first-order desire at the basis of the trying. *Unsuccessful* trying, by contrast, is embodied in a causal-dynamically

[36] See, e.g., Gendler and Hawthorne (eds.), *Perceptual Experience*, esp. chs. 3, 7, and 10.
[37] Audi, *Action, Intention, and Reason*, 91.

distinct neurobiological process that intrinsically *lacks* the overt body movements targeted by the conscious, intentional effective first-order desire at the basis of the trying. And *mere non-performance*, because it lacks any trying or essentially embodied effective first-order desire, and thereby also lacks any overt body movements targeted by such a desire, is both phenomenologically and causal-dynamically distinct from both successful trying and unsuccessful trying.

Moreover, given the essential embodiment of conscious, intentional minds_{lo}, we would also predict that in arm-raising act-deception cases the subjective experience of the failed essentially embodied trying process—understood as primarily manifest in desire-based emotion and originally given in primitive bodily awareness—is also in itself *sharply phenomenologically different* from the phenomenology of the essentially embodied trying process in an actual arm-raising. Given the Deep Consciousness Thesis, moreover, this phenomenology will be sharply different in a *pre-reflectively conscious* or sensorimotor-subjective way, even if the subject cannot *self-consciously or self-reflectively* discriminate between the two. If the subject forms a self-conscious or self-reflective belief or judgment about the two cases, then she may not be able to find a discriminable difference, and so can be fooled. But just because one can be fooled by a certain type of experience, or indeed even *constantly* fooled by a certain type of experience, it does not follow that the misleading experience is phenomenologically like its real counterpart, *except* in the superficial respect that remains stubbornly resistant to self-conscious or self-reflective discrimination. Indeed, the fact of "change-blindness" or "difference blindness"—e.g., our inability to notice the difference between two complex pictures, scenes, or sequences of sounds presented at different times, one of which in fact contains some extra colors, objects, shapes, or sounds folded cleverly into the overall pattern; or one of which lacks some colors, objects, shapes, or sounds that the other includes—is both well-documented by cognitive psychologists[38] and increasingly noted by philosophers of mind,[39] and of course the whole art of magic or illusionism is entirely based on this fact.

[38] See, e.g., O'Regan, Rensink, and Clark, "Change Blindness as a Result of 'Mudsplashes' "; Rensink, O'Regan, and Clark, "On the Failure to Detect Changes in Scenes Across Brief Interruptions"; and Simons and Levin, "Change Blindness."

[39] See, e.g., Dretske, "Change Blindness"; and Noë, *Action in Perception*, 51–3.

In this way, act-deception cases will necessarily include misdirecting features that can trigger a temporary state of *act-intentional change blindness or difference blindness* in the agent. But given the essential embodiment of consciousness$_{lo}$, it necessarily will be the case that in comparison with a successful trying to (say) raise one's arm, a failed trying to raise one's arm will be a very affectively etiolated, disconnected, and hollow pre-reflectively conscious and sensorimotor-subjective experience at the level of primitive bodily awareness, precisely because in such experiences we are necessarily alienated in certain definite ways from our own bodies and their neurobiological and overt movements. Correspondingly, a successful trying to raise one's arm will also have to be *a much richer subjective experience at the level of pre-reflectively conscious primitive bodily awareness* than a failed trying to raise one's arm, even if we cannot help being fooled by act-deception cases.

The same point would hold, mutatis mutandis, for the phenomenology of illusions on the one hand, and the phenomenology of correct, veridical perception on the other. Given the essential embodiment of perceptual consciousness$_{lo}$, the phenomenology of *imagining or hallucinating* (say) a dagger seen before you must be a very sensorily etiolated, disconnected, and hollow subjective experience at the level of pre-reflectively conscious primitive bodily awareness compared to actually *seeing* a dagger before you. And correspondingly, a correct veridical perception of a dagger seen before you must be a much richer subjective experience at the level of pre-reflectively conscious primitive bodily awareness than the illusion of a dagger seen before you, even if you are unable to discriminate between them at the level of self-conscious or self-reflective judgment or belief.

This in turn helps to substantiate the direct realist doctrine that correct, veridical perception is an unmediated, relational, sensorily conscious, intentional openness to real objects and properties in the external world.[40] Our perceptual openness to the world, just like our causally efficacious ability to make intentional body movements, is directly confirmed via a rich pre-reflectively conscious primitive bodily awareness. So if this line of reasoning is sound, then our essential embodiment approach to consciousness and intentional action also provides significant support for direct realist disjunctivism in action theory and the philosophy of perception alike.

[40] See, e.g., Campbell, *Reference and Consciousness*; and Johnston, "Better then Mere Knowledge? The Function of Sensory Awareness."

4.3 Is Trying an Epiphenomenal Illusion? *No.*

We will now finish up this chapter by addressing a serious worry, related to the possibility of act-deception, about all volitionalist approaches to intentional action—including ours. This worry derives from contemporary cognitive science, and more specifically from a series of famous (and controversial) neuroscientific experiments carried out by Benjamin Libet.[41] Here is the worry, as crisply and lucidly formulated by Daniel Wegner:

> The celebrated experiments of Benjamin Libet provide . . . evidence that conscious will can be experienced that does not correspond to causation. In spontaneous, intentional finger movement, Libet found that a scalp-recorded brain readiness potential (RP) preceded the movement (measured electromyographically) by a minimum of ~550 ms. This finding indicates only that some sort of brain activity reliably precedes the onset of voluntary action. However, participants were also asked to recall the position of a clock at their initial awareness of intending to move their finger, and this awareness *followed* the RP by some 350–400 ms. So, although the conscious intention preceded the finger movement, it occurred well after whatever brain events were signaled by the RP. This finding suggests that the experience of consciously willing an action begins after brain events that set the action into motion. The brain creates both the thought and the action, leaving the person to infer [unsoundly] that the thought is causing the action.[42]

Wegner's overall conclusion, based heavily on the Libet results, is that there is good reason to believe that the subjective experience of conscious willing is epiphenomenal and illusory, and that intentional action is instead caused solely by deterministic non-conscious brain processes.[43]

In reply to Wegner, what we want to argue is that the Libet findings, interesting and important as they are, do *not* in fact provide a sufficient

[41] See Libet, "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action"; Libet, Gleason, Wright, and Pearl, "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Unconscious Initation of a Freely Voluntary Act"; Libet and Haggard, "Conscious Intention and Brain Activity"; and Haggard, "Conscious Intention of Awareness and Action."

[42] Wegner, "The Mind's Best Trick: How We Experience Conscious Will," pp. 65–66.

[43] Ibid., 65 and 68. See also Wegner, *The Illusion of Conscious Will.*

reason for asserting that the subjective experience of conscious willing is epiphenomenal.[44] There are two simple reasons for this.

First, Wegner consistently fails to distinguish between

(i) someone's desire-based emotive pre-reflective sensorimotor-subjective consciousness of willing—that is, her effective first-order desire, or effortless trying, to make intentional body movements,

and

(ii) someone's higher-order, self-conscious or self-reflective beliefs about her own first-order consciousness of willing.

But it seems obvious to us that (i) and (ii) are sharply different.[45] Not only can an ordinary adult human intentional agent perform pre-reflective or spontaneous aimless or impulsive intentional acts without forming any higher-order, self-conscious belief about her own first-order conscious volitional states, but also there are many non-human animals and young human children who operate as conscious, intentional agents without even having a *capacity* for making higher-order, self-conscious or self-reflective beliefs about their own first-order conscious mental states. So if we are correct that it is effective first-order desires, i.e., pre-reflective effortless tryings, that are the primary and universal mental causes of intentional body movements, and *not* self-conscious or self-reflective, deliberative intentions, which are only secondary or derivative mental causes of action, then it is obvious that (i) does not entail (ii).

Going in the converse direction, and by Wegner's own admission, it is also possible for someone to have a higher-order, self-conscious or self-reflective belief that she is causing some overt movement of her own body, yet actually *lack* a corresponding pre-reflectively conscious sensorimotor-subjective effective first-order desire, or effortless trying, to make that movement.[46] Indeed, this is just another form of act-deception—i.e., act-intentional change blindness or difference blindness—but now at the level

[44] For a set of closely related critical responses to Libet's experiments and Wegner's interpretation of them, see also Pockett, Banks, and Gallagher (eds.), *Does Consciousness Cause Behavior?*, esp. essays 6–11 by Gallagher, Ross, Pacherie, Bayne, Mele, and Malle.

[45] See also Jeannerod, "Consciousness of Action as an Embodied Consciousness."

[46] See, e.g., Brasil-Neto, et al., "Focal Transcranial Magnetic Stimulation and Response Bias in a Forced-Choice Task"; and Wegner and Wheatley, "Apparent Mental Causation: Sources of the Experience of Will."

of pre-reflectively conscious sensorimotor-subjective willing, rather than at the level of intentional body movements. Just as you can mistakenly think that you are intentionally moving a limb that is in fact paralyzed, so too you can mistakenly think that you are consciously willing to move a limb that is in fact moved by a non-standard causal mechanism whose control panel and power source are both outside your own living body. So (ii) does not entail (i) either, and thus (i) and (ii) are mutually logically independent of one another.

Now it seems clear to us from Libet's descriptions of his experiments that what the subjects are reporting is *only their higher-order, self-conscious or self-reflective beliefs or judgments about* their first-order pre-reflectively conscious sensorimotor-subjective experience of effortless trying. Therefore Libet's time-delay data measure only the temporal difference between the onset of the readiness potential and *higher-order, self-conscious or self-reflective beliefs about* the subject's first-order pre-reflectively conscious effortless trying, *not* a temporal difference between the onset of the readiness potential and her trying. So it is perfectly possible to hold, consistently with Libet's results, that first-order pre-reflectively conscious effortless trying and the onset of the readiness potential are *synchronous*. And since the onset of the readiness potential precedes the beginning of the overt body movement by at least 550 milliseconds, this is also perfectly consistent with our thesis that a synchronous effortless trying and its active guidance are essentially embodied in the covert neurobiological processes that precede overt body movements. Therefore, on our view, a synchronous effortless trying is just the mental aspect of a causal-dynamically complex essentially mental-and-physical living organismic event consisting of a first-order desire-based emotive pre-reflective sensorimotor-subjective consciousness of mental directedness towards overt body movements, together with its essential embodiment in a neurobiological process and in the overt body movements that subsequently arise from this process and accompany it.

Second, granting our distinction between first-order pre-reflectively conscious willing, and higher-order, self-conscious or self-reflective beliefs about first-order conscious willing, Libet's time-delay data are then best explained *not* by Wegner's hypothesis, but instead by our alternative hypothesis. Our hypothesis says that from the moment of the synchronous beginning of first-order pre-reflectively conscious willing on the one hand and of the onset of the neural activity that is measured by the readiness

potential on the other, it takes between 350 and 400 milliseconds for the experimental subject to form the higher-order, self-directed, self-conscious or self-reflective psychological judgment to the effect *that* she is indeed trying to move her finger and also to correlate this judgment with her visual perception of the clock, and then yet another 200 milliseconds for the overt intentional finger movement to arise from the ongoing trying-guided neurobiological process that began between 550 and 600 milliseconds earlier.

So if we are correct, then in the experimental situation the subject causes her own finger to move by synchronously pre-reflectively effortlessly trying to move it and by actively guiding her finger movement from its covert neurobiological beginnings to its overt behavioral manifestation. Then, at least 350 milliseconds after the beginning of pre-reflective effortless trying and its active guidance and at least 200 milliseconds before the beginning of her overt intentional finger movement, she also manages to judge self-consciously or self-reflectively, for the benefit of the experimenter, *that* she is trying to move her finger, and also correlates this higher-order, self-conscious or self-reflective belief with her visual perception of a clock. This neatly explains Libet's time-delay data, and involves no appeal whatsoever to an "error theory" of conscious willing.

We conclude that the Libet experiments provide no sufficient reason for asserting that the subjective experience of conscious willing, or trying, is an epiphenomenal illusion. On the contrary, given the Essentially Embodied Agency Theory of intentional action, Libet's findings merely provide us with some important empirical information about how the intentional body movements of essentially embodied agents like us are efficaciously caused by our synchronous pre-reflectively conscious effortless trying and its active guidance.

This page intentionally left blank

5

Essentially Embodied Agency III: Emotive Causation

> The heart has reasons of its own that reason knows nothing about.
>
> Blaise Pascal[1]

> Dasein's Being reveals itself as *care*.
>
> Martin Heidegger[2]

> Caring, insofar as it consists in guiding oneself along a distinctive course or in a particular manner, presupposes both agency and self-consciousness. It is a matter of being active in a certain way, and the activity is essentially a reflexive one. This is not exactly because the agent, in guiding his own behavior, necessarily does something *to* himself. Rather, it is more nearly because he does something *with* himself.
>
> Harry Frankfurt[3]

5.0 Introduction

Here, again, is the big philosophical story we have been telling. The goal of this book is to present and prove the Essential Embodiment Theory of the mind–body relation, mental causation, and intentional action. The Essential Embodiment Theory says that creatures with conscious, intentional minds_{lo} are *essentially embodied minds*, or *minded animals*, which in turn are *self-organizing thermodynamic systems*. Our core ideas are

(1) that conscious, intentional minds_{lo} are the irreducible global intrinsic structures of motile, neurobiologically complex, situated, forward-flowing, living organisms,

[1] Pascal, *Pensées*, section 4, # 277, our translation. [2] Heidegger, *Being and Time*, H. 182, 227.
[3] Frankfurt, "The Importance of What We Care About," 83.

and

(2) that because organismic life is basically causally efficacious and $minds_{lo}$ are alive, then $minds_{lo}$ are basically causally efficacious too.

And if these core ideas are correct, then assuming favorable inner and outer natural conditions, we can intentionally move our own living bodies when we want to, which of course is just what we were trying to explain.

The Essential Embodiment Thesis emerged from Chapters 1 and 2, in which we argued from neurophenomenological premises that conscious, intentional $minds_{lo}$ are necessarily and completely neurobiologically embodied. In Chapters 3 and 4 we argued that the pre-reflectively conscious intentional activity of effortless (as opposed to effortful, self-conscious or self-reflective, deliberative) trying, together with its active guidance of body movements, is the mental cause of all basic intentional action, and that this fact adequately explains the difference between actions (the things we intentionally do) and unintentional body movements (mere bodily events, or the things that just happen to us). If so, then I act intentionally if and only if I move my own living body by pre-reflectively conscious effortless trying and its active guidance.

In this chapter, we argue that the subjective experience of effortless trying and its active guidance is grounded in *pre-reflectively conscious desire-based emotions*, rather than in beliefs, judgments, thoughts, or in self-conscious or self-reflective, deliberative intentions. As a result, whatever action-initiating and action-guiding causal powers that self-conscious or self-reflective, deliberative intentions have are derived from the action-initiating and action-guiding causal powers of pre-reflectively conscious desire-based emotions. If this argument is sound, then it establishes the Emotive Causation Thesis and completes the Essentially Embodied Agency theory of action.

In Chapters 6, 7, and 8 we will spell out the background metaphysics that supports this theory of action—in particular, the metaphysics of mental causation, mental-physical property fusion, dynamic systems and the dynamic world, non-logical or strong metaphysical a priori necessity, neo-Aristotelian hylomorphism, and dynamic emergence—and defend that metaphysics against some possible objections. That will bring us back full circle to where we began, with our robust pre-theoretic intuitions about essentially embodied agency and the categorical difference between, e.g., arm-*raisings* and mere arm-*risings*.

Of course you already *knew*, without ever having to philosophize about it, *that* you can intentionally raise your arm to wave to a loved one when you want to, other things being equal, and also *that* Dr Strangelove's spasmodic salute is a completely different kind of body movement, even if it is behaviorally indiscriminable from your loving wave. But by the end of this book, you will also know precisely *how* and *why*.

5.1 Essentially Embodied Agency and the Emotions

The Essentially Embodied Agency theory of action has three basic elements. The first basic element is that the essential embodiment of conscious, intentional minds_{lo} rules out any *metaphysical gap* between the mental causes of action and the intentional body movements that are their effects. This holds whether the metaphysical gap is generated by the substance dualism of agent-causal approaches, or by Jaegwon Kim's causal-explanatory exclusion problem for non-reductive materialist approaches that are based on supervenience.

The second basic element is that effortless trying and its active guidance is synchronous with all our intentional body movements, whether these movements are the covert neurobiological processes that necessarily combine with trying to begin a basic act, or the overt movements that arise from those processes and accompany them until that act is completed. The synchronicity of effortless-trying-and-its-active-guidance and body movements entails that no *temporal gaps*, into which deviant causal chains might be inserted, can ever arise between the mental cause and the physical effect in a basic act.

The third basic element of the Essentially Embodied Agency theory of action is that we understand pre-reflectively conscious effortless trying and its active guidance as a phenomenon of mental directedness that is identical to pre-reflectively conscious effective first-order desire. This in turn is the same as the act of *willing* in animals with $\text{consciousness}_{lo}$ and $\text{intentionality}_{lo}$. More precisely, on our view the will itself, considered as a psychological fact about animals minded like us, is nothing more and nothing less than a dynamic hierarchy of desires. The will is specifically a fact about *desires* because it bottoms out in pre-reflectively conscious effective first-order desires. The will is specifically a fact about a *hierarchy* of desires

because it is a structured complex of higher-order or reflexive desires, especially including second-order volitions, along with pre-reflective first-order desires. And the will is specifically a fact about a *dynamic* hierarchy of desires because it is actively configured by a pre-reflectively conscious, and sometimes also self-conscious or self-reflective, subject over time and inherently open to gradual or even radical re-configuration, as she continually "makes up her mind" and "changes her mind," and sometimes even transforms her will and thereby "changes her life." Here Augustine's *Confessions* provides probably the most brilliant and moving first-person narrative of a life-changing transformation of the will. But anyone who has ever permanently stopped smoking will also have experienced the fact of volitional transformation in a minor key.

It is crucially important *not* to over-intellectualize the will. As we have repeatedly emphasized, conscious effective first-order desires are *pre-reflective* in that they need not necessarily be accompanied by any occurrent or even dispositional *self-conscious or self-reflective* consciousness of any sort, whether this takes the form of deliberative self-consciousness or self-reflection, higher-order thoughts, or some other form of higher-order self-representation, such as the body image (see Sections 1.2 and 2.1), or even episodic memories—i.e., memories of events in one's own life, as opposed to memories of impersonal facts. Therefore the *reflexivity* of the hierarchy of desires that constitutes the will of an animal minded like us does not necessarily entail the *self-representation* of a special hierarchy of self-conscious or self-reflective representations. This is clearly shown by the fact of pre-reflective or spontaneous aimless or impulsive actions intentionally performed by rational animals like ourselves (see Section 3.4), and also by the existence of proto-rational intentional agents, such as normal human toddlers or Great apes. So the willing of an essentially embodied intentional agent is a pre-reflectively conscious effective first-order desire in a dynamic hierarchy of reflexive desires that might be, but need not necessarily also be, self-representations. Moreover, if we must avoid over-intellectualizing the will, then we must also avoid over-intellectualizing intentional action. Hence it is pre-reflectively conscious effective first-order desire in a dynamic hierarchy of reflexive desires—a.k.a. willing, a.k.a. effortless trying and its active guidance—rather than belief, judgment, thought, or self-conscious deliberative intention, that causes action.

As we said above, we want to argue that the subjective experience of willing, or effortless trying and its active guidance, is grounded in pre-reflectively conscious *desire-based emotions*[4]—"emotions$_d$" for short—or *conative affect*, and not in beliefs, judgments, thoughts, or self-conscious or self-reflective, deliberative intentions. By emotions$_d$ or conative affect we mean *caring* of all sorts, including salient drives of all sorts, inclinations of all sorts, liking and disliking of all sorts, love and hate, lust and disgust, moods of all sorts, passions of all sorts, pleasures and pains of all sorts, feelings of all sorts, sensations of all sorts, and sentience of all sorts. Emotions$_d$ differ intrinsically from the mental states, events, or processes involved in instrumental action, precisely because emotions$_d$ can produce pre-reflective or spontaneous actions—whether aimless or impulsive—which are *purposive* acts but not necessarily also *purposeful*, self-consciously or self-reflectively deliberative acts. Emotions$_d$ are in this way fully pre-reflectively conscious and sensorimotor-subjective, and can operate independently of what Searle called the "world-to-mind direction of fit"[5] that is characteristic of all self-conscious or self-reflective, deliberative intentions and motivation by instrumental reasons. On the other hand, emotions$_d$ always include the *body-to-mind* direction of fit that is characteristic of essentially embodied intentional agency,[6] such that an intention to act is always and originally a desire *to move one's own living body*. As a consequence of their pre-reflectiveness and independence from instrumental reasons, however, emotions$_d$ cannot be adequately accounted for by the classical decision-theoretic, or means-end, model of practical reasoning.

Instead, we think that the emotions$_d$ associated with pre-reflective or spontaneous actions are much more accurately understood in terms of the classical psychological concepts of *appetition, drive, impetus*, or *urge*. These classical notions all have a close affinity to the basic idea of our action theory that willing, or effortless trying and its active guidance,

[4] It is a commonplace since Freud that at least some emotions are *in some sense* non-conscious. We accept this, but also hold, by virtue of the Deep Consciousness Thesis, that no mental facts are *absolutely* non-conscious, and that all mental states or acts are at least miminally occurrently conscious. See Section 1.2. See also Searle, *Rediscovery of the Mind*, ch. 7.

[5] See Searle, *Intentionality*, 7–13.

[6] Corresponding to the world-to-mind direction of fit for self-conscious deliberative intentions is the mind-to-world direction of fit for the conceptual and propositional contents of beliefs and judgments. On our view, corresponding to the body-to-mind direction of fit for essentially embodied intentional agency, is the mind-to-body direction of fit for the essentially non-conceptual contents of perceptions. See Hanna, "Kantian Non-Conceptualism."

is a synchronous *unmoved motor* of intentional body movements, a synchronous *creative ground, origin, or source* of intentional action—the time when it is ultimately *up to me*. But at the same time, to borrow Yeats's language, it is impossible to tell the dancer from the dance; and to borrow Wittgenstein's language, it is impossible to subtract the fact that *I try to raise my arm* from the fact that *my arm goes up*. So the desire-based emotions that constitute pre-reflective willing are both causal-dynamic unmoved motors *and* essentially embodied, just like all consciousness$_{lo}$ and intentionality$_{lo}$.

Intentional action without reasons is neither *ir*rational nor *a*rational, however, precisely because the *authenticity*—the overall psychological coherence and integrity—of our rational agency depends on it. We have to be *able* to (which is not to say that we often or even usually actually *do*) resist the almost irresistible attraction of all egoistic, self-interested, hedonic, and consequence-based desires, as well as all instrumental reasons, and spontaneously *plump* for self-transcendence, thereby satisfying the heart's *deepest* desire, whether for better or for worse. This is a direct consequence of the Desire-Overriding Internalism about reasons that we defended in Section 3.4. Our capacity for pre-reflective or spontaneous *impulsive* intentional actions, or actions without instrumental reasons, therefore has a uniquely motivating and psychologically ineliminable role in the constitution of our rational intentional agency. But at the same time, the uniquely motivating and psychologically ineliminable role of pre-reflective or spontaneous impulsive actions is not always directly or easily accessible to our self-conscious or self-reflective rationality. To use Pascalian language, the head cannot always see the heart's own reasons. Yet at the same time, the head also cannot self-consciously or self-reflectively cause any action *without* the heart's own reasons. Or otherwise put, rational minded animals, including of course all rational *human* minded animals, are by their very nature sentient, sapient, sane, and *impulsive*.

So describing emotions$_d$ in terms of appetition, drive, impetus, or urge should not lead us to think that the intentional agent is somehow merely passive, compulsive, or obsessive with respect to her motivational desire-based emotions and the intentional body movements resulting from them. On the contrary, rational minded animals are capable not only of impulsive actions, but also of various forms of emotional self-control through our ability to configure and re-configure the complex structure of the dynamic

hierarchy of desires that constitutes our will. As a consequence of this fact, emotions$_d$ are at work not just in cases of impulsive actions but also in *all* cases of intentional action, including our most self-conscious or self-reflective, deliberative, and deliberately-planned intentional movements. In short, our claim is that it is emotion$_d$, in the form of pre-reflectively conscious effective first-order desires in a dynamic hierarchy of desires, that is identical to our effortless trying and its active guidance, and thereby is the mental cause of intentional actions of *all* kinds. More specifically, this is the case when those actions *also* include a reasons-driven, self-conscious or self-reflective, and deliberative superstructure built on the autonomous foundations of our pre-reflectively conscious desire-based emotions.

In Section 5.2, we critically engage with contemporary philosophy of the emotions and attempt to get clearer about just what emotions are. We argue that they bear a much closer structural resemblance to first-order or higher-order desires than they do to beliefs, judgments, or thoughts. Indeed, there seems to be no serious distinction worth making between

(i) a particular emotion,

and

(ii) a particular hierarchically organized set of pre-reflectively or meta-conscious desires, normally together with a further pre-reflectively conscious effective first-order desire to impulsively move one's body in such a way as to express that very set of desires.

The description provided by (ii) is the *definition* of an emotion$_d$. The three distinct and distinctive components of this definition are the necessary connections we are asserting between emotions and

(i) the hierarchical desire conception of the will,
(ii) the notion of a pre-reflective or spontaneous impulsive intentional act,

and

(iii) in normal cases, a supplementary pre-reflectively conscious effective first-order desire to make intentional body movements that express the agent's current hierarchy of desires.

The third factor, in particular, re-introduces a certain minimal or weak dimension of Behaviorism into the analysis of the nature of an emotion. Correspondingly, we reject accounts that over-intellectualize the emotions and understand them either as belief-desire pairs (i.e., Davidsonian reasons) or as evaluative judgments, and claim instead that emotions are essentially emotions$_d$. This brings us to our Emotive Causation Thesis, to the effect that, first, it is pre-reflectively conscious effective first-order desire that is identical to our effortless trying and its active guidance, and thereby the primary mental cause of intentional actions of all kinds, and second, that the causal powers of self-conscious or self-reflective, deliberative intentions are founded on and derived from the causal powers of this pre-reflective willing.

In Section 5.3, we analyze the intentionality or "aboutness" of the emotions, and introduce the notion of *affective framing*, which is how the feelings that partially constitute our emotions$_d$ sufficiently determine the finegrained and hyper-finegrained intentional *targets* of cognitive attention, especially including the *goals* of intentional action.

And finally, in Section 5.4, we discuss the role of emotional self-control in the desire-based emotive causation of intentional action. For us, the pre-reflectively conscious and meta-conscious mental process by which we configure and re-configure our first-order and higher-order desires in order to constitute our wills is *one and the same* as the process of emotional self-control. So, far from being normally the passive *victims* of our emotions, as classical theories of the emotions often assert or assume, according to our view we are instead normally the essentially embodied active *shapers* of our emotions by causing intentional body movements. Then, switching from an inference-to-the-best-explanation strategy to an a priori conceptual argument strategy, and using an example based on the classic 1956 sci-fi film, Don Siegel's *Invasion of the Body Snatchers*, we conclude the chapter by arguing that a conscious creature *without* emotions$_d$, or what we will dub an *Emotional Zero*, would necessarily be incapable of *our* sort of conscious cognition and intentional action, which necessarily requires capacities for what we call *attentive focusing* and *goal focusing*. An Emotional Zero therefore could not possibly be an animal that is minded *like us*.

5.2 What is an Emotion?

So what is an emotion? Not surprisingly, there have been many attempts by philosophers of mind and psychologists to provide a *reductive* analysis of the emotions. These include:

(a) the passive affect theory, which asserts that emotions are nothing but a species of receptive feeling;
(b) the drive-based or motivation theory, which claims that emotions are nothing but certain innate urges or species of motive;
(c) the behaviorist theory, which holds that emotions are nothing but certain types of overt bodily movement or dispositions to overt bodily movement;
(d) William James's theory to the effect that emotions are nothing but an awareness of certain changes in our physiology;

and

(e) the (recently popular) cognitivist theory which says that emotions are nothing but certain kinds of belief-desire pairs or evaluative judgments.

We think that each of these accounts is prone to the same fairly obvious sort of objection, namely that it leaves out some or another component that appears to be intrinsic to our emotional experience. Indeed, the inability of reductive theories to capture the nature of our emotional experience solely in terms of one or another of the isolated components of passive affect, drive or motive, behavior, neurophysiology, or cognition strongly suggests, as Peter Goldie has pointed out, that emotion is essentially a complex state, event, or process involving *all* of these elements.[7] Furthermore, says Goldie, the various elements of emotion are normally held together, in part, by belonging to a first-personal history or narrative, so that emotional experience cannot be understood apart from the rest of a person's character and life.[8]

[7] Goldie, *The Emotions*, 11. [8] Ibid., 16.

Instead of reducing emotion to some single factor, we follow Goldie's lead and take a thoroughly non-reductive and multi-factored approach. But to make this a substantive and explanatory claim, we must not merely endorse holism and pluralism and then wave our hands. On our view, there is indeed an *essential* factor in all emotion, and it is *pre-reflectively conscious desire*, or more precisely the pre-reflective consciously *felt need* for something. We then explain emotions in terms of hierarchies of conscious intentional desires, including both pre-reflectively conscious effective first-order desires and also second-order volitions. More generally, and to borrow a usefully comprehensive psychological notion exploited by both Heidegger and Frankfurt,[9] we will say that the emotions are essentially how minded animals and especially human persons *care*. Caring is how we feel about objects of all sorts, how we feel about each other, how we feel about our own feelings, and how we feel about our own lives, in part or as a whole. Then willing, or effortless trying and its active guidance, is just that special type of essentially embodied animal caring that causes intentional actions. But caring and the desire-based emotions could not exist at all, or at least could not exist at all in minded animals, without an intrinsic connection to intentional agency and the ability to perform intentional body movements.

It should be noted that caring in this comprehensive sense does not necessarily imply concern or solicitude in the normal sense of those terms. Anxiety, boredom, mild depression, disdain, disgust, dislike, embarrassment, fear, hatred, loathing, and even supremely cool indifference (what the French aptly call *je-m'en-foutisme*) are all special types of caring, since they each project some definite degree and mode of first-person valuation onto the intentional object of caring. Otherwise put, the one thing that minded animals necessarily are *not* is *Emotional Zeroes*—i.e., conscious creatures without conative affect, and thus without the ability to consciously desire something or another in some way or another. It is true that human beings and other animals can sometimes approach this non-caring or emotionally zeroed-out condition in automatism, brain trauma, nervous breakdowns, catatonic states, severe depression, seizures, and so on. But even here frequently it is a surfeit of caring—caring too much and

[9] See Frankfurt, "The Importance of What We Care About."

too intensely, without self-control—that directly leads to this relatively desireless condition. In any case, facts about these unfortunate conditions constitute the *special* domain of abnormal psychology, neuropathology, or psychiatry—or otherwise put, these facts (as striking as they are) indicate only the outermost limits or margins of our caring, and not its essence.

It is also true that a conscious, intentional animal in one of these unfortunate conditions might be unable to tell whether she is currently in a real desire-based emotional state or in a relatively desireless pathological state—as in the condition of *anosognosia*, or the inability to recognize one's own psychological illness. Similarly, as we noted in Section 4.2, act-deception and perceptual hallucination are both possible. But just as, on our view, act-deception and perceptual hallucination are categorically different from real intentional acts and correct, veridical perceptions respectively, despite their epistemic non-discriminability, so too for us anosognosic relatively desireless states are categorically different from real desire-based emotional states. So just as we are direct realist disjunctivists in the theory of action and the philosophy of perception, we also defend an *emotive direct realist disjunctivism*, to the effect that real desire-based emotional states and anosognosic relatively desireless states are categorically different and share only what is needed to guarantee the bare possibility of their epistemic indiscriminability.

What then is distinctive about the emotions$_d$, or caring, of animals minded like us? As we have already mentioned, our desire-based theory of the emotions does bear an affinity to the old-fashioned drive-based or motivation theories which claim that emotions are nothing but certain innate urges or species of motive. But while we do hold that emotions are *essentially* desires, our view is also explicitly non-reductive and multifactored, so we do not hold that emotions are *nothing but* desires. Our theory not only explicitly postulates the intrinsic presence of irreducible consciousness$_{lo}$ and intentionality$_{lo}$ in all desires, but it also explicitly postulates irreducible hierarchies of pre-reflectively and meta-representationally conscious desires (first-order, second-order, and so on), and explicitly allows for these different types of reflexive desires to be *sometimes* recorded in self-conscious or self-reflective judgments, under appropriate conditions of volitional rationality. Our theory thereby fully accommodates the irreducible

presence of judgment-based factors (belief, thought, evaluation, etc.) in the specifically self-conscious or self-reflective, deliberative superstructure of the desire hierarchy. So in other words, our theory fully accounts for the way in which we endorse some of our desires and reject others, up to and including our being able to be motivated by desire-overriding non-instrumental reasons (see Section 3.4). Moreover, the factors of neurophysiology and behavior are non-reductively incorporated into our theory of action via the necessary relation between effortless trying and its active guidance, neurobiological processes, and overt body movements. Indeed, one way of formulating the upshot of our theory of the emotions is that *desire-based emotions* and *intentional agency* are necessarily mutually connected.

Otherwise put, although not all desire-based emotions are connected to self-conscious or self-reflective, deliberative action, nevertheless all caring and all emotion$_d$ are inherently poised for essentially embodied effortless trying and its active guidance in the production of intentional body movements.[10] *But this is not Behaviorism about the emotions*, or at least it is not *classical* or *full-strength* metaphysical and methodological Behaviorism about the emotions. Classical Behaviorism about the emotions is wrong that specific emotions are necessarily correlated with some specific type of overt body movements (or with dispositions to make such movements), much less identical with them. Indeed, we accept that the old joke about Behaviorism—Two behaviorists meet on the street, and one says to the other: "You're fine. How am I?"—is amusingly and correctly pointing up an absurd consequence of classical Behaviorism, to the effect that my emotions are strictly determined by my (dispositions to) behavior and therefore are necessarily directly accessible to external observation, but necessarily also *not* directly accessible to conscious introspection, because both consciousness$_{lo}$ and introspection alike are nothing but epistemological and metaphysical illusions.[11] That really *is* absurd. Consciousness$_{lo}$ and introspection are certainly not what Descartes (when wearing his substance dualist hat) thought they were—but on the other hand they are not *nothing*. So we think that classical Behaviorism is deeply wrong that either consciousness$_{lo}$ or introspection is eliminable.

[10] See also Freeman, "Emotion is Essential to All Intentional Behaviors."
[11] See, e.g., Lyons, *The Disappearance of Introspection*.

Nevertheless, we also think that classical Behaviorism *did* contain something of fundamental importance for the philosophy of mind. This idea was fully recognized by Wittgenstein:

"But doesn't what you say come to this: that there is no pain, for example, without *pain-behaviour*?"—It comes to this: only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious.[12]

Only of what behaves like a human being can one say that it *has* pains. For one has to say it of a body, or if you like of a soul which some body *has*. And how can a body *have* a soul?[13]

"But you will surely admit that there is a difference between pain-behaviour accompanied by pain and pain-behaviour without any pain"—Admit it? What greater difference could there be?—"And yet you again and again reach the conclusion that sensation itself is a *nothing*."—Not at all.

It is not a *something*, but not a *nothing* either.[14]

If one sees the behaviour of a living thing, one sees its soul.[15]

The human body is the best picture of the human soul.[16]

The behaviorist recognizes that conscious, intentional minds_{lo} are necessarily and completely embodied in motile, spatially situated, forward-flowing living organisms. And we pre-reflectively and directly respond to the presence of another conscious, intentional mind_{lo} by *empathically mirroring*[17] its intentional body movements in our own overt and covert body movements. Hence the kernel of truth in Behaviorism is captured by the Essential Embodiment Thesis. So our unsolicited answer to Wittgenstein's rhetorical question, "And how can a body *have* a soul?," is this:

A body can *have* a soul—by which we mean a conscious, intentional mind_{lo}—*but only by essentially embodying that soul*. The Latin word for 'mind' or 'soul' is '*anima*', and this aptly captures the sense in which a mind or soul like ours is just that which *animates* a suitably neurobiologically complex living body. Even more precisely, a mind or soul like ours is just the *animating truly global or inherently dominating dynamic intrinsic structure* of a suitably neurobiologically complex living

[12] Wittgenstein, *Philosophical Investigations*, §281, 97. [13] Ibid., §283, 98.
[14] Ibid., §304, 102. [15] Ibid., §357, 113. [16] Ibid., 178.
[17] See ch. 1, n. 3 above; and also Gallese, Keysers, and Rizzolatti, "A Unifying View of the Basis of Social Cognition."

body. A conscious, intentional $mind_{lo}$ is irreducible to its own living body, and therefore it is not a "nothing." But as an animating structure, it is also not a Cartesian "something" either. Thus the intentional body movements of animals minded like us, and especially their pre-reflectively or spontaneous impulsive intentional body movements, as empathically mirrored by us, provide the best *picture* of their essentially embodied souls.

In this way, for animals minded like us, all emotion is essentially embodied, and furthermore there is no emotion that cannot be impulsively expressed in overt intentional body movements—even if this is only grimacing for a moment or making an inarticulate noise under one's breath. To be sure, the bodily expressions of emotions often can be stifled or suppressed to some extent; and there is also the special case of bodily paralysis, which we will consider in Section 5.4. Still, most emotions are plainly visible on the features or in the gestures of the people who are experiencing them, or plainly audible in the sound of their voices as they talk. Some emotions—e.g., terror—even have a bodily stereotype, as Munch's iconic painting *The Scream* clearly shows.

Recently popular *cognitive theories* of emotion that center on belief-desire pairs or evaluative judgment tend to shunt pre-reflective desires, feelings, neurobiology, and intentional body movements to the sidelines. But we think that, on the contrary, they are all intrinsic parts of the complex essence of emotion, understood as $emotion_d$. Since our desire-based approach to the emotions is the more unfamiliar one, obviously the burden of proof is on us. In order to motivate our own approach, we will critically discuss the belief-desire account of emotion, and also Robert Solomon's theory of "emotion as judgment," and maintain that such accounts do not adequately account for the desiderative, affective or felt, and embodied dimensions of emotion. And while, as we have said, we fully agree that emotions cannot be explanatorily or ontologically reduced to desires, feelings, neurophysiology, or (dispositions to) behavior, we do also hold

(i) that all conscious desires in minded animals are consciously felt needs,

and

(ii) that all conscious desires in minded animals are necessarily and completely neurobiologically embodied,

and will also argue

(iii) that an emotion is essentially a set of pre-reflectively or meta-conscious desires in a dynamic hierarchy, normally together with a further pre-reflectively conscious effective first-order desire to impulsively move one's body in such a way as to express that very set of desires.

Cognitive Theories 1: Emotions as Belief-Desire Pairs

Hume famously asserted that reason is fully subservient to the passions, that practical rationality is instrumental, and that reason's task is only to represent or infer the means whereby one can attain the ends established by desires. It is commonplace to criticize the Humean view that reason is fully subservient to the passions by pointing out reason's autonomously self-conscious or self-reflective and deliberative function of evaluating reasons by weighing the considerations offered up by different desires, and then deciding for or against the possible ends, goals, or courses of action picked out by desires. It is less common, however, to criticize Humean accounts of *desire* per se. Humean desires are generally taken to be psychologically basic facts having what Searle calls a "world-to-mind direction of fit," and thereby having propositionally-stuctured satisfaction-conditions which project ways of changing the world to suit our self-conscious deliberative ends. Indeed, the notion that desires are ultimately ways of propositionally representing instrumentally-attainable purposes and goals is the conventional wisdom in the philosophy of action. But can *emotions* be accommodated within this Humean model? Or otherwise put, can emotions be explained in terms of belief-desire complexes and instrumental reasons? If not, then *either* the Humean model of desire must go, *or else* emotions and desires are not essentially connected. We will pursue the former option.

Classical causal theories of action of the Davidsonian variety hold that providing a causal explanation of an action is a matter of positing appropriate motivational factors (desires, intentions) and related epistemic factors (beliefs, judgments) in the mental states of an agent. According to Davidson, as we have seen in Chapter 3, citing the primary reason or the relevant belief-desire pair yields a mental cause and causally explains an agent's behavior. At first glance, it may be tempting to analyze emotion in this way too, as a pro-attitude in favor of some action (the motivational

component) together with an evaluative appraisal (the cognitive component). However, there are many examples of cases for which such an analysis proves inadequate.

One salient range of examples is provided by cases in which subjects have inappropriate or "inert" emotions that do not change in light of changed beliefs. For example, suppose that Mary regards her co-worker Mike with utter disdain and continues to find him at best "a complete loser," even though he never treats her with anything but polite affection, collegiality, good humor, and kindness. And further suppose that Mike is instantly deeply attracted to Mary, falls hopelessly in love with her, and continues to find her utterly irresistible even though she never treats him with anything but manipulative callousness or supreme indifference—or even worse, *both*, in alternating sequence, depending on her mood. This sort of situation, as unhappy as it is, is not at all unusual. In such cases it would be manifestly implausible and inaccurate to explain Mike's and Mary's inappropriate emotions by attributing belief-desire pairs that would rationalize them.[18] While false beliefs and irrational intentional sets are of course possible, they by no means account for every instance of conservation of the emotions.

Another salient range of examples arises from cases of akrasia or so-called weakness of the will—or more accurately, what we have called *impulsiveness of the will*—in which we act in a sudden, uncalculating, and unplanned way, without or against the rulings of our instrumental judgments. We have seen that the traditional belief-desire model cannot account for cases of akrasia, nor can it provide an adequate characterization of action that is inherently driven by impulsive desire-based emotion. Indeed, pre-reflective or spontaneous impulsive action is a paradigm case for the present discussion.

One problem with postulating a belief-desire explanation of impulsive action is that it significantly over-intellectualizes the emotions.[19] The belief-desire account would hold, e.g., that the intentional elements involved in fear are desires and beliefs that are related syllogistically. When asked, "Why did Deirdre suddenly run away from Dan?," we might answer that Deirdre wanted to get away from Dan and believed that this was the best way of achieving this in the given circumstances. However, it seems that this sort of explanation is consistent with Deirdre's subjectively experiencing no emotion at all. Yet, other things being equal, surely the

[18] Rorty, "Explaining Emotions," 104.

[19] See also Goldie, *The Emotions*, 38.

best explanation is that Deirdre is terrified by Dan. There seems to be a crucial difference between an action that is done impulsively on the basis of pre-reflectively conscious desire-based emotion, and an action that results from the more humdrum self-conscious or self-reflective, calculative, or deliberative emotions involved in means-end reasoning. Thus the crucial role of pre-reflectively conscious desire-based emotion in action is forced out of the picture from the start by the supposed explanatory sufficiency of belief and desire to explain actions of all kinds.

Consider now the often-cited example of Jane's scratching out the eyes in a photograph of a person she hates with an intensely jealous passion—call her Joan. The belief-desire model attempts to rationalize such an action by attributing means-ends reasoning to the agent. However, it seems quite clear that means-ends reasoning is altogether absent from this sort of "heart-on-one's-sleeve" type of impulsive action.[20] Some have claimed that the belief-desire model can be retrofitted to account for such cases. One might claim, for example, that what needs to be added to the belief-desire explanation of Jane's behavior are more beliefs and desires: Jane's desire to scratch out Joan's actual eyes, her belief that scratching out the eyes in the photo will allow her to vent her intensely jealous hatred of Joan, and her belief that the photo represents Joan. But surely this sort of explanation grossly over-intellectualizes Jane's action. Her intentional body movement is far more akin to an *improvisational dance* than it is like an *instrumental act*. We believe that the most plausible explanation of Jane's act is that she is just *her own intense jealousy incarnate* by scratching out the eyes in Joan's photograph. In other words, Jane's intentional body movements constitute a sort of *self-depicting diorama*, or to use Wittgenstein's term, a *self-depicting picture*—"the human body is the best picture of the human soul"—of her jealousy of Joan. Or in still other words, Joan's intentional body movements, as spontaneously or pre-reflectively and impulsively *self-expressive* and *self-referring*, are a necessary part of that very emotion.

According to Goldie, instead of giving up on the belief-desire model so easily, we should look around for a better retrofitting of the model that will account for such cases. In that spirit he then offers the thesis that some of our desires, which we do not believe it possible to satisfy, are idle wishes,[21] and that some of our actions serve as symbolic expressions of these wishes.

[20] Doring, "Explaining Action By Emotion," 215. [21] Goldie, *The Emotions*, 129.

For example, we might then say that Jane has a passionate desire to scratch out Joan's eyes, but believes it impossible to satisfy this desire in the actual world, and then imagines she is doing this through her action. Goldie claims that there is a symbolic match between the object of Jane's emotion (i.e., Joan) and the object towards which her scratching activity is directed (i.e., the photograph of Joan). While this sort of symbolic action certainly is possible, does it account for all or even most cases of heart-on-one's-sleeve emotional action? Goldie himself notes that we sometimes "take out" our emotions on the nearest objects at hand, which may have no symbolic relation whatsoever to the object of our emotion. For example, in a fit of frustration about one's finances, one might slam the door or kick over a chair. Goldie proposes that such behavior might be rendered intelligible by the desire to vent one's emotions.[22] His idea seems to be that we have a primitive standing desire to vent our emotions, and then recognize that slamming the door or kicking over the chair is the way to accomplish this on some particular occasion.

But it seems to us that in many or even most cases it will be far more plausible to describe heart-on-one's-sleeve cases as actions done pre-reflectively, impulsively, self-expressively, and self-referringly rather than as an attempt to satisfy one's desire to vent.[23] In the examples above, surely it is just intense frustration, consciously felt at least in part as a pre-reflectively conscious urge to move one's living body in a way that constitutes a self-depicting diorama or picture of one's own intense frustration, which mentally causes the door-slamming or table-kicking movements.

It seems clear, then, that our pre-reflectively conscious desire-based emotions are often very different from ordinary self-conscious or self-reflective desires that seek to bring about some concrete change in the world. It also seems clear that in such cases the intentional body movements caused by these pre-reflectively conscious desire-based emotions will lack any clear further goal or purpose, and therefore will not support any further non-basic act that is supposed to be brought about by means of those movements. For example, we can see the obvious contrast between

(1) suddenly raising one's arm while freestyle hip-hop dancing,

[22] Goldie, *The Emotions*, 134.

[23] It is certainly true that on some occasions, we may have such beliefs and desires. However, it is implausible to think that this is always the case.

and

(2) waving to a loved one by suddenly raising one's arm,

even if the body movements would be indiscriminable to a decontextualized outside observer.[24] This can be shown in at least three ways.

First, because one can have an impulsive self-expressive desire to move one's body in a certain way and yet also *not* want the world to be changed to fit a relevant corresponding self-conscious or self-reflective desire to move one's body in that way, we should be cautious about applying the idea of direction of fit to such actions. In the case of Jane, no doubt the intensely jealous desire to scratch out Joan's eyes is a conscious first-order desire that Jane actually *feels*, but in fact she does *not* self-consciously or self-reflectively endorse that pre-reflectively conscious first-order desire in a corresponding second-order volition. Jane impulsively scratches the photograph, not Joan, and this could be true even if Joan were in the next room. In self-expressive impulsive action there is always a *body*-to-mind direction of fit, but not necessarily or perhaps even usually a *world*-to-mind direction of fit.

Second, there are many cases of pre-reflectively conscious effective first-order desire in which no goal-directed desire or instrumental reason whatsoever is involved. For example, there are various characteristic body movements associated with excitement that are not instrumentally purposeful at all. Agents do not usually jump for joy in order to advance some further goal. A more plausible explanation is that they just impulsively move their own living bodies, and thereby create self-depicting dioramas or pictures of their own excitement.

And third, if an agent, call her Anne, feels very proud about how things have turned out (say, she has won an award) and impulsively smiles, her pre-reflectively conscious desire-based emotions are thereby directed toward the way things *just are* and not toward some different non-actual way she wants them to be.[25] The proud agent can pre-reflectively desire things to be *just as they are*, hence endorsing the actual world in that context, and so not desire to change the world to fit her goals.

Two further important points should be noted here.

First, desire is an integral part of the emotions associated with both excitement and pride. Anne's deep sense of accomplishment, for example,

[24] Doring, "Explaining Action By Emotion," 219.

[25] Goldie, *The Emotions*, 78.

is of course partially based on knowing that she has won an award, but depends primarily on her desiring things to be just as they are. If she did not have this desire, then she would not feel proud in that way. Although emotions need not involve any forward-looking, goal-directed, or instrumental desires, it does not in any way follow from this that emotions can ever lack desires altogether.

Second, and in direct opposition to the thesis that there can be emotions without desires, we hold that in normal circumstances it is impossible for a minded animal to feel an emotion without *also and thereby* desiring to impulsively move its own body self-expressively in some way or another—where this can include motionless intentional orientations and positionings (see Section 4.1). If this is correct, then all emotions are inherently poised to cause basic intentional actions, and in particular to cause body movements that create self-depicting dioramas of those very emotions. To be sad is normally also to have a pre-reflectively conscious effective first-order desire to impulsively move one's body sadly; to be happy is normally also to have a pre-reflectively conscious effective first-order desire to impulsively move one's body happily; to be frustrated is normally also to have a pre-reflectively conscious effective first-order desire to impulsively move one's body frustratedly; and so-on.

We are *not*, however, saying that for each emotional type necessarily there is some *specific* way of moving one's body—as it were, the Sad Way, the Happy Way, the Frustrated Way, and so-on. But although we think that classical Behaviorism about the emotions is false, at the same time we also think that classical behaviorists saw an important truth about minded animals—our essential embodiment and embodied agency—as through a glass, darkly. So we *are* saying that for minded animals, necessarily for every emotion there is normally a further pre-reflectively conscious effective first-order desire to give impulsive bodily self-expression to that very emotion, *in some way or another.*

It is true that the pre-reflective impulse to move one's body in some emotionally self-expressive way often can be stifled or suppressed to some significant extent. One can do one's best not to cry hot tears of frustration or to raise one's voice in anger. But even if it is true that the dampening modulation of the impulsive self-expressive acts associated with emotions is always possible to some extent, then that in turn presupposes that it is indeed normally a necessary component of every such emotion to pre-reflectively

consciously *want* to move our bodies in some self-expressive way. Indeed, it seems to us that the intentional stifling or suppression of the pre-reflectively conscious effective first-order desire to move one's body in an emotionally self-expressive way is usually only partially successful. For example, putting on a poker face when extremely angry or desperately disappointed is usually still, in some subtle way, a self-depicting bodily picture of one's anger or disappointment. To recognize this, one need only look closely at people in social situations, and then compare and contrast the poker face of actual poker games with the poker faces of extreme anger or desperate disappointment, or with the sort of poker face that is put on when one is desperately bored at a department meeting, or again with the sort of poker face that is put on when one is trying desperately not to laugh at a funeral.

But what about the possibility of a completely successful stifling or suppression? Is that a problem for our theory? In his famous critique of Behaviorism, Hilary Putnam used the thought-experiment of a race of *Super-Spartans* who, by dint of generations of training plus some adaptive evolution, have learned to eradicate the bodily expression of being in pain.[26] Super-Spartanhood does seem to be at least logically possible. But Super-Spartanhood is still no counterexample to our view since it is only the *pre-reflectively conscious effective first-order desire* to impulsively move one's body in an emotionally self-expressive way that is required by us, which of course means that it is a desire that does, or will, or *would* move us all the way to action—if not actually impeded by something else. And by the hypothesis of the Super-Spartan example, this *would-be* effective first-order desire to give bodily expression to feelings of pain still exists even for the Super-Spartans, since it is that very desire that has to be actually *stifled or suppressed* by them.

And this vividly brings out yet another sharp contrast between the Essential Embodiment Theory and classical Behaviorism. While classical Behaviorism holds that all mental states, to the extent that they exist at all, happen only at the *surfaces* of animal bodies, the Essential Embodiment Theory holds that all mental states are necessarily and completely *neurobiologically* embodied, and thus that the mental life of a conscious, intentional creature necessarily happens *in-and-through* the brain and necessarily

[26] See Putnam, "Brains and Behavior."

in-and-through all the other vital systems as well, whether or not any overt body movements can or do occur.

In any case, the bottom line here is that agents do not need to be acting in pursuit of some goal whenever they intentionally move their bodies. While our actions do indeed sometimes reflect further symbolic desires or instrumental goals, in many other cases we do things in an impulsive way and on the basis of pre-reflectively conscious desire-based emotions alone. Such impulsive actions moreover, while often trivial—say, suddenly frowning when the sun goes behind a cloud—might on the other hand completely revolutionize the agent, turn her motivational world inside out, and change her life. For example, Dashiell Hammett's *The Maltese Falcon* contains the tersely beautiful inserted story of Flitcraft, a man whose life was radically changed by a close encounter with a falling beam:

> The life [Flitcraft] knew was a clean orderly sane responsible affair. Now a falling beam had shown him that life was fundamentally none of those things. He, the good citizen-husband-father, could be wiped out between office and restaurant by the accident of a falling beam. He knew then that men died as haphazard like that, and lived only while blind chance spared them. . . . By the time he had eaten his luncheon he had found his means of adjustment. Life could be ended for him at random by a falling beam: he would change his life at random by simply going away.[27]

In a brilliant literary expression of Existentialism in the guise of hard-boiled pulp fiction, Hammett is telling us, it seems, that Flitcraft impulsively incarnates the emotional recognition of the real possibility of his own death[28] by "flitting" off in order to "craft" a different life for himself. When we act in an impulsive way and as a result of pre-reflectively conscious desire-based emotions, neither reasons in general nor rationalizing means-end descriptions in particular are able to account for the nature of our intentional performances.

Cognitive Theories 2: Emotions as Judgments

Given the fundamental importance of pre-reflectively conscious desires, both in the psychological constitution of the emotions themselves, and also in the etiology of basic intentional acts, it may then seem unclear

[27] Hammett, *The Maltese Falcon*, 64.
[28] See Heidegger, *Being and Time*, §50, 250/294, and §53, 263–307.

how we should account for the specifically cognitive element involved in the emotions. In "Emotions and Choice" and *The Passions*, Solomon famously argued that both the intentionality of the emotions and their active status can be captured by characterizing emotions as evaluative judgments. According to the early Solomon, an emotion is an evaluative judgment about one's situation, a "personal evaluation of the significance of [a particular] incident" that projects one's values and ideals.[29] Note that Solomon does not regard emotions as judgments *simpliciter*, but rather as *constitutive* judgments that supply standards of interpretation and evaluation to our experience and constitute the framework within which those experiences and facts have meaning. For example, our choosing to get angry is what makes a comment offensive. Likewise we constitute, not find, the charms and virtues of the person we choose to love. Because agents shape and structure their world according to these constitutive judgments, Solomon finds it plausible to consider them judicative actions. Emotions are not occurrences that merely happen to us, but instead are rational and purposive evaluative judgments.

One reason that Solomon gives for identifying emotions with evaluative judgments is their common logic or formal structure. Because emotions have a characteristic formal structure, this structure can be explicitly described and regimented, like any other logical or conceptual system.[30] Solomon defines each emotion according to its characteristic sort of judgment and claims that the logic of an emotion dictates the logic of the resulting emotional expression: joy demands a joyful expression, love demands a loving expression, and so on. For example, because anger is essentially a judgment of condemnation, there can be no anger without the desire to punish. Moreover, the relationships between beliefs, opinions, and emotions are in some sense a matter of logic, so that a change in beliefs typically entails a change in emotions. Tony cannot be angry at Tom for something Tony believes Tom did not do. According to Solomon's early work, then, Tony's anger should vanish immediately upon the refutation of the putative fact he was angry about.[31]

[29] Solomon, *The Passions,* 126. [30] Ibid., 195.

[31] Everyday experience suggests that the beliefs and emotions of agents do sometimes conflict. Sue may recognize that her husband Steve did not really say horrid things to her and that this occurred only in her dream the night before. However, she may nevertheless still feel angry at Steve for

But as other philosophers of emotion have noted, it seems clear that the formal structure of the emotions differs from that of judgments. Patricia Greenspan, for example, explores the everyday phenomenon of *mixed feelings*[32] and points out that a rational person can easily have contrary emotions about the very same object. For example, if Judy is happy that John won the award (because she likes and respects him), yet she also is unhappy that John won the award (because she did not win it), we can then regard Judy's happiness and unhappiness as contrary attitudes towards the same object. However, if we follow Solomon's recommendation, we are led to identify Judy's emotions with the following evaluations:

A. John's winning the award is good.
B. John's winning the award is bad.

Assuming that Judy is a normal rational, sane person, it is unlikely that she will hold both of these contrary beliefs at the same time. If asked about her evaluative view of the situation, Judy will likely qualify her judgments with distinct uses of "insofar as," so that they are no longer genuine contraries:

A.* John's winning the award is good insofar as I like him and he deserves it.
B.* John's winning the award is bad insofar as I really wanted to win it myself.

One can have contrary emotions, on the other hand, without qualification, and also without "blending" them together into a single intermediate emotion.[33] Moreover, an emotion can persist even when it is accompanied by much stronger opposing feelings—e.g., Mike's undying love for Mary, even when he is astounded and made deeply unhappy by her alternating callous manipulation of him and supreme indifference towards him—so that the notion of contrariety takes on a different meaning in the case of emotions. The fact that contrary emotions *can* both be "true"[34] (insofar as they are *appropriate*) while contrary judgments *cannot* shows, as against

saying horrid things to her. Emotions do sometimes persist in the face of evidence that suggests they should disappear.

[32] Greenspan, "A Case of Mixed Feelings: Ambivalence and the Logic of Emotion," 223.

[33] Ibid., 232.

[34] The very fact that it is odd to speak of "truth-makers" with respect to the emotions supports the claim that the "logic" of emotion differs from that of judgment.

Solomon, that the logic of judgments is actually quite different from the logic of emotions.

Arguing along similar lines, and in a direction that points towards our theory of the emotions, Jenefer Robinson claims that the logic of emotion in fact conforms much more closely to the logic of *desires* than it does to the logic of judgments. To show this, she spells out four basic ways in which the logic of emotion seems isomorphic to the logic of desires:

(1) Emotions and desires both allow degrees of intensity.
(2) Resistance to summing: neither conflicting desires nor conflicting emotions can be "summed up" into one intermediate emotion or desire.
(3) Tolerance of inconsistency: inconsistent desires and emotions can exist in a basically rational person. Even when one succeeds in ranking one's desires, rationality does not require that one drop the second-ranked desire altogether.
(4) Resistance to change: Two inconsistent desires or emotions can persist unchanged in a basically rational person. There may be adequate reasons for both of the two conflicting desires or emotions.[35]

Taken together, we think, these considerations collectively show that emotion is much more likely to be intrinsically connected to desire than it is to belief or judgment. As William Lyons points out, this is not at all surprising in view of the obvious fact that emotional reactions are typically deeply related to our basic wants, and also deeply informed by a broader set of wants, goals, and values.[36] So, because an agent's desires serve as the primary determining factor of the content of emotions, it seems merely Pickwickian to call them "judgments".[37] Furthermore, if emotions really are determined by our wants, then giving an adequate analysis of the nature of emotion will necessarily involve an appeal to desires in order to make sense of the core fact that an emotional subject always wants things to be a certain way, or not to be a certain way.

Against this critical backdrop, we can see that several basic problems arise for Solomon's account in particular and for cognitive theories of the

[35] Robinson, "Emotion, Judgment, and Desire," 735–6.
[36] Lyons, *Emotion*, 186.
[37] Robinson, "Emotion, Judgment, and Desire," 737.

emotions more generally. First, it is unlikely that emotional judgments are constitutive in the way Solomon suggests.[38] It does not seem correct that Barbara renders Ben's comments offensive by choosing to get angry during Ben's utterance-act, but rather that she becomes angry just because Ben's comments are offensive. Indeed, she may even be responding more directly to Ben's offensive body-language and derisive tone of voice, and not to what he says. Everyone with eyes to see and ears to hear knows how "Oh, what a nice dress (haircut, hairdo, pair of shoes, tie, etc.)!," when said in a certain way and with a certain look on one's face, can be a horrid insult. On this point, Solomon seems to get things just backwards. To be sure, we are sympathetic with Solomon's claim that emotions play a basic role in shaping our cognitive and practical interpretation of the world by rendering certain facts or possibilities salient and thereby focusing our attention and goals. But we also hold that desire-based emotions can very frequently respond immediately to the way the world and other people are presented in sense perception, and have nothing to do with full-fledged evaluative judgments in those contexts, even if desire-based emotions in other contexts do include such judgments.

Second, and even more importantly for the present discussion, it seems highly doubtful that emotions ever *are* judgments, although of course judgments can be involved in various emotions. For one thing, there seem to be many cases of emotions without corresponding judgments. An individual may just feel sad, for example, without her sadness involving any sort of judgment, and can remain sad even if she has made a self-conscious judgment that there is nothing to be sad about. This reveals the fact that our emotions intrinsically have a certain *cognitive impenetrability*,[39] which is to say that a change in judgment need not result in any change in the corresponding emotion. An emotion can persist even once an individual has recognized that some of her past judgments are false. For example, Randy can believe that spiders will not harm him and yet still be intensely afraid of spiders, and this is because emotional experience often is impervious to an agent's holding certain relevant beliefs. Similarly, we can make self-conscious rational judgments calling out for a certain emotion, and yet not feel the expected emotion. Renée might, for example, make

[38] See Roberts, "Solomon on Control of Emotions."

[39] See Greenspan, "A Case of Mixed Feelings: Ambivalence and the Logic of Emotion," 233–4; and Goldie, *The Emotions*, 76.

a self-conscious or self-reflective rational judgment to the effect that she is guilty of some crime, and yet not *feel* guilty.[40]

Our account of emotion, on the other hand, is well-suited to make sense of cognitive impenetrability. We hold that an emotion is a set of pre-reflectively conscious first-order desires, together with a meta-representational superstructure of hierarchically-ordered desires, normally also together with a further pre-reflectively conscious effective first-order desire to impulsively move one's body in such a way as to express the relevant desire-set. So sadness is a way of wanting the world, oneself, or other people to be a certain way, or not to be a certain way, plus the higher-order desires one has about those desires, normally also plus the pre-reflectively conscious effective first-order desire to impulsively move one's own body so as to express one's desire that the world, oneself, or other people be that certain way. For example, for Colleen to be sad about the break-up of her relationship with Chris is for Colleen to want various things about the world, herself, and Chris to be different, and also for her pre-reflectively to want to impulsively move her body sadly, whether by crying, frowning, snapping at other people, moping about listlessly, or whatever. Given this account, it is easy to see how Colleen's sadness can persist even if she judges that the relationship was not worth continuing and believes that the break-up with Chris was a good thing for both of them.

Third and finally, it seems very plausible that it is the spontaneous upsurge of pre-reflectively conscious effective first-order desires, rather than judgments, that motivate and cause the intentional actions normally associated with the various kinds of emotion. For example, a person who does not have the desire-based emotion of guilt is unlikely to be moved to apologize, express regret, or make amends. (Of course, someone might apologize without really feeling guilty, because she sees that is in her own best interests or otherwise advances her instrumental desires. However, a disingenuous apology or expression of regret is likely to be different in character from sincerely apologetic or guilty behavior and may well be detected as inauthentic by others.) A person who is pre-reflectively and phobically afraid of spiders, on the other hand, is likely to avoid them regardless of the judgments she has made about them. Judgment, belief, or

[40] Roberts, "Solomon on the Control of the Emotions," 399.

thought on its own is *not* what drives intentional action. Reasons alone are *never* mental causes of action.

As a coda to this discussion, it is worth noting that several theorists, including Solomon, have denied that *feelings* are a central or intrinsic feature of the emotions. Solomon, e.g., claims in his early work that emotions are conceptual structures rather than feelings and asserts that feeling is the ornamentation of emotion rather than any part of its essence.[41] He motivates this claim by offering three objections to the thesis that for each emotion there is a distinctive set of feelings. First, the feelings associated with one emotion often are no different from the feelings associated with another emotion. Because we cannot always tell emotions apart by how they feel, emotions cannot be individuated on the basis of feelings. Second, we often have an emotion without experiencing any particular feeling. In the most extreme indignation, for example, one finds oneself completely numb, which suggests that one can have an emotion without feeling anything. Third and finally, emotions are much more rich and complex than mere feelings. While we may be mistaken about our emotions, we cannot be mistaken about our feelings. Similarly, while emotions can be appropriate or inappropriate, and must have objects, feelings, like headaches, cannot be inappropriate and are not about anything.

We fully agree that emotions should not be identified with objectless feelings, or viewed as analogous to headaches. As we stated at the outset, emotions are inherently complex and irreducible to any one simple factor. However, it simply does not follow from the fact that emotions are non-identical with feelings, and can be associated with different feelings in different contexts, that emotions do not necessarily involve some feeling *or another*. Even feeling completely numb and unresponsive is itself a special kind of feeling.[42] Only the strange creatures we call *Emotional Zeroes* would lack all feelings whatsoever, and as we already mentioned above, we will argue later in the chapter that it is a priori impossible for an Emotional Zero to have a consciousness *like ours*. So according to our theory that emotion is essentially desire-based emotion, necessarily every emotion$_d$ involves some feeling or another.

This leads directly to an even more important point. Cognitive theories of the emotions must characterize the intentional contents, objects, and

[41] Solomon, *The Passions*, 60 and 97. [42] See Section 2.3.

reference of emotions as intrinsically conceptual or propositional, and at best extrinsically phenomenal. But if emotions are essentially conscious effective first-order desires (in a dynamic hierarchy of desires, normally together with the further pre-reflectively conscious effective first-order desire to impulsively move one's body in such a way as to express that very set of desires), then those desires intrinsically contain phenomenal character. It follows that the intentionality of the emotions is necessarily connected with the role of feeling in emotion$_d$, because the phenomenal character of consciousness$_{lo}$ just *is* the felt dimension of consciousness$_{lo}$. Now as we argued in Section 1.2, the felt dimension of consciousness$_{lo}$ is grounded in primitive bodily awareness. If that is correct, then the phenomenal character or felt dimension of emotion is precisely *what-it-is-like-to-be, for an essentially embodied mind, during the occurrence of desire-based emotions*. And that, of course, brings us right back to the Essential Embodiment Thesis. In the next section, we will argue that all desire-based emotions have intentionality$_{lo}$, that this intentionality$_{lo}$ is essentially embodied, and that it necessarily includes affect or feelings.

5.3 The Intentionality$_{lo}$ of Desire-Based Emotions

To describe *how* and *what* an emotion is about, or to describe an emotion's being directed in *some way* or another at *some target* or another, is to describe that emotion's intentionality$_{lo}$. Otherwise put, as we said in the Introduction and Sections 2.1 and 2.2, intentionality$_{lo}$ is the *directedness* and *aboutness* of minds$_{lo}$. We thus adopt the classical phenomenological view of intentionality (common to the work of Brentano, Meinong, Husserl, early Heidegger, early Sartre, and Merleau-Ponty) which says that all intentionality necessarily involves

(i) mental *episodes*, whether mental *acts* or mental *states*,
(ii) mental *targets* of directedness, whether objects, actions, locations, events, other conscious creatures, or itself,

and

(iii) shareable mental ways of *representing* or being about those targets, or *contents*.

In this section, we argue that a particular way of representing an intentional target—where the notion of a "target," again, is understood in the maximally broad sense that comprehends objects, actions, locations, events, or itself, including targets which may or may not actually exist[43]—necessarily partially constitutes every desire-based emotion, and that the desire-based emotional representation of an object, in turn, is partially constituted by the subject's essentially embodied feelings about that object. On our view, the intentionality$_{lo}$ of emotions is neither reducible to nor requires the intentionality$_{lo}$ of belief, judgment, or thought. Otherwise put, the intentional content of emotions need not be understood as the content of a belief or as the object of a propositional attitude. Indeed, we want to describe the intentional content of a desire-based emotion in terms of its egocentrically-centered and spatially oriented, thermodynamically irreversible, and essentially embodied *focus* rather than in terms of something that necessarily falls under a conceptual or propositional description.

Desire-Based Emotions and Propositional Content

Solomon was certainly not alone in assigning propositional intentional contents or objects to all emotions. For example, in his classic study *Action, Emotion, and Will,* Anthony Kenny held that all desires and emotions have content in the way that belief has content. He argued that emotions are essentially directed to objects and that in general it is not possible "to ascribe a piece of behavior to an emotional state without at the same time ascribing an object to the emotion."[44] We cannot, e.g., ascribe an agent's flight to his state of fear without ascribing an object to that state of fear. Similarly,

[43] There is of course a deep problem about how to understand singular cognition and reference directed at non-existent targets. One possible solution, which we favor, is to say that all singular cognition and reference are target-dependent and essentially indexical, although we need not identify this target-dependence and essential indexicality with the properties of an *actual referent*. Then in the case of singular cognition of and reference to non-existent targets, the target-dependency and essential indexicality attach to the actual intentional *episode*, not to an actual referent. So it will be possible to say that every singular thought has an *actual target*, even if not every singular thought has an actual referent. And since intentional episodes can be treated either as tokens (occasions of thinking) or types (ways of thinking), then these actual episodic entities can be treated as either private or shared. Thus Mr Pickwick, who obviously does not exist, can be treated as either *some particular person's* (say, Dickens's, or my, or your) *occasion of thinking about Pickwick* or as *a shared way of thinking about Pickwick*, depending on the context. Correspondingly, the propositions expressed by the sentences used in acts of singular thought that have an actual object, but not an actual referent, can be interpreted as either false or truth-valueless, depending again on the context. In any case, we will assume that singular intentionality$_{lo}$ is possible, whether or not the intentional target actually exists.

[44] Kenny, *Action, Emotion, and Will*, 60.

William Lyons claims that a person is not really in a state of fear until he both believes that he is in danger and also wants not to be.[45] And Eddy Zemach argues that anger is caused by a belief that a bad thing happened, which causes the agent to view the situation as outrageous, which in turn justifies some action, say, an attack.[46]

However, we think that the thesis that emotions necessarily have propositional objects and are essentially a matter of rationalized mental predication is mistaken. If emotions do indeed all require objects of desire, as we think, nevertheless these need not always be propositional in structure. Moreover, we also have some allies here. Annette Baier, e.g., points out that when music arouses our emotions, ''there are not a series of reportable belief states with any particular propositional content'' associated with those emotions.[47]

Fred Dretske also has made similar claims with respect to perceptual experience. In *Seeing and Knowing,* he presents an argument for what he calls ''non-epistemic seeing'' that suggests that perception does not necessarily involve either belief or propositional content. Because the statement ''S sees D'' does not logically entail ''S believes P,''[48] it is perfectly consistent to say that someone saw something without believing himself to be visually aware of anything.[49] To demonstrate this, Dretske draws attention to preoccupied states in which we see things without being aware of them, and to the conscious perceptual states of human infants or non-human animals. He points out that while you very likely saw most of the leaves on, say, the tree in front of your house this morning as you left to go to work, it is very unlikely that any belief or propositional content accompanied the seeing of any particular leaf.[50] It also seems obvious enough that a normal human toddler (say, a boy named 'Clyde') or a non-human animal (say, a cat named 'Otis') living across the street could pre-reflectively consciously see the same leaves on that same tree without having any beliefs or propositional contents at all.

In much the same way, it is highly doubtful that belief or propositional content is required in all cases of desire-based emotion. There is nothing logically inconsistent in supposing that someone experiences an emotion without any sort of belief or propositional content accompanying that

[45] Lyons, *Emotion*, 94. [46] Zemach, ''What is Emotion?'', 202.
[47] Baier, ''What Emotions are About,'' 12. [48] Dretske, *Seeing and Knowing*, 6.
[49] Ibid., 10. [50] Ibid., 11.

emotion. And there do also seem to be many real-world cases of this. For example, Sarah can directly experience her passionate love for Sam simply as a sort of ongoing pre-reflective emotional buzz or high, without even having to think about him. More mundanely, anyone can be in a bad mood or a good mood without any accompanying beliefs or propositional thoughts at all. Another good example is people at dance clubs. Lost in the throbbing music, in a semi-darkened room, and rhythmically moving along with many other dancers, individuals on the dance floor often find themselves pre-reflectively and unself-consciously experiencing a wide variety of vivid emotions: amusement, excitement, free-floating sexual desire, nervousness, nostalgia, or sheer joy. And depending on their desire-based emotional state, they will make impulsive efforts to move their bodies in a number of different ways, but this is rarely accompanied by beliefs or thoughts about their body movements. So in impulsive and, as it were, *dionysian* dancing we have a clear case of desire-based emotion, willing, and intentional body movement without accompanying beliefs or propositional content, which illustrates the fact that emotional intentionality and basic action are both non-epistemic in Dretske's sense. And finally, it also seems very plausible that both normal human infants and toddlers, as well as at least some non-human animals, are capable of experiencing desire-based emotions without having any corresponding beliefs or propositional thoughts, whether occurrent or dispositional.

Another reason for doubting the thesis that desire-based emotions entail beliefs or propositional contents is the potential for error. We can be mistaken about precisely what we are desiring or feeling, and as a consequence can be quite confused as to precisely which desire-based emotion we are actually experiencing in that context. For example, in experiencing a certain intense emotion, Ken may think that he passionately loves Karen, when in reality he mainly feels very bitter and resentful towards her. Or alternatively Ken may feel very sad and firmly believe that this sadness is all about missing Karen, when in reality his sad feelings are mainly about an unresolved Freudian conflict with his mother.

Such examples are important in part because both non-epistemic perceptions and desire-based emotions have a unique content, structure, and psychological function, and cannot be wrong in the same way that our propositions, beliefs, judgments, or thoughts can be mistaken. If the direct

realist disjunctivism about sense perception and intentional action that we briefly sketched in Section 4.2 is true, then all non-epistemic perceptions are at least non-conceptually correct and veridical (even if not *conceptually or descriptively* correct), and cannot wholly fail to detect the real world. This is because, according to direct realist disjunctivism about perception, only correct, veridical perceptions are authentic perceptions, and because all correct, veridical perceptions put us in correct, direct, non-descriptive conscious contact with real objects and their properties. In the examples presented above, Ken is experiencing a "false emotion." But this is a misleading label. While desire-based emotions may be deemed *appropriate* or *inappropriate* (in relation to their objects and surrounding contexts), *genuine* or *phony* (in relation to the integrity and sincerity of the emotional agent), and *self-aware* or *self-deceived* (in relation to the agent's level of self-knowledge), a desire-based emotion is not accurately deemed true or false. This is because a desire-based emotion has a unique content, structure, and psychological function, and thus it is not subject to the same logical constraints of correspondence-to-the-facts, consistency, and consequence to which propositions, beliefs, judgments, and thoughts must adhere.

Emotional Intentionality$_{lo}$ and Perception

It is the sorts of considerations just rehearsed that have led many theorists of emotion to claim that emotional intentionality has more in common with sense perception that it does with belief or thought. Robert C. Roberts, for example, proposes "construal" as an alternative to judgment and maintains that when a person judges himself to be guilty without experiencing the emotion of guilt, what he lacks is a non-visual analogue of "seeing-as."[51] He judges himself to be guilty but does not construe himself as guilty. A man with obsessive fears about his house burning down, on the other hand, might construe his house as subject to great danger despite the fact that he also judges it to be highly improbable that this will happen.

In this connection, one obvious parallelism between sense perceptions and desire-based emotions is that even when the latter are under our control, they usually are experienced as simply *arising* and thus bear a resemblance to the sensations associated with external perception. Along

[51] Roberts, "Solomon on the Control of the Emotions," 399.

these lines, Sabine Doring holds that an emotion's representational content resembles the content of sense-perception in that (as, e.g., the fact of the persistence of the Müller-Lyer illusion and others even under changing conceptual and propositional information shows) it might not be revisable in light of belief or better knowledge.[52] In her view, emotions have a unique formal structure that makes them unlike beliefs, and more similar to perceptual evaluations that also have an inherently felt dimension. She therefore proposes that we understand an emotion's motivational force in terms of "affective perception."[53]

While we quite agree that the intentionality$_{lo}$ of the emotions has a greater structural resemblance to the intentionality$_{lo}$ of sense perception than it does to the intentionality$_{lo}$ of beliefs or thoughts, we also think that it would be fallacious to conclude that an emotion is a special *sub-species* of sense perception. Just because two kinds of things are similar, it obviously does not follow that one of them is a sub-species of the other. On the contrary, from the standpoint of our desire-based theory of the emotions, the fact that emotions are more similar to perceptions than they are to beliefs or thoughts depends on the deeper fact that emotions are *desire-based*, and on the further fact that *desires* are more similar to perceptions than either of them is to beliefs or thoughts. But because emotions and perceptions are indeed similar in certain respects, we can still exploit relevant analogies between emotional intentionality$_{lo}$ and perceptual intentionality$_{lo}$ in order to understand better the intentionality$_{lo}$ of desire-based emotion. One of these relevant analogies is between the role of affect or feeling in desire-based emotion, and the role of sensations or sensing in perception. Just as sensations and sensing in perception pinpoint the objectively real properties of perceived objects in a highly finegrained way, and explicate their salience—e.g., enabling me to see the characteristic red of that apple even in dimly-lit conditions or with a shadow thrown across it, or enabling me to see the characteristic shape of the apple even when it is partially occluded—so too affects and feelings in desire-based emotion pinpoint the cognitive objects and practical goals of those emotions in a finegrained or hyper-finegrained way, and explicate their salience. For example, the sharp difference between a friendly kiss and a lovers' kiss—here Rodin's *The Kiss* provides another bodily stereotype—is obvious to anyone old enough

[52] Doring, "Explaining Action By Emotion," 223. [53] Ibid., 220.

to care about different kinds of kissing, and yet it seems highly unlikely that this salient difference could be pinpointed or adequately explicated in terms of anything but affects or feelings in desire-based emotion.

In his later work, Solomon admitted that in *The Passions* he was too dismissive of feeling. However, while he recognized that feelings had indeed been "left out" of his earlier theory, Solomon still believed that cognition or judgment, properly understood, could capture this missing component. So despite retaining his cognitivism, he made a real effort to avoid the over-intellectualization of emotions that we isolated earlier as a serious problem for cognitive theories of emotion. In this way, he came to hold—correctly, from our point of view—that the cognitive component of emotion need not be articulate, self-conscious, or self-reflective.[54] Because beliefs and thoughts are propositional attitudes, while many emotions are not, belief or thought is not the right sort of thing to identify with an emotion. The sorts of appraisals that we make and the construals we perform are sometimes made without any articulation or reflection. So according to the revised version of Solomon's thesis that "emotions are judgments," although emotions involve recognition and response, this does not entail that emotional judgments are all doxic or propositional.

Solomon then went on to claim that the element of affect or feeling that is missing from cognitive accounts can be identified with the body. He pointed out that many of our cognitive responses have more to do with the habits and practices of embodied and motile beings than with reflective judgments, and that the phenomenology of emotional experience can be understood as intentional states directed towards the condition of one's body. Similarly, Lyons suggests that evaluations and wants together cause unusual bodily changes associated with the central nervous system and subjective feelings, and that this is sufficient for an emotion.[55] And on Antonio Damasio's view, while the evaluative process is very much a part of the emotion, this process is separate from feeling.[56] The evaluative process comes first, followed by a certain neurobiological state on which an emotion supervenes, and then a feeling. Such accounts all imply that emotional intentionality$_{lo}$ generates neurobiological processes and then associated affects or feelings.

[54] Solomon, "Thoughts and Feelings: What is a 'Cognitive Theory' of the Emotions, and Does it Neglect Affectivity?"

[55] Lyons, *Emotion*, 60.

[56] Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*.

Of course we welcome the proposal that emotions and embodiment are closely linked. On our account, since conscious, intentional $minds_{lo}$ are essentially embodied, and since emotions are desire-based, and since the phenomenal character or felt dimension of consciousness is grounded in primitive body awareness, there is no doubt whatsoever that the feelings associated with bodily changes and movements are a necessary part of emotion. Now as Solomon pointed out, feelings grounded in body awareness alone cannot reveal what that emotion is about or even which emotion you are experiencing. The feeling of one's heart racing, e.g., may be associated with anger, fear, rapturous love, or too much coffee. These bodily feelings are merely some of the bodily symptoms or manifestations of emotion and not what the emotion is about. The content and object of emotional $intentionality_{lo}$ are still required to explain the specific nature of that emotion. By identifying emotional feelings with bodily symptoms or manifestations, the theories described above imply that feeling has at best a secondary role in our emotional experience. However, to suppose that affective feeling is simply a matter of our awareness of bodily changes that occur *after* emotional $intentionality_{lo}$ has taken place is to adopt a very narrow view of emotional feelings. On our view, by sharp contrast, the essentially embodied $intentionality_{lo}$ of the emotions is necessarily *infused* with feeling. This affective infusion, in turn, plays a definite and ineliminable role in emotional $intentionality_{lo}$ by determining the cognitive and action-oriented *focus* of emotions.

Emotional $Intentionality_{lo}$, Feeling, and Affective Framing

If we are correct, then desire-based feelings play a definite and necessary role in the constitution of desire-based emotional experience in minded animals and its $intentionality_{lo}$, by determining both the *attentive focus* of the perceptual and cognitive element in desire-based emotions as well as the *goal focus* of the volitional and practical element in desire-based emotions. Or otherwise put, of all the contextually-presented options for cognition and intentional action, something must determine *precisely* what we attend to in cognition and *precisely* what we aim at in action—and for us this is the special role of affect or feeling. The issue of precise determination arises in the following way.

It is clear that cognitive attention and intentional goal selection are either finegrained or hyper-finegrained, in the sense that the natural stimulus bases

of such attention and selection significantly underdetermine them. By the notion of a "natural stimulus" we will mean a combination of causal impacts from the environment on a conscious, intentional animal together with the relevant neurobiological processes triggered by these impacts. Then we can ask: How many different ways are there of attending to any given natural visual stimulus? In other words, given a stream of incoming information, what parts of the stream will the conscious animal focus her attention on? For an individual cognizer with a consciousness$_{lo}$, the mapping from attention to the natural stimulus—i.e., the mapping from attention to the information stream—seems self-evidently to be many-to-one. Familiar examples of perceptual multistability like the Necker Cube phenomenon and the Jastrow duck-rabbit phenomenon show that the mapping from attention to the natural stimulus can even be uniformly underdetermined across our species, including all cognizers who possess the concepts CUBE, DUCK, RABBIT, and PICTURE. It is because cognizers attend to different aspects of a visual stimulus, *even* when their eyes are foveated so as to focus the same perceptual shape (as in the case of the Necker Cube phenomenon, which switches aspects spontaneously[57]), that they may see either face of the cube as standing forward towards them, or that they may see either the duck or the rabbit.

Similarly, then, we can also ask: How many different ways are there of selecting a given natural state of affairs as one's particular goal? Suppose that the natural state of affairs is an arm going up. The very same natural arm movement can be part of a wave, part of a dance, part of a stretching exercise, part of a political rally, part of a committee meeting, and so on and so forth. Presumably there is virtually no upper bound on the number of possible non-basic acts that can be associated with a given intentional body movement. So again, the mapping from goal selections to natural states of affairs is self-evidently many-to-one.

Therefore, it seems highly implausible that a natural stimulus could ever even remotely approach the sufficient determination of Sam's attentive visual focus on the shape as opposed to the color of Sarah's eyes when he gazes lovingly at her face. Similarly, it seems highly implausible that a natural state of affairs could ever even remotely approach the sufficient determination—shades of Buridan's Ass!—of my reaching out for this as

[57] See Hanna and Thompson, "Neurophenomenology and the Spontaneity of Consciousness".

opposed to that type-identical bottle of Buffalo Trace Kentucky Straight Bourbon Whisky on a closely-packed shelf. Only pre-reflectively conscious desire-based emotional feeling, with its myriad of kinds, shades within kinds, and degrees of intensity seems to be as *finegrained* and even as *hyper-finegrained* as attentive focusing in cognition, and as finegrained and even as hyper-finegrained as goal focusing in action.

Just to be as clear as possible, by "finegrainedness" in this context we mean

> differences in the *internal structures* of any two representational contents R_1 and R_2, consistently with *the co-extensionality of* R_1 *and* R_2 *in the actual world or across all possible worlds*,

and by "hyper-finegrainedness" in this context we mean

> differences in the intentional-agent-centered *presentation or evaluation* of any two representational contents R_1 and R_2, consistently with *the same internal structure and co-extensionality of* R_1 *and* R_2 *in the actual world or across all possible worlds*.

So, e.g., the difference between the representations *creature with a heart* and *creature with a kidney* is a finegrained difference, and so too is the difference between the representations *triangular* and *trilateral*, while the difference between *friendly kiss* and *lovers' kiss* is a hyper-finegrained difference. Insofar as feeling plays a finegrained and hyper-finegrained focusing role in emotional intentionality$_{lo}$, we call it *affective framing*.

As the previous discussion will have made clear, while all desire-based emotional intentionality$_{lo}$ does involve an element of appraisal, this element is not best understood in terms of evaluative *judgment*, for the content of an emotion need not be understood as the content of a belief or the object of a thought. The fact that emotional experience can be attributed to normal human infants and at least some non-human animals demonstrates that desire-based emotion sometimes involves non-propositional, non-judicative, non-belief-based, and non-thought-based engagement with the external world. Goldie notes that while it is possible to recognize something as dangerous without feeling fear, there is a special affect-charged way of recognizing something as dangerous that does entail fear.[58]

[58] Goldie, *The Emotions*, 36.

This affect-charged mode of recognition is what he calls "feeling-toward," or thinking of with feeling, which is for him a matter of perceiving or imagining an object in some way. Goldie points out that the content of a thoughtful recognition that is infused with affect is quite different from the content of a recognition not infused with affect. Somewhat similarly, as we have already seen, Doring offers her notion of affective perception as a way to conceive of emotional sensitivity to the world and its opportunities for action. On her view, an emotion's intentional content is representational content directed at a particular object or target, which resembles the content of sense-perception.

While we are sympathetic with the general spirit of such accounts, we also believe that they are mistaken in two important respects. First, given that the mode of recognition involved in emotions need not be in any way articulate or reflective, it is mistaken to associate it with thinking, as Goldie does. Second, while emotion is more like perception than it is like judgment, belief, or thought, nevertheless emotion is *unlike* sense perception insofar as the intentional objects or targets of emotion need not be immediately present or actually given. Desire-based emotions arise not merely in response to causal impacts upon the senses, but also frequently as a result of having memories or imagining scenarios. So it is a mistake to construe desire-based emotion as affective perception, as Doring does.

But if emotional intentionality$_{lo}$ is fundamentally neither judgment, nor belief, nor thinking, nor perception, then what is it? Our answer should be easy enough to anticipate. For us, emotional intentionality$_{lo}$ is pre-reflectively conscious desire-based intentionality whose attentive focusing and goal focusing are both sufficiently determined by affective framing. Something must sufficiently determine, right down to the most hyper-finegrained levels, what we specifically attend to in emotional perception, memory, or imagination. Correspondingly, something must also sufficiently determine, right down to the most finegrained and hyper-finegrained levels, what we specifically project as a goal in willing or effortless trying and its active guidance. These are obviously jobs that the affective frames of feeling are well qualified to carry out, since they (and, it seems, they alone) are as finegrained and hyper-finegrained as phenomenal character itself. To be sure, the affective frames of feeling are not required *only* for finegrained or hyper-finegrained differences in attentive focusing and goal focusing, but are operative at *all* levels of grain in attentive focusing and goal focusing,

including the most roughgrained levels. For example, if Raymond is very shy, or socially insecure, and has to attend a large and noisy party, he may find the whole experience just one big terrifying blur and screen out most of the important differences.

As animals with consciousness$_{lo}$ and intentionality$_{lo}$ navigate their way through the world, obviously they do not sequentially process all of the cognitive and practical information that is potentially available to them, but instead almost always home in on certain very specific things rather than others. If intentionality$_{lo}$ in general and emotional intentionality$_{lo}$ in particular did not involve an underlying process of affective framing, then agents in the world would be faced with a potentially endless array of possible cognitive and volitional options, and presumably would merely shut down like so many massively overloaded word-processors. So an affective frame is an *egocentrically-centered and spatially oriented, thermodynamically irreversible, essentially embodied, finegrained and hyper-finegrained emotional map,* or an *emotional sensorium*, that conscious, intentional creatures rely on for finding definite points, lines, and contours of salience in the complex world around them, in order to orient themselves in that world, reduce its otherwise overwhelming clutter to something first-personally manageable, confer upon it specific cognitive significance and specific purpose, and then get on with their forward-flowing lives.

Now of course, affective framing is not our *only* method of reducing the clutter in order to make way for cognition and action. Concepts help us to organize complex information into coherent categories, allowing for the logical organization and simplification of descriptive information. Or in other words, concepts get our cognitive and practical encounters with the world ready for judgment, inference, and self-conscious or self-reflective deliberative intentions. Affective framing, on the other hand, occurs during an essentially embodied sensorimotor-subjective experience of the world that is independent of conceptual and propositional information processing, and yields a *pre-reflectively conscious* finegrained and hyper-finegrained emotive mapping of that world, so that we can immediately focus our cognitive attention and our practical goals. Here is another beautiful example, this time from Austen's *Pride and Prejudice*:

As they walked across the lawn towards the river, Elizabeth turned back to look again; her uncle and aunt stopped also, and while the former was conjecturing as

to the date of the building, the owner of it himself suddenly came forward from the road, which led behind it to the stables. They were within twenty yards of each other, and so abrupt was his appearance, that it was impossible to avoid his sight. Their eyes instantly met, and the cheeks of each were overspread with the deepest blush. He absolutely started, and for a moment seemed immovable from surprise; but shortly recovering himself, advanced towards the party, and spoke to Elizabeth, if not in terms of perfect composure, at least of perfect civility. She had instinctively turned away; but, stopping on his approach received his compliments with an embarrassment impossible to be overcome. Had his first appearance, or his resemblamce to the picture they had just been examining, been insufficient to assure the other two that they now saw Mr Darcy, the gardener's expression of surprise, on seeing his master, must immediately have told it. They stood a little aloof while he was talking to their niece, who, astonished and confused, scarcely dared lift her eyes to his face, and knew not what answer she returned to his civil inquiries after her family.[59]

This of course is the moment when Elizabeth Bennett finally begins to realize something the reader has known for a long time—that she loves Darcy. Until that very moment, and indeed still *at* that very moment, she self-consciously or self-reflectively believes that she does not love Darcy. But Austen's pellucid prose shows us that the opposite is true. Note that Elizabeth's radically new way of emotionally representing Darcy occurs impulsively, and does not involve analysis, inference, or reflection. Affective framing—the world as intended through an emotional sensorium—provides a way of discriminating, filtering, and selecting information that is immediate, direct, Gestalt-like, non-inferential, and pre-reflective.

Just to wrap up this phase of our argument, we will now briefly compare and contrast our account of the emotions as essentially embodied, desire-based, weakly Behaviorist, and pre-reflectively affectively framed, with the "embodied appraisal" theory recently worked out by Jesse Prinz. Prinz rightly denies that emotions require conceptualized appraisals or necessarily are constituted by propositional attitudes. In his view, a state is cognitive just in case it involves representations that are controlled by structures in executive systems, and which are activated, maintained, and manipulated by the organism rather than by the environment. Because he believes that emotions are passive and often not under the organism's control, he

[59] Austen, *Pride and Prejudice*, 365.

concludes that most of the time, emotions are not cognitive. Instead, he understands emotions as mental states that detect and register bodily changes, represent objects or events as having some bearing on one's interests and concerns, and thereby track organism-environment relations. For example, fear registers an array of bodily changes, tracks danger, and represents events as posing a threat to one's interests and concerns. As Prinz points out, danger is a relational property, for "something can be dangerous only to some creature or other, and whether or not something is dangerous depends on the creature in question."[60] Drawing from the work of Paul Lazarus, Prinz describes the relational properties that pertain to well-being as "core relational themes," and claims that emotions track these core relational themes by registering changes in the body.[61] Insofar as certain bodily changes reliably co-occur with certain organism-environment relations (i.e., the core relational themes), emotions use our bodies to tell us how we are faring in the world.

So emotions represent organism-environment relations by registering and tracking patterned physiological responses. In this way, they are just "gut reactions" that make it unnecessary to think in some cases, for our bodily feelings convey information about our well-being. For example, according to Prinz, a fear representation becomes active when a sufficient number of the bodily changes that can occur in a dangerous situation is detected.[62]

But how we are faring in the world, as Prinz readily points out, is not entirely an objective matter, for core relational themes are grounded in our needs and interests. Because our concern for our well-being goes well beyond our instinctive desire to survive, whether organism-environment relations are important and whatever significance they have depend to a large extent on what the creature in question cares about and desires. This is to say that emotions are elicited by things as they relate to us, or, as *we* would phrase things, only in relation to what we pre-reflectively desire. When Prinz characterizes sadness, e.g., as a representation of the loss of something valued, we would describe this as the loss of something one pre-reflectively *cares* about. What is crucial is that whether one experiences bodily changes that provide the basis for sadness when one becomes estranged from a co-worker depends on whether and to what extent one desired and cared

[60] Prinz, *Gut Reactions,* 63. [61] Ibid., 68. [62] Ibid., 73.

about that relationship. Likewise, whether one experiences the bodily changes that provide the basis for fear when one sees a large spider depends necessarily on whether one affectively frames the spider as something that undermines one's desires and whether one cares about being as far away from spiders as possible.

In other words, if I have not affectively framed the situation as being in conflict with my finegrained and hyper-finegrained pre-reflectively conscious desires, my heart will not race and my pulse will not quicken in just the way it does. Once these bodily changes occur, pre-reflectively conscious desire already has been on the scene. Therefore, the roughgrained organism-environment relations that Prinz posits will always *underdetermine* the specific character of emotional responses in human minded animals, and only pre-reflectively conscious affective framing can adequately explain the finegrained and hyper-finegrained variation that actually occurs.

If this line of argument is sound, then what we pre-reflectively desire plays a necessary role in explaining the total range of mappings from core relational themes to specific emotional responses in us, and thus the particular bodily changes we undergo cannot be separated from pre-reflectively conscious affective framing. Indeed, if our claim that emotions are essentially embodied is correct, it is strange to suppose that bodily changes come first, and that emotions come along subsequently to monitor and detect these changes. Instead, we believe that *pre-reflectively conscious desire and caring* are at the root of Prinz's core relational themes, and that these desires and caring must always be present if the bodily changes Prinz describes are to occur in all the multifariously different ways that they do occur. So while it may very well be true that an emotion corresponds in a roughgrained way to a broad body state prototype and that there is a range of bodily changes that reliably co-occur with core relational themes, the supposition that every emotion has distinctive patterns of activation mistakenly implies that bodily changes come first and are then followed by emotional experience. On the contrary, the desires involved in emotion are essentially embodied, and when these desires seem to be either fulfilled or thwarted, the sorts of changes in bodily profile that Prinz describes directly correspond to an affectively framed change in pre-reflectively conscious desire-based emotions.

Interestingly, our view of emotion as pre-reflectively desire-based and affectively framed is supported by Prinz's discussion of "valence," or the positive or negative tone of emotion. Valence is essential to emotionality,

in Prinz's view, and he equates it with inner reinforcements that serve as imperatives that have an impact on future behavior. Emotions call for "More of this!" or "Less of this!," so that negative emotions encourage us to withdraw from situations that elicit them, and positive emotions encourage us to seek out the situations that elicit them. We believe that this characterization of emotions as valent, embodied appraisals is just a way of saying emotions are in fact grounded in finegrained and hyper-finegrained intentional wants and needs and that they always involve pre-reflectively conscious desires to move our bodies in some highly specific way or another, in order to express that very emotion. Indeed, we find it reasonable to suppose that pre-reflectively conscious desire and caring are at the foundation of valence, and that this is ultimately what allows emotions to play an active and essential role in shaping intentional behavior.

Prinz, of course, denies the necessary link between emotion and desire and maintains that although the bodily changes that emotions involve are "action enabling," emotions are not simply action tendencies or motivations.[63] In his view, while valence tells us to change how we are feeling, motivation tells us to change how we are acting. We, on the other hand, have argued that emotions always involve a desire to move one's body in some way or another, either in order to change how we are feeling, how our bodies are oriented or positioned, or how we are acting. Such bodily movement need not be goal-oriented, and may amount to no more than impulsively changing one's posture, tone of voice, or facial expression. But desire-based emotions normally do impel and drive one to act or move in some way or other so as to express the precise hierarchically conative constitution of one's will in that situation, and this is what makes emotions intrinsically motivating for the intentional actions that do flow naturally from one's emotions.

5.4 Invasion of the Body Snatchers: Emotional Self-Control and Emotional Zeroes

Solomon repeatedly—and we think very correctly—criticizes the reason vs. emotion dichotomy. Indeed, the very idea that there *really is* a reason

[63] Prinz, *Gut Reactions,* 194.

vs. emotion dichotomy constitutes one of the most deeply entrenched of all our philosophical pictures, going back at least as far as Plato's doctrine of the tripartite soul in the *Republic* and *Phaedrus*, where it appears as the dichotomy between the masterly "reasoning" part of the soul on the one hand, and the subservient "spirited" and "appetitive" (or passionate) parts of the soul on the other. Hume's famous eighteenth-century assertion that

> reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them,[64]

Schopenhauer's nineteenth-century doctrine of the basic metaphysical and psychological contrast and conflict between "representation" and "will," and Freud's early twentieth-century distinction between the primitive psychic functions of the *superego* and *ego* on the one hand, and the *id* on the other, all assert the classical Platonic dichotomy. It is also particularly interesting to note how in Hume, Schopenhauer, and Freud the "reasoning" part of the mind has become fundamentally subservient to the "emoting" part. In any case, according to the general picture of the reason vs. emotion dichotomy, emotions are taken to be inherently disruptive and overwhelming, psychic compulsions or forces not under our direct control. Correspondingly, our emotive feelings are also taken to be intrinsically passive, or as if they were always on the verge of being helplessly victimized by another part of ourselves. Emotions are held to be at best *arational* and at worst downright *irrational*.

In short, the dichotomous picture tells us both that our rationality is inherently *non*-emotional, and also that our emotions are inherently *non*-rational. This deeply-entrenched philosophical picture has had, and still has, some very profound and not always altogether beneficial or benign implications for philosophy of mind, action theory, and ethics.

Obviously it cannot be denied that our emotions do sometimes overcome us and cause us to deviate from ideals or standards of rationality, either by merely failing to do rational things (arationality), or by doing downright perverse things (irrationality). As Goya famously pointed out in the text of *Los Caprichos*, "the sleep of reason breeds monsters." However, this obvious fact should not lead us to conclude either that the emotions are

[64] Hume, *Treatise of Human Nature*, book II, part III, section iii, 415.

in any way *intrinsically* opposed to rationality, or that the experience of the emotions is *intrinsically* passive. It is a crucial fallacy to think that the true proposition

> (1) Emotions sometimes cause us to act in ways that deviate from our self-conscious or self-reflective deliberative intentions and good reasons for action,

entails the proposition

> (2) Emotions are not the primary causal factors in all intentional action, including self-conscious or self-reflective, deliberative action.

According to our Desire-Overriding Internalism about reasons (see Section 3.4), it is the desire-based emotions that move us when we intentionally move our bodies in pre-reflective or spontaneous actions by means of effortless trying; it is the desire-based emotions that move us when we impulsively deviate from our self-conscious or self-reflective, deliberative intentions; and it is the desire-based emotions that move us *also* when we act in accordance with self-conscious or self-reflective, deliberative intentions—sometimes even contrary to egoistic, self-interested, hedonic, or consequence-driven desires. Therefore the second proposition is clearly false.

Indeed, for us, even *relatively desireless* emotional states are desire-based emotive causes of action. For example, even Eve's feeling highly depressed and listless, and so unable to choose or do some things she would normally self-consciously and deliberatively choose or do, *still* causes her to spontaneously and impulsively perform various depressed and listless intentional body movements—crying, frowning, groaning, sighing, slouching, slumping, staring off into the middle distance, and so on. In other words, she is still moved by the emotion-based desire to impulsively move her body in such a way as to express her current hierarchy of desires. She is highly depressed and listless, but not *catatonic*. In this way, even to be adversely affected by one's desire-based emotions and thereby fail to act on self-conscious deliberative intentions through a corresponding "lack of desire," is *not* thereby to fail to act intentionally on the basis of desire-based emotions. Rather, it is only to fail to act intentionally on the basis of those special desire-based emotions that are normally mobilized by self-conscious deliberative intentions.

According to our Emotive Causation Thesis, pre-reflectively conscious effective first-order desires (in a dynamic hierarchy of conscious desires, and normally along with the further impulsive pre-reflectively conscious effective first-order desire to move one's body in such a way as to express that very set of desires), and thus desire-based emotions, are the basic mental causes of all kinds of intentional action, in the guise of effortless trying and its active guidance of intentional body movements. So if we are correct, then desire-based emotion is the unmoved but also essentially embodied motor of all rationality in action

The belief in the supposed intrinsic non-rationality and passivity of the emotions has been recently and rightly opposed by cognitive theories of emotions, which as we have seen, insist that emotions are belief-desire pairs or judgments. But as we also have seen, this approach commits the equal and opposite error of over-intellectualizing the emotions. From the standpoint of the Essentially Embodied Agency theory, it seems to us very likely that the basic problems of both the non-rationalist/passivist and the cognitivist/activist approaches to emotion alike stem from a *single* mistaken philosophical tendency that has been and still is rife in the philosophy of mind, action theory, epistemology, and ethics, not to mention the metaphysics of free will and personhood—namely, *the mistaken tendency to deny or depreciate the role of embodiment*. Of course feminist philosophers and existential phenomenologists have been saying this—without much response from mainstream analytic philosophers—ever since the publication of Edith Stein's *On the Problem of Empathy* in 1917 and Maurice Merleau-Ponty's *Phenomenology of Perception* in 1945, and now many contemporary cognitive scientists are saying it too. But it seems to us that when philosophical push comes to philosophical shove, *only* the metaphysics of essential embodiment and the Essentially Embodied Agency Theory of action can adequately elaborate and justify these important claims. So if we are correct that conscious, intentional minds_{lo} are necessarily and completely neurobiologically embodied, and if all mental causation is emotive causation via synchronous pre-reflectively conscious effortless trying and its active guidance of our intentional body movements, then non-rationality and passivity are merely derivative and secondary features of *some* of our emotions, and not intrinsic to *any* of our emotions, much less an intrinsic feature of *all* of them.

Otherwise put, the intentional movements of our living organismic body are not inherently removed from the control of reason by our emotions,

such that we are thereby their passive victims. *Desire-based emotions are* ***not*** *body-snatchers*. On the contrary, according to our view, necessarily for creatures with consciousness$_{lo}$ and intentionality$_{lo}$, the emotions are desire-based, and thus they live, move, and have their being only in essentially embodied agency. The intentional movements of our living body are the inherent *realization* of our emotions, of which we are then the active *shapers*, and when rational and self-conscious or self-reflective, the *rationally* active shapers. It is not that our emotions are the *same* as our actual or possible intentional movements—that would be classic Behaviorism about the emotions, which we think is false. Instead, we are saying that without the possibility of intentional body movements *we would have no emotions at all,* and also that the structure of each desire-based emotion is completed and perfected by the intentional movements that inherently would, and normally do, pre-reflectively impulsively express it.

Bodily paralysis is no counterexample to this thesis. Partially paralyzed people are still able to shape, complete, and perfect their emotions via overt intentional body movements, even though they are incapable of making *certain* overt intentional body movements. But everyone knows that even just a *wink* can be highly emotionally expressive. More generally, *any* sort of overt or covert intentional body movement can shape, complete, or perfect an emotion. The harder case of *complete* paralysis together with conscious alertness—e.g., in locked-in syndrome, or curare poisoning—is of course quite rare, and so there is little empirical evidence to go on. But there is at least a rough analogue of this provided by the subjective experience of (often extremely vivid) emotions in dreams during deep sleep, in which there is a kind of temporary complete paralysis of the body. Here the distinction we made in Section 3.0 between the *covert* intentional body movements of synchronously actively guided neurobiological processes, and the *overt* intentional body movements of synchronously actively guided behavior, is crucial. The subjective experience of desire-based emotions in complete paralysis or during dreams in deep sleep, then, is just when, whether by catastrophic accident or by natural biorhythmic cycles, our *overt* intentional body movements have been reduced to zero or nearly to zero, even though our pre-reflective or spontaneous, impulsive, and self-expressive *covert* neurobiological intentional body movements are still occurring. So our desire-based emotions are still essentially embodied even when complete bodily paralysis prevents our overt movements.

In this way, the Emotive Causation Thesis directly yields a theory of emotional self-control. If emotions are essentially desire-based, and if the will is essentially a dynamic hierarchical structure of desires, and if willing or effortless trying and its active guidance, as the mental cause of all action, is the same as pre-reflectively conscious effective first-order desire, then the conscious mental process by which we are constantly configuring and re-configuring our desires in order to constitute our wills for the purpose of mental causation, is *identical* to the process of emotional self-control. We will now work out some of the further details of this theory.

To be sure, desire-based emotions sometimes give rise to various bodily events that are relatively involuntary, and merely passively enjoyed or suffered. Butterflies in the stomach; changes in the tone, steadiness, or volume of one's voice; chills; flushing; grimaces; sexual arousal; nervous laughter; quickening of the pulse; tears; and so on—of course, these frequently accompany emotional episodes. Solomon claims that these bodily events are mere involuntary *symptoms* of the emotions rather than voluntary *expressions* of the emotions.[65] But while this is true to some extent, it also seems to be obviously true that there are varying degress of voluntariness, even within the standard cases of the so-called symptoms of the emotions. It seems to us to be obviously true that these so-called symptoms can always, *to some non-trivial extent*, be modulated into expressive *vehicles* of the desire-based emotions in the very same episodes that begin involuntarily. For example, children—and sometimes also lachrymose undergraduates—can accentuate and prolong their initially involuntary crying in order to get what they want, or to avoid criticism and punishment. Stage actors, and especially Stanislavsky "method" actors, can learn how to put themselves in states in which they begin to cry or laugh involuntarily, and then can carefully control these emotional episodes in order to play their parts. And parents or teachers can exaggerate and prolong their initially involuntary frowns in order to signal serious disapproval with their children or students.

Again, consider normal anger. At some downstream point in the process of an initially involuntary episode of normal anger, as evidenced by various passively suffered bodily symptoms (say, a sudden involuntary increase in the loudness of voice), it seems that we can *always* choose either to continue

[65] Solomon, "On the Passivity of the Passions," 221.

to work ourselves up into an absolute fury (say, by trying to shout even louder) or to begin to calm ourselves down (say, by trying to lower one's voice). Thus it seems that, even if it begins involuntarily, the process of normal anger can always eventually be voluntarily escalated or suppressed to some extent. To be sure, for people with pathological anger management problems, it is a very different story. But that is *abnormal* anger, and not what we are talking about.

This point generalizes. Although the onset of the bodily symptoms of emotions is sometimes very rapid, and indeed too fast for us to self-consciously or self-reflectively recognize what is happening, it seems to us to be always the case that at some point downstream in the very same emotional process, we become able, *to some non-trivial extent*, to modulate our bodily movements, or sometimes even to end or interrupt the whole process. To take a trivial example, the familiar butterflies in the stomach normally experienced by most college and university lecturers gradually modulate into animated lecturing. But even prior to lecturing, one way of interrupting or ending the butterflies is to jump up and down in the privacy of one's office to the point of absurdity, followed by relaxing self-amusement.

Less trivially, many people are able to configure and re-configure their wills, modulate their emotions, and thereby significantly improve their lives, by the use of breathing, relaxation, or rhythmic body movement techniques learned and honed in dancing classes, long distance running, Buddhist meditation, yoga, *tai chi*, and the martial arts. This is also direct empirical evidence for the role of essential embodiment in emotion, since it is precisely by intentionally *moving their bodies in a certain way* in dancing classes, long distance running, Buddhist meditation, yoga, and so on, that people shape and re-shape their emotions. Indeed, in many ways the essential embodiment thesis fits well with Eastern philosophical perspectives that emphasize these meditative practices.

If this general line of argument is sound, then emotions are fundamentally *not* like allergic reactions, fevers, or spasms, even if their phenomenology can occasionally be indiscriminable to the conscious subject. Furthermore, if some episode that is self-consciously or self-reflectively subjectively indiscriminable from an emotion does in fact operate like an allergic reaction, fever, or spasm, and is completely out of someone's control, then it is merely *pathological* and not a genuine desire-based emotion in our

sense. So we are direct realist disjunctivists about the emotions, just as we are direct realist disjunctivists about perception and intentional action. In short, we are saying that even in clear cases of emotional passivity normally we can always *eventually* volitionally catch up with the neurobiological processes and overt movements of our bodies within the same emotional episode, and then actively guide our body movements. In this way, normal desire-based emotions are always either directly open to the pre-reflective active guidance of body movements or else *on the verge* of this sort of pre-reflective active guidance.

Even given our direct realist disjunctivism about the emotions, however, this is *not* to say that there cannot be borderline cases or vagueness, at least of an epistemic sort, where it is going to be very and perhaps even almost impossibly difficult to tell whether a given psychological state, event, or process is a genuine desire-based emotion, or a pathological counterpart. The phenomenon of *addiction*, for example, is a rich source of just such borderline cases. It seems correct to say that some and perhaps even many cases of addiction are genuine desire-based emotions, experienced as extremely intense pre-reflectively conscious effective first-order desires, hence effortless tryings, that can be actively guided to some extent and so are open to some level of emotional self-control, via the configuration and re-configuration of the will, and effortful trying. It also seems correct, however, to say that some other cases of addiction are pathological counterparts of such emotions, and uncontrollable. Given the possibility of epistemic indiscriminability, it may sometimes be impossible for the first person to be able to tell a genuine addictive desire-based emotion that is to some extent controllable, from a pathological counterpart that is uncontrollable. So one source of the intensely controversial personal, social, and political difficulties associated with addiction, it seems, is the fact that the concept ADDICTION is a semantic mongrel that indiscriminately picks out cases on either side of the divide between desire-based emotions and pathological counterparts.

In any case, it is also manifestly *not* true that all of our emotions are experienced as passive feelings or accompanied by involuntary bodily symptoms. On the contrary, one can "work oneself up into" a particular emotion just by imagining, thinking, or remembering. Solomon notes that there is a variety of ways that we might influence our emotions—by seeking new influences, placing ourselves in the appropriate

circumstances, striving to understand our prejudices, provoking argument, and looking for evidence.[66] One can choose mental activities or external situations that are conducive to certain emotions, and also choose to spend time with people who tend to prompt particular emotions. So while it is true that an emotional agent cannot simply choose to hate someone else, he can make himself feel malice towards that person by performing other activities. For example, by mentally replaying, over and over again, the experience of covertly watching his girlfriend Kate talking to his friend Kevin at a party, Karl can turn himself into a monster of jealousy and so move himself all the way to cold hatred. For whatever reason, Karl *wants* to hate Kevin or Kate, and so chooses and acts accordingly.

This exemplifies the extremely important point that even though every emotion is itself a hierarchically-ordered set of conscious desires (normally together with a further pre-reflectively conscious effective first-order desire to impulsively move one's body in such a way as to express that very set of desires), we can also form further higher-order desires concerning our *emotions* themselves. For example, Theresa can desire to stop feeling angry at Tom. Then, perhaps, by remembering how sweet Tom usually is, by thinking about how life is much too short to waste any of it by engaging in pointless and unproductive fits of anger directed at people she truly loves, and by telling herself how absurd and laughable her anger is, she does indeed manage to stop feeling angry at Tom. When such higher-order desires do indeed determine changes in our desire-based emotions, this is a paradigmatic kind of emotional self-control that operates in essentially the same way as ordinary second-order volitions in relation to effective first-order desires. At some higher-order level, we want our emotions$_d$ to be different from what they are, and also for the effective first-order desires associated with these preferred emotions$_d$ to guide our intentional body movements. Indeed, since every desire-based emotion normally includes a further pre-reflectively conscious effective first-order desire to impulsively move one's body in an emotionally self-expressive way, the second-order desires involved in paradigmatic emotional self-control constitute merely a special species of *will-restructuring* desire-based emotions.

[66] Solomon, "Emotions and Choice," 261.

The Impossibility of Emotional Zeroes

So far in this chapter we have been employing an inference-to-the-best-explanation strategy by criticizing various contemporary philosophical approaches to the emotions, and by noting the explanatory payoffs of our theory that all emotions in animals minded like us are desire-based and thereby necessarily connected with intentional agency. Now we want to switch methods, and supplement our case by developing an a priori conceptual argument for the thesis that necessarily all creatures with consciousness$_{lo}$ have desire-based emotions. This will also set us up for the use of several other a priori conceptual arguments in Section 6.3.

The appeal to a priori conceptual arguments in the philosophy of mind goes back as far as Descartes's property dualist argument for the "real distinction between mind and body" in the sixth *Meditation*. More recently, Descartes's arguments have been retrofitted for compatibility with contemporary modal logic and modal semantics by Saul Kripke, Thomas Nagel, Frank Jackson, David Chalmers, and others under the rubrics of *the modal argument, the gap argument, the knowledge argument, the zombie argument*, and *the inverted qualia argument*.[67] Since the mid-90s, however, all these a priori conceptual non-reductive arguments have come under a concerted and extended series of counter-attacks by contemporary empiricists. In particular, using the *other* Kripkean idea of necessary truths known only by empirical means—the necessary a posteriori—contemporary empiricists have argued both

(i) that there is no generally valid inferential step from conceivability to logical possibility,

and

(ii) that even if there were a generally valid inferential step from conceivability to logical possibility, non-reductivists still confuse imaginability with conceivability.

Conceding the existence of the necessary a posteriori, Chalmers has undertaken to reply to the contemporary empiricist critics of a priori conceptual non-reductive arguments by arguing that under certain epistemically

[67] See, e.g., Kripke, *Naming and Necessity*, 144–55; Nagel, "What is it like to be a bat?"; Jackson, "Epiphenomenal Qualia"; and Chalmers, *The Conscious Mind*, ch. 4, and 263–6.

ideal conditions of conceivability there is a valid inferential step from conceivability to logical possibility, and that the classical a priori conceptual arguments for non-reduction do indeed occur under these epistemically ideal conditions of conceivability.[68] This debate is currently still very vigorous, although unresolved, and shows little evidence of being resolved in the near future.

Another line of reply to the contemporary empiricist critics would be to reject the very idea of the necessary a posteriori.[69] But except for a few contemporary Kantians prepared to argue for the necessary equivalence of necessity and apriority, this is a decidedly unfashionable and uphill route to pursue.

Nevertheless there is at least one *other* line of reply to contemporary empiricist critics of a priori conceptual arguments that seems more promising. Instead of focusing on conceivability or the necessary a posteriori, this line of reply focuses on the very idea of *logical possibility*.

Formal or symbolic logic is the science of the necessary relation of consequence that holds between the premises and conclusion of a valid argument, and one of the central purposes of logic is to capture and express in a formalized and rigorous way our basic intuitions about valid inferences—including, of course, valid inferences concerning necessity and possibility. In contemporary philosophy of logic, it is a widely-accepted fact that there exist *non-classical* logics, including both *extended* logics and *deviant* logics.[70] Now classical logic is elementary logic: bivalent polyadic first-order predicate logic with identity. By contrast, extended logics preserve all of the theorems of classical elementary logic, although they add new theorems, rules of inference, rules of interpretation, or rules of syntax. Examples of extended logics would include modal logic (which introduces modal operators), deontic logic (which introduces deontic operators), epistemic logic (which introduces epistemic operators), second-order logic (which allows for quantification over properties, sets, or functions), and free logic (which allows for the occurrence of non-referring names). The crucial point is that extended logics are all *conservative* with respect to classical or elementary logic.

[68] Chalmers, "Does Conceivability Entail Possibility?"

[69] But see, e.g., Hanna, *Kant, Science, and Human Nature*, chs. 3–4; and Putnam, "Is Water Necessarily H_2O?".

[70] See Haack, *Deviant Logic*; and Priest, *An Introduction to Non-Classical Logic*.

Deviant logics, on the other hand, are *radical* with respect to classical or elementary logic, insofar as they reject at least some of the theorems of classical elementary logic, and add new theorems, rules of inference, rules of interpretation, or rules of syntax. Examples of deviant logics include intuitionist logic (which rejects the universal law of excluded middle), three-valued logic (which rejects the universal law of bivalence), and dialetheic logic (which rejects the universal law of non-contradiction). The only putative non-classical logics that seem to be ruled out altogether are those that have the effect of entailing that *every* sentence or proposition whatsoever is *both true and false*.[71] That would be utter cognitive chaos, the end of logical rationality, and presumably also the End of the World As We Know It.[72] But short of that, anything goes.

If non-classical logics exist, as most contemporary philosophers of logic fully admit, and if a central purpose of logic is to capture and express in a formalized and rigorous way our basic intuitions about valid inferences, including valid inferences concerning necessity and possibility, then it seems that there must be *some* non-classical logics that capture and express our basic intuitions concerning valid inferences from conceivability to possibility. Indeed, there must be *some* non-classical logics that capture and express an *identification* of possibility with conceivability. Let us assume that conceivability is construed liberally, so as to allow for conceivability under either non-ideal or ideal epistemic conditions, and also to include imaginability, short of permitting any explosive contradictions in any line of reasoning in which it occurs. Call this *liberal conceivability*. Therefore, in some or another non-classical logic, in which liberal conceivability just *is* logical possibility, there will *automatically* be a generally valid inference from liberal conceivability to logical possibility.

Let us now dub this logic *the A Priori Argument Logic*, or the APA logic for short. Precisely what sort of deviant logic the APA logic is, and the exact details of its formalization, do not matter for our purposes. The crucial facts are only, first, that the APA logic clearly falls within the range of minimally acceptable non-classical logics *by containing (we hereby stipulate) an "anti-End-of-the-World principle" that prevents it from ever being the case that every proposition or sentence in the system is both true and false*, and, second,

[71] See Putnam, "There is at Least One *A Priori* Truth."
[72] See Hanna, *Rationality and Logic*, chs. 2 and 7.

that we will be using the APA logic whenever we offer a priori conceptual arguments in this book.

The metaphysical effect of our adopting the APA logic is to allow for finegrained or hyper-finegrained identities and differences between properties, and also for connections of logical necessity and logical (im)possibility that track these finegrained or hyper-finegrained identities and differences. Again, by "finegrainedness" in this context we mean

> differences in the *internal structures* of any two representational contents R_1 and R_2, consistently with *the co-extensionality of* R_1 *and* R_2 *in the actual world or across all possible worlds*,

and by "hyper-finegrainedness" in this context we mean

> differences in the intentional-agent-centered *presentation or evaluation* of any two representational contents R_1 and R_2, consistently with *the same internal structure and co-extensionality of* R_1 *and* R_2 *in the actual world or across all possible worlds*.

So, e.g., it is liberally conceivable and therefore logically possible in the APA logic that something could be a creature with a heart but not a creature with a kidney (and conversely). It is also liberally conceivable and therefore logically possible in the APA logic that something could be trilateral but not triangular (and conversely).

In short, to adopt the APA logic is simply to open up the notion of logical possibility wide enough for it to let in liberal conceivability. Any intensional difference short of complete intensional identity can be picked out by liberal conceivability, just as any complete intensional identity can be picked out by liberal inconceivability. For example, it is liberally *in*conceivable and therefore logically *im*possible that something could be an oculist and also fail to be an eye-doctor (or conversely). But surely *that is exactly what a priori conceptual arguments were designed to do*. So adopting the APA logic permits us to do just what a priori conceptual arguments were designed to do, without any fear of a general inferential gap ever appearing between conceivability and logical possibility.

Please notice—and this bears emphatic repetition—*please notice* that we are *not* saying that by adopting the APA logic it follows that anything anyone ever says about conceivability or inconceivability is automatically acceptable as a philosophical claim about possibility or impossibility. Substantive

justification of these claims is required in each case. In order to be legitimate or valid, the inferential step from conceivability to possibility must yield substantive information about corresponding properties. But our adoption of the APA logic defuses the *in-principle* worry about the general validity of the inferential step from conceivability to possibility, so that the difference between conceivability intuitions and *logic* is never going to stand in the way of our chains of conceptual reasoning.

Therefore the methodological effect of our adopting the APA logic is just to level the dialectical playing field for our debate with contemporary empiricists. When they challenge the general validity of the inferential step from conceivability to logical possibility, they pre-empt any fair debate by uncharitably tipping the playing field and by unfairly shifting the burden of proof to those who *explicitly* engage in a priori conceptual reasoning. By doing so, they fail to comply with the basic rules of the practice of a priori conceptual reasoning. For this is a practice in which contemporary empiricists also *implicitly* engage, when they argue that it is a priori conceivable, and therefore possible, that any particular inferential step from conceivability to possibility could be invalid. No philosopher, including any contemporary empiricist, would ever have engaged in such an inferential practice if the inference from liberal conceivability to logical possibility were not *generally valid in the APA logic at the very least*.

So short of rejecting contemporary philosophy of logic and showing that non-classical logics are simply impossible from the get-go, a contemporary empiricist cannot rationally fail to grant us the general validity of inferences from liberal conceivability to logical possibility in the APA logic. This in turn forces the contemporary empiricist critic to admit explicitly that there *are* some a priori conceptual arguments, and that the testing of alternative philosophical intuitions against the premises and conclusions of such a priori conceptual arguments for the purposes of evaluating their soundness is something over and above the *general validity* of such arguments. If they want to criticize the arguments, they will now have to engage explicitly in the practice of a priori conceptual reasoning with us and address the specific arguments on their own terms.

Assuming the general validity of these inferences in the APA logic, what we now want specifically to argue is that it is logically impossible for there to be *Emotional Zeroes* that also have a consciousness like ours. By the

notion of an "Emotional Zero" we mean a living organism *O* that meets the following two conditions:

(i) *O* is physically *exactly* like me right now, and also shares my causal history, so that it pairs causal power for causal power, body movement for body movement, neurobiological process for neurobiological process, living-cell-for-living-cell, molecule-for-molecule, atom-for-atom, etc., from the beginning of my life right up to the present moment,

and

(ii) *O* constitutively lacks all desire-based emotions.

Here is the argument. Suppose that for any current reader of this book, there exists some or another Emotional Zero corresponding to that reader. To make this more vivid, imagine a scenario from Don Siegel's classic 1956 science fiction film *Invasion of the Body Snatchers*: Your body has been snatched away by aliens and replaced by an emotionless but otherwise perfect physical duplicate organically grown in a pod. As one of the characters in the film, the psychiatrist Dr Kaufman, who has himself been replaced by a pod person, says:

> Your new bodies are growing in there. They're taking you over cell for cell, atom for atom. There is no pain. Suddenly, while you're asleep, they'll absorb your minds, your memories and you're reborn into an untroubled world . . . . Tomorrow you'll be one of us. There's no need for love. . . . Love. Desire. Ambition. Faith. Without them, life is so simple, believe me.[73]

In order to make the scenario more philosophically robust, imagine that the aliens also have found some way of replicating the causal history of your body from the beginning of your life right up to the present moment. So for the purposes of evaluating the relevant modal hypothesis, we imagine that the duplicate differs from you *only* in that it is devoid of desire-based emotions. In general, let us call such a creature "[Your name here] the Zero" To make the example not only more philosophically robust, but also more personally meaningful for one of us, we will call him "Bob the Zero," or for short, "Bob the Z" (pronounced, in the Anglo-Canadian way, as "Bob the Zed").

[73] Siegel, *Invasion of the Body Snatchers*.

It should be noted that the Emotional Zero hypothesis is *not* a zombie hypothesis. Zombies are supposed to be exact physical duplicates of us that lack consciousness in some logically possible world. And in fact, our essential embodiment metaphysics of the mind-body relation (see Chapters 6 to 8) entails that while zombies are logically possible, they are nevertheless non-logically or strongly metaphysically a priori *impossible*, which of course entails that they are also nomologically impossible. But that is irrelevant for the present purposes. For the present purposes of this argument, given the duplication of all his physical properties, Bob the Zero not only *can* be "conscious," just as a matter of straight logical possibility in the APA logic, but it is also logically possible that Bob the Z is *necessarily* "conscious," whether by materialist identity or materialist logical strong supervenience. What we will very shortly learn, however, is that it is logically impossible for Bob the Z to have a consciousness *like ours*. That is why we have scare-quoted the word 'conscious'. Even if Bob the Z *were* "conscious" in some liberally conceivable way, we would not have the slightest idea what his "consciousness" *is like*. Similarly, it is liberally conceivable and therefore logically possible in the APA logic that the desk in front of you, the book you are reading, or your laptop computer, is "conscious." If panpsychism were true, *everything* would be "conscious." But we do not have the slightest idea what a desk-ish, book-ish, or laptop computer-ish kind of "consciousness" would ever *be like*.

In any case, the crucial and most relevant question to ask about Bob the Zero is what he will do next. The obvious answer is—*nothing*. This is because Bob the Z cannot have any desire to do anything, hence cannot have any will to do anything, hence can never undertake any intentional body movements. This can be clearly seen in the following way.

By hypothesis, since Bob the Zero is a perfect physical duplicate of me up to the present moment, then in the near future, other things being equal, Bob the Z will continue to live, move, and have his being. Yet Bob the Z cannot cognize or act intentionally, precisely because he lacks the characteristic affects and feelings that necessarily accompany desire-based emotions, and thus he lacks any ability to focus his attention or focus his goals by means of affective framing. Any movement that Bob the Z makes will be strictly underdetermined by natural stimuli and natural states of affairs, and yet his continuing life and body movements cannot be explained by appealing to his intentional agency. So Bob the Z's continuing

life and body movements are all bodily events that merely endogenously *happen* to him. Bob the Z is nothing but a *puppet*. In short, Bob the Z is not only an Emotional Zero but also an intentional agency zero.

But not only that: this conclusion directly entails that Bob the Zero cannot be a creature with a consciousness$_{lo}$. *Reading* is an intentional act, and therefore Bob the Z cannot read even if he can mechanically apply suitable information-processing algorithms to the text. *Yet all rational creatures with a consciousness$_{lo}$ are capable of reading, and we have all actually and intentionally read the book up to this point, or at the very least we have all actually and intentionally read this sentence*, thereby actually exemplifying both attentive focusing and goal focusing in our conscious, intentional activity. It is liberally inconceivable that Bob the Z could ever do this. Hence it is logically impossible for Bob the Z or any other Emotional Zero to have a consciousness$_{lo}$, and it directly follows by contraposition that logically necessarily all creatures with consciousness$_{lo}$ are not Emotional Zeroes and therefore have desire-based emotions.

This completes our project of extending the Essentially Embodied Agency theory of action to the philosophy of the emotions by way of the Emotive Causation Thesis. We have argued in this chapter that the subjective experience of effortless trying and its active guidance is grounded in pre-reflectively conscious desire-based emotions, rather than in beliefs, judgments, thoughts, or self-conscious or self-reflective, deliberative intentions. We have also argued that the action-initiating and action-guiding causal powers of self-conscious or self-reflective deliberative intentions are derived from the action-initiating and action-guiding causal powers of pre-reflectively conscious desire-based emotions. But what ultimately makes this line of reasoning possible is the application of the background metaphysics of essential embodiment to intentional agency, to which we now explicitly turn in the last three chapters.

6

The Metaphysics of Agency I: The Problem of Mental Causation

> If it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . , if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.
>
> Jerry Fodor[1]

6.0 Introduction

The mind–body problem, as we understand it, is the problem of explaining the *existence* and *specific character* of conscious, intentional minds_{lo} in a physical world. But since the physical world is also a world of causally efficacious events in spacetime, any attempt to solve the mind–body problem leads directly to another, even deeper problem: How can we explain the *causal relevance* and *causal efficacy* of conscious, intentional minds_{lo} in a physical world? That is the problem of mental causation, and it obviously relates directly to the problem of action. Comprehensively and ultimately, then, what we really want to know is this: How can we explain the existence, specific character, causal relevance, and causal efficacy of conscious, intentional minds_{lo} in a physical world, insofar as these minds are engaged in *intentional actions*?

According to our Essential Embodiment Thesis, conscious, intentional minds_{lo} are necessarily and completely neurobiologically embodied. In turn, according to our Essentially Embodied Agency theory of action, intentional

[1] Fodor, "Making Mind Matter More," 156.

body movements are caused by our synchronous effortless trying and its active guidance of those movements. And according to our Emotive Causation Thesis, effortless trying and its active guidance are primarily activities of our pre-reflectively conscious desire-based emotions, and only derivatively self-conscious or self-reflective, deliberative intellectual activities.

These three theses, we believe, jointly provide the basic ingredients for an adequate unified solution to the mind–body problem and the problem of action. But in order to show that our solution is fully adequate, we must also face up to the problem of mental causation. We have claimed that the fundamental mental properties of conscious, intentional minds_{lo} are irreducible to fundamental physical properties of the natural world. But if conscious, intentional minds_{lo} are non-physical, then how can they cause physical events? Or more bluntly put: If conscious, intentional minds_{lo} cannot be plausibly shown to have causal *efficacy* in a physical world, then it seems that all our talk of essential embodiment, effortless trying, active guidance, and pre-reflectively conscious desire-based emotions will have been in vain. Because something's causal relevance is consistent with its *really doing nothing*, causal relevance alone will not be good enough. We will not be keeping faith with our basic intuitions about intentional action unless it is true that conscious, intentional minds_{lo} can also *really do something* in a physical world.

This chapter is all about mental causation. In Section 6.1, we work out some important preliminary points about the concept of causation. In Section 6.2, we carefully formulate the problem of mental causation as a philosophical paradox that we call *the Amazingly Hard Problem*. In Section 6.3, we show how each of the basic premises in the Amazingly Hard Problem is independently rationally well-supported. And then finally in Section 6.4 we spell out Jaegwon Kim's two well-known *Causal Exclusion Problems* for the dualist and non-reductive materialist solutions to the mental causation problem, and show why recent attempts by non-reductive materialists to solve the Causal Exclusion Problems also fall short. This will set the stage for our new solution to both the Amazingly Hard Problem and the Causal Exclusions Problems alike in Chapter 7.

6.1 Some Preliminaries about Causation

It seems that there are three and only three sorts of efficacious causal relations a conscious, intentional $mind_{lo}$ can enter into:

(i) *mental-to-mental* causation (e.g., making inferences),
(ii) *physical-to-mental* causation (e.g., visual perception),

and

(iii) *mental-to-physical* causation (e.g., basic intentional actions such as impulsively raising one's arm while freestyle hip-hop dancing, or self-consciously or self-reflectively and deliberatively raising one's arm for a reason).[2]

If the causal efficacy of conscious, intentional $minds_{lo}$ were to exist, then it would be *centrally or pre-eminently* expressed in the mental-to-physical causation of intentional body movements. This in turn is for two reasons.

First, since external causal impacts on minded animals naturally and normally lead to responsive intentional action of some sort—e.g., turning one's head slightly when foveating a visual stimulus, or when moving towards something seen—then it seems that normal cases of physical-to-mental causation presuppose the possibility of mental-to-physical causation. Indeed, according to the *enactive* theory of perception recently defended by Alva Noë, the possibility of intentional body movement at least *partially constitutes* the content of sense perception, or as he puts it: "all perception is intrinsically active."[3] Noë also holds the significantly stronger thesis that perceptual content is *wholly* constituted by intentional action,[4] but from our standpoint that seems *too* strong. As disjunctivist direct perceptual realists

[2] What about cognitive priming effects, post-hypnotic suggestion, automatism, psychosomatic illness, placebo effects, and so-on? Given the Deep Consciousness Thesis, it follows that these effects are mentally caused in a pre-reflectively conscious, but not self-consciously or self-reflectively conscious, way. Nevertheless, just because they are pre-reflectively conscious, of course it does not automatically follow that they are *freely* or *intentionally* caused. Mental causation can also occur in cases of unfreedom of the will and unintentional action. So mental causation is a necessary but not sufficient condition of intentional causation and intentional agency.

[3] See Noë, *Action in Perception*, p. 3. [4] Ibid., chs. 1 and 3.

who are also perceptual enactivists, we think that perceptual content is partially constituted by the possibility of intentional action and *also* partially constituted by the real objects of perception and their real properties.[5] So for our purposes here, we need only the weaker enactive perception thesis in order to establish the thesis that normal cases of physical-to-mental causation presuppose mental-to-physical causation.

Second, since on our view all conscious, intentional events, states, or processes in minded animals are essentially embodied, then mental-to-mental causation can be defined as a special case of mental-to-physical causation in which either a covert neurobiological body movement or an overt intentional body movement, caused by effortless trying and its active guidance, *also* produces a further conscious, intentional event, state, or process as the intended effect. In other words, I make a valid inference by intentionally moving my body (remembering now that the range of my possible intentional body movements can also include the limiting case of holding my body in a single orientation and position—see Section 4.1) from the conscious, intentional situation in which I rationally assert the premises to the conscious, intentional situation in which I assert the conclusion, in self-conscious conformity with a priori normative laws of logical consequence.[6] Indeed and more generally, as Kim has correctly argued, except for substance dualists, dualist parallelists, and causal anti-realists, who reject the Principle of the Causal Closure of the Physical, or CCP, and think that efficacious causation can somehow bypass the physical world altogether, mental-to-mental causation is always explained in terms of mental-to-physical causation.[7]

In this way, both normal physical-to-mental causation and mental-to-mental causation alike presuppose mental-to-physical causation. Therefore mental-to-physical causation is the basic or pre-eminent fact of mental causation, and unless otherwise specified, for the rest of the book that is what we will mean by the label 'mental causation'.

But before we discuss mental causation in particular, we need to say something about causation in general.[8] Our overall approach to causation is *realist, non-reductive*, and *theoretically inclusive*. It has five parts.

[5] See, e.g., Campbell, *Reference and Consciousness*, chs. 6 and 12.

[6] See Hanna, *Rationality and Logic*, esp. chs. 4–7.

[7] See Kim, *Philosophy of Mind*, ch. 6.

[8] See Sosa and Tooley (eds.), *Causation*; and Schaffer, Lewis, Hall, Collins, and Paul, "Special Issue: Causation."

First, we hold that the primary fact of causation is a real metaphysical relation between *singular* events in spacetime, such that a singular event e_1 causes a singular event e_2 which is not earlier than e_1, under an intrinsic nomological constraint or law.

Now it is important for our later line of argument to recognize that singular events may or may not be *simple*. So just to be as clear as possible, here are some intuitively plausible theses and definitions about events.

Extension: Every event has a four-dimensional spacetime volume (including the familiar three Euclidean dimensions in globally orientable space, and one asymmetric temporal dimension) or an *extension*.

Cohabitation and divorce: Two events *cohabit* the same spatial or temporal extension if and only if they either partially or completely overlap the same extension; otherwise they are non-cohabiting or *divorced*.

Simultaneity: Two events are simultaneous if and only if they are cohabiting and completely overlap the same temporal extension.

Simplicity and complexity: An event is *simple* if and only if it contains no other cohabiting events as proper parts; otherwise, if an event contains some other cohabiting events as proper parts, then it is *complex*.

Compoundness: An event is *compound* if and only if it contains at least one divorced event as a proper part.

Singularity: An event is *singular* if and only if it is either simple, or else complex such that all its cohabiting events are simultaneous.

In this way, the several simple events that make up a *complex singular event* must all be simultaneous and cohabiting. For example, the simple event of my wearing a belt and the simple event of my wearing suspenders, if they happen at the same time, make up the complex singular event of my wearing a belt and suspenders. By contrast, all simple or complex events that are spread out over asymmetric successive time, or that occur separately in space at the same time, are compound events. For example, the event of my putting on my belt followed soon after by my indecisively taking off my belt again, is a compound event by virtue of asymmetric temporal succession. Correspondingly, the two-part event consisting of my wearing my belt, together with the simultaneous event of my next door neighbor John wearing his suspenders, is a compound event by virtue of spatial separation.

For us, causation is at bottom an objective fact about nature, and causal laws occur inherently in the singular events they constrain and govern. Or in other words, nature is filled with causally-related events and *processes*. Here, then, is one last definition to add the above list:

> **Process:** A set of events is a *process* if and only if it constitutes a temporally extended compound event that is inherently constrained and governed by at least one causal law.

For our present purposes, we will not try to argue against either causal anti-realism or the extrinsic approach to causal laws. Our present aim is just to spell out a general framework for thinking about causation, and then test the truth of that framework by the indirect method of demonstrating its ability to contribute to the best overall explanation of mental causation.

But the question of the nature of causal laws raises a seminal issue. We are officially leaving it open whether the laws that intrinsically constrain and govern all natural causal relationships are

(i) deterministic,
(ii) probabilistic or statistical,

or

(iii) context-sensitive ("hedged" or *ceteris paribus*).[9]

Corresponding to these three types of laws, we also hold

(1) that there are some *completely deterministic* events or processes, i.e., events or processes that are logically or causally entailed by rough-grained general deterministic laws plus antecedent facts,
(2) that there are some *completely indeterministic* events or processes, i.e., events or processes that are logically or causally entailed by roughgrained general statistical laws plus antecedent facts,

and

(3) that there are some *natural causal singularities*, which are events or processes that are *neither* logically or causally entailed by rough-grained general deterministic laws plus antecedent facts *nor* logically

[9] See, e.g., Fodor, "Making Mind Matter More"; and Rupert, "*Ceteris Paribus* Laws, Component Forces, and the Nature of Special Science Properties."

> or causally entailed by roughgrained general statistical laws plus antecedent facts, and which exert their finegrained and hyper-finegrained causal powers consistently with whatever roughgrained general deterministic or statistical laws there are.

Examples of completely deterministic events or processes include the acceleration of falling bodies due to gravitational attraction, and water boiling at 100 degrees centigrade/212 degrees Fahrenheit at standard pressure. Examples of completely indeterministic events or processes include quantum-mechanical phenomena, and (arguably) batting streaks in baseball. And examples of natural causal singularities include the Big Bang and black holes, the roiling movements of boiling water (as opposed to water boiling at 100 degrees centigrade/212 degrees Fahrenheit at standard pressure, which is deterministic), weather systems, traffic systems, ecosystems, planets like the Earth, solar systems, stars, star systems, and the biological processes and endogenously produced overt movements of living organisms, including the intentional body movements of conscious, intentional animals and real persons. Indeed, General Relativity predicts the existence of natural causal singularities such as the Big Bang and black holes.[10] But even more importantly from our point of view, the biological processes and overt body movements of individual living organisms constitute a class of naturally creative *little bangs* with the same essential properties as the dramatically larger natural causal singularities.[11]

What are the essential properties of natural causal singularities, whether large or little? Natural causal singularities are nomologically unique, actual-world dependent, unprecedented, unrepeatable, situated, forward flowing, non-random processes with thermodynamic self-organization, existing in a natural world that also has some roughgrained general deterministic laws and some roughgrained general probabilistic or statistical laws. This sets natural singularities sharply apart from both completely deterministic events *and* completely indeterministic events. On the one hand, completely deterministic events or processes are such that from the set of settled facts about the past together with the laws of nature, both the existence and the specific character of all future events are logically or causally neccessitated

[10] See, e.g., Hawking, *A Brief History of Time*.

[11] See, e.g., Nicolis and Prigogine, *Self-Organization in Nonequilibrium Systems*; and Prigogine, *Being and Becoming: Time and Complexity in the Physical Sciences*.

and they can in principle be predicted a priori. So completely deterministic events obey what we will call *the Closed Future Rule*:

> necessarily if any two events e_1 and e_2 have exactly the same past, then e_1 and e_2 will also have exactly the same future.

On the other hand, completely indeterministic events are such that from the set of settled facts about the past together with the laws of nature, neither the existence nor the specific character of those events is necessitated and they cannot even in principle be predicted a priori. So completely indeterministic events obey *the Open Future Rule*:

> necessarily even if two events e_1 and e_2 have exactly the same past, then possibly and with some definite degree of probability, e_1 and e_2 each will have a different future.

But by sharp contrast to competely deterministic and completely indeterministic events alike, natural causal singularities are such that they have neither a closed future nor an open future. Instead, *they naturally create their own future by what they actually do in the present,* consistently with all the roughgrained general deterministic or probabilistic laws there actually are. The naturally creative character of natural causal singularities follows from their being inherently context-sensitive and actual-world dependent (a.k.a. "essentially indexical"[12]), naturally purposive or teleological, and holistic. In response to actual initial conditions, by internally generating novel dynamic patterns that are globally shared by all the parts of that dynamic system, a natural causal singularity produces an inherent "demand" or "need" to ramify or sustain those very patterns in order to maintain itself, and thus anticipates the later course of its own natural development.

In this sense, a natural causal singularity is a "law unto itself" or "nomologically one-off," precisely because its intrinsic constraining and governing law is a special hedged or *ceteris paribus* law whose actual-world-dependent contextual conditions allow for exactly and necessarily one instance. A natural causal singularity does not *violate* any roughgrained general deterministic or probabilistic natural causal laws, and thus it is perfectly *consistent* with all the roughgrained general deterministic or probabilistic natural

[12] See, e.g., Perry, "The Problem of the Essential Indexical."

causal laws that are causally relevant to it. But at the same time, all the roughgrained natural causal laws do not *suffice* to fix its finegrained or hyperfinegrained nomological essence. In this sense, a natural causal singularity is not only naturally *law-abiding*, but by virtue of its natural creativity it is also naturally *self-legislating*.

This absolutely crucial point needs more elaboration. As a general conceptual, logical, and metaphysical matter, it is entirely possible for a dynamic system X to *comply with* all the relevant roughgrained general deterministic or probabilistic laws such that they *partially fix* its dynamic trajectory, even though all those relevant laws together with antecedent facts do not *fully fix* what X actually does. For example, if I accidentally fall off a bridge, I cannot fall towards the earth slower or faster than 10 meters per second^2. That is the law of falling bodies. Call this law L_1. In this context, L_1 determines the velocity of my body. But L_1 does *not* itself precisely determine my complex neurobiological state as I fall, nor indeed does it determine my complex feelings about the whole sad state of affairs (presumably, absolute terror together with my whole life suddenly passing before my eyes). Nor does the fact of my actually instantiating the law of falling bodies itself necessitate that there exists some *further* law L_2 that precisely determines my complex neurobiological state or my complex feelings about the whole sad state of affairs.

More generally, the fact that I actually instantiate *some* roughgrained general deterministic or probabilistic laws, and indeed also actually instantiate *all* the roughgrained deterministic or probabilistic laws that are *causally relevant* to me, does *not* itself necessitate either that every event, including all those events that make up my life, is determined (Universal Natural Determinism); or that every event, including all those events that make up my life, is indeterministic (Universal Natural Indeterminism); or that every event including, all those events that make up my life, is either completely deterministic or completely indeterministic (Natural Mechanism). And this is because, despite the compliance of any given event with all the roughgrained general deterministic or probabilistic laws that are causally relevant to it, it simply does not follow that every single intrinsic structural property of that given event is precisely fixed by either a deterministic or a probabilistic law. So there could still be *further* intrinsic structural properties of at least some events that are not precisely fixed by those laws, and that

also naturally create and self-legislate the finegrained or hyper-finegrained dynamic trajectory of those events. Events that have these further naturally creative and self-legislating intrinsic structural properties are the natural causal singularities.

If this account is correct, then natural causal singularities are neither logically nor causally necessitated by the past, and their effects are not predictable from a standpoint external to the event itself. So they are not competely deterministic. But on the other hand, as we pointed out in Section 3.3, their effects *could* be predicted if one *were able* to adopt a standpoint that is literally internal to the self-organizing thermodynamics of the event itself. So they are not indeterministic either. Adopting a standpoint literally internal to the self-organizing thermodynamics of the event itself would not, of course, be practically feasible in the case of the Big Bang, black holes, the roiling movements of boiling water, and many individual living organisms—e.g., plants. But adopting such a standpoint would indeed be practically feasible in the case of minded animals, for that is precisely what an essentially embodied effectively desiring intentional consciousness$_{lo}$ *is*. As a first-person standpoint literally internal to the self-organizing thermodynamics of a motile, situated, forward flowing suitably neurobiologically complex living organism, our capacity for conscious intentionality is the capacity of an animal to predict its own intentional body movements by means of *pre-reflectively conscious effortless trying and its active guidance*.

Second, we also are officially leaving open the possibility, within our event-causation framework, of both property-causation and substance-causation. On our account—which we will spell out in some detail in Sections 7.2 and 7.3—all physical substances are *complex dynamic systems*, and thus physical substances are intrinsically structured singular or compound events that cause efficaciously via their constituent simple or complex singular events. In turn, the intrinsic structural *properties* of those dynamic systems are also causally efficacious via the causal powers of their constituent simple or complex singular events. Thus our account of causation is highly inclusive. In fact, the only logically possible sorts of causation that are metaphysically ruled out by our account are

(1) the sort of non-physical causation involved in dualist interaction or "noumenal" causation, as e.g., in Cartesian or Kantian versions of Agent Causation,[13]

and

(2) the sort of non-standard (i.e., non-physical) systematic causal overdetermination postulated by non-reductive materialism—according to which the very same event has both an individually sufficient but not individually necessary *physical* cause and also an individually sufficient but not individually necessary *non-physical* cause that is upwardly determined by nomological strong supervenience.

This is because non-physical causation clearly violates any reasonable interpretation of the principle of the Causal Closure of the Physical or CCP; because non-standard systematic causal overdetermination also violates Kim's plausible Explanatory Exclusion Principle or EEP; and also because Kim has adequately shown that any version of mental causation according to Non-Reductive Materialism, since it is based on strong supervenience, leads to Epiphenomenalism. (See Sections 6.2, 6.3, and 6.4 below for more discussion of CCP, EEP, strong supervenience, and their implications for Non-Reductive Materialism). Hence the fact that our account metaphysically rules out both of these types of causation while remaining otherwise highly inclusive is another important point in its favor.

Third, it should also be especially noted that our account allows for causation to occur not only over a temporal sequence of successive moments but also *simultaneously over continuous time*, or *synchronously*. More precisely, we hold that two temporally extended simple singular events e_1 and e_2 can be simultaneous and *also* such that e_1 causes e_2. These two events e_1 and e_2 thereby make up a complex singular event e_3. The everyday natural world appears to supply many examples of synchronous causation:

1. A lead ball is resting on a cushion; the presence of the ball causes an indentation in the cushion.

[13] See, e.g., Chisholm, "Human Freedom and the Self;" and Watkins, *Kant and the Metaphysics of Causality*, esp. part III.

2. A locomotive is pulling a truck; the movement of the engine is responsible for the movement of the truck.
3. An iron bar is glowing because of its high temperature.
4. The lowering of one end of a seesaw causes the other to go up.
5. Moving one end of a pencil causes the other end to move.[14]

To be sure, those who defend a sequential view of causation will try to argue that all apparent examples of synchronous causation can be explained away using sequential causal relations. But even if that were so, the sequential view still faces the general problem of explaining how one event can cause another if the first event expires in time *before* the second event begins. How can causal power be transmitted over an absolute temporal gap? This seems to be every bit as metaphysically mysterious as the transmission of causal power over an absolute spatial gap, which of course is the classical problem of *action-at-a-distance*.

Our synchronous causation view not only allows for simultaneous and continuous causation, but also postulates this synchronous causation as an underlying ground for all real sequential causation, which always occurs in compound events consisting of successive series of partially overlapping or cohabiting spatiotemporally extended simple or complex singular events with simultaneous causation in the overlaps, like the links in a chain. We call this *relatively* sequential causation. We think that the notion of simultaneous and continuous causation, taken together with the notion of relatively sequential causation, provides a much better overall characterization of causation in everyday life and natural science alike than the sequential view does. Indeed, synchronous causation seems to be a tacit but under-acknowledged feature of classical physics itself. For the Newtonian principle $F = ma$ can be read as saying precisely that the force of a material body *at a given time* is directly proportional to its mass and its acceleration *at that time*, and that this relationship holds fixed over *successive time*.

Fourth, provisionally granting us the three other elements of our approach to causation, what then is the defining essence of the causal relation? The usual suspects—i.e., the standard candidates—are these:

(i) e_1 is a necessary condition of e_2,
(ii) e_1 is a sufficient condition of e_2,

[14] Huemer and Kovitz, "Causation as Simultaneous and Continuous," 557.

(iii) e_1 is a necessary and sufficient condition of e_2,
(iv) e_1 is an insufficient but non-redundant part of an unnecessary but sufficient cause (or otherwise put: e_1 is an INUS cause[15]) of e_2,
(v) e_1 is a counterfactual condition of e_2,
(vi) e_1 is neither a necessary nor sufficient nor counterfactual condition of e_2 but instead tokens of e_1-type events are just regularly followed by tokens of e_2-type events according to a lawlike generalization covering successions of events of those sorts.

Options (i), (iii), (iv), (v), and (vi) however all seem questionable as candidates for capturing the essence of the causal relation, for there are fairly robustly intuitive cases that falsify each. Of course, even accepting the counterexamples, it will remain entirely possible that each of the standard candidates still accurately captures *some* aspects of *some* or even of *a great many* causal relations. This seems obviously true, e.g., of (v), causation as counterfactual influence. In any case, we will briefly canvass the counterexamples because they are philosophically instructive.

Mere *background or standing conditions* for causation supply examples in which e_1 is a necessary condition of e_2 but e_1 does not cause e_2, thereby falsifying (i). For example, although the presence of air is a necessary background or standing condition for the production of sounds, it does not itself usually cause sounds.

Standard causal overdetermination cases supply examples in which e_1 causes e_2 but e_1 is not a necessary condition of e_2, thereby falsifying option (iii). For example, two gun-toting assassins simultaneously shoot someone and kill him although either one of the assassins' bullets alone would have been sufficient to bring about the victim's death. A less lurid example is how a belt and suspenders can each simultaneously hold up the same pair of trousers.

Fallacy of causal composition cases, in which an intrinsic proper part of a whole singular event is illegitimately substituted for that whole singular event in a causal attribution or causal inference, supply examples in which e_1 is an INUS cause of e_2, but e_1 is clearly not *the* cause of e_2, thereby falsifying option (iv). For example, while someone might well claim or say that slapshooting the puck caused the goalie's nose to break, strictly speaking

[15] See Mackie, "Causes and Conditions."

only the *whole* singular event consisting of the slapshot, the flying puck and its actual trajectory, local gravitational facts, facts about the composition of the goalie's human body, facts about the thinness of the goalie's plastic mask, etc., caused the goalie's nose to break.

Trumping preemption cases provide examples in which e_1 is a counterfactual condition of e_2 but e_1 does not cause e_2, thereby falsifying (v). For example, to recur to the sort of case used against causal theories of action in Chapter 3, we can appeal to deviant causal chains. If someone is an Olympic sprinter, his belief that the race is about to begin together with his strong desire to start running right at the sound of the gun might make him so nervous that he jerks forward just as the gun simultaneously goes off. Although the sprinter's movement would not have happened if the belief-desire pair had not been in place, it was his *jumpy nerves* that caused his movement, thereby causally trumping and preempting the belief-desire pair.

Cases of *constant lawful mere coincidence* provide examples in which tokens of e_1-type events are regularly followed by tokens of e_2-type events according to a lawlike generalization covering successions of events of those sorts, but e_1 does not cause e_2, thereby falsifying option (vi). For example, two symptoms of a disease can regularly and systematically follow each other without the first being the cause of the second.

And finally *natural causal singularities* provide examples of cases in which e_1 causes e_2, but tokens of e_1-type events are not regularly followed by tokens of e_2-type events according to a lawlike generalization covering successions of events of just those sorts, thereby falsifying (vi) again. For example, the Big Bang, black holes, the roiling movements of boiling water, and the covert biological processes and overt body movements of living organisms do not fall *precisely* or *specifically* under any roughgrained deterministic or indeterministic lawlike generalizations over regular sequences of events. Natural causal singularities are perfectly consistent with all such laws, but their finegrained or hyper-finegrained causal powers and operations are not fully fixed by those laws, precisely because some of their intrinsic properties are naturally *created* and thereby *self*-legislated by those very events.

Nevertheless, nomological sufficiency-relations between singular events do appear to hold in *all and only* possible cases of natural causation in our actual world. How causation might operate in possible worlds far from ours, or how divine or angelic causation might work, are things we will not consider in this context. So, otherwise put, it seems to us that in all and

only possible cases of *natural causation in our actual world*, one singular event *lawfully necessitates* another event that is not earlier than the first event, in the following sense: Given the occurrence of the first event, it is *no mere coincidence* and *no mere accident* that the second event also occurs. The first event thereby *predictably produces* the second event. Therefore we adopt a version of option (ii), according to which a cause is a nomologically sufficient condition of its effect. To emphasize this, we use the term *nomologically sufficient cause*.

But we also include in our overall analysis a version of option (v), the counterfactual condition, as a necessary but not sufficient condition of causation. Just to round things out, then, here is a summary of our analysis of causation. By *causation* we mean a relation between two singular events, e_1 and e_2, such that

> (i) e_2 is not earlier than e_1, (ii) e_1 nomologically sufficiently guarantees the existence and specific character of e_2, and (iii) e_2 would not have existed if e_1 had not existed.

Then e_1 is *a nomologically sufficient cause* and e_2 is *its effect*.

Now while the classical Humean or regularity/covering law theory of causation has some natural causal singularity counterexamples that isolate and undermine the condition of *regularity*, it still effectively brings out the plausible thesis that each instance of the singular causal event relation falls under some causal law of nature that intrinsically governs the precise connection between events of those types.[16] This remains true *whatever* one's view on the nature and modal status of causal laws of nature happens to be. So as we mentioned above, it seems to us that such laws could in principle be either deterministic, probabilistic, or context-sensitive (hedged or *ceteris paribus*). Such laws could thus in principle describe universal deterministic LaPlacean regularities in nature, stochastic regularities, regularities under contextual constraints, or unique dynamic trajectories in the limiting case of "one-off" laws with essentially indexical context-sensitivity—namely, the dynamic patterns of natural causal singularities. The only requirement is that the laws intrinsically connect types of events and carve out a class of relevant possible worlds containing all and only the relevant singular

[16] This is what Davidson calls "the Principle of the Nomological Character of Causality." See Davidson, "Mental Events," 108.

causal event-connection tokens. In this way, we are committed only to the broadest possible interpretation of the thesis that an event-effect is nomologically necessitated by its event-cause, or alternatively, that an event-cause is nomologically sufficient for its event-effect. According to our view then, for all and only possible natural events in our actual world, e_1 causes e_2 if and only if

> (i) e_2 is not earlier than e_1 (ii) e_1 nomologically sufficiently guarantees (in the maximally broad sense just described) the existence and specific character of e_2, and (iii) e_2 would not have existed if e_1 had not existed.

Fifth and finally, we also hold that there is a crucial conceptual and metaphysical distinction to be marked between (a) *causal efficacy* and (b) *causal relevance*.[17] In order to capture this distinction, we will say that a singular event e_1 is causally efficacious if and only if

> either (i) e_1 is itself a nomologically sufficient simple singular event cause of some physical event e_2 or (ii) e_1 is a necessary proper part of e_3, which itself is a nomologically sufficient complex singular event cause of e_2.

We can also extend this notion of causal efficacy to properties and physical substances. Then a property P is causally efficacious if and only if P is instantiated as an intrinsic property by events that are causally efficacious, and a physical substance S is causally efficacious if and only if S is constituted by causally efficacious events and properties.

By sharp contrast, an event e_1 is causally relevant if and only if

> either (i*) e_1 is a necessary condition for some event e_3's being a nomologically sufficient cause of some physical event e_2 or (ii*) some correct description of e_1 enters directly into an informative characterization of e_3's being a nomologically sufficient cause of e_2.

And we can also extend this notion to properties and physical substances. Then a property P is causally relevant if and only if some of P's instantiations are causally relevant; and a physical substance S is causally relevant if and only if S is constituted by causally relevant events and properties.

In other words, the causal relevance of an event e, property P, or physical substance S means only that e or P or S has a definite logical or

[17] See, e.g., Jackson, "Mental Causation," 397.

informational bearing on an efficacious causal process, which is perfectly consistent with *e*'*s* or *P*'s or *S*'s *really doing nothing at all*. By contrast, the causal efficacy of a simple or complex singular event *e* or property *P* or physical substance *S* means that *e* necessarily belongs to an efficacious causal process itself either as a nomologically sufficient condition on its own or as a necessary proper part of a nomologically sufficient condition. In that case, *e really does something*, *P* is the property by virtue of which *e* has precisely these efficacious causal powers, and *S* is a substance made up of several causally efficacious *e*s and *P*s.

6.2 The Amazingly Hard Problem

So much for the preliminaries about the concept of causation in general—now back to mental causation in particular.

If we were to restrict our attention narrowly to phenomenal consciousness$_{lo}$, and also assume that all the other kinds of consciousness$_{lo}$ and intentionality$_{lo}$ can be materialistically explained, then the mind–body problem becomes what Chalmers calls "the hard problem."[18] But if the version of the mind–body problem that narrowly restricts it to phenomenal consciousness$_{lo}$ is *the Hard Problem*, then surely the problem of mental causation must be *the Amazingly Hard Problem*,[19] for at least three reasons. First, the problem of mental causation, as we are understanding it, does *not* construe consciousness$_{lo}$ as merely *phenomenal* consciousness$_{lo}$, and explicitly includes both conscious intentionality$_{lo}$ and intentional agency. Second, as a direct consequence of the first point, the problem of mental causation expresses, as it were, the "compleat" mind–body problem—i.e., a problem about how *all* aspects of minds$_{lo}$ can be adequately accounted for in a physical world—and not only one special part of it. Third and perhaps most importantly, the problem of mental causation, as we are understanding it, takes the form of a genuine philosophical paradox, and not just that of a philosophical puzzle.

[18] See Chalmers, *The Conscious Mind*.

[19] This is to be distinguished from what Block calls "the harder problem of consciousness," which is an epistemic variant on Chalmers's hard problem. See Block, "The Harder Problem of Consciousness."

To see all this, let us look closely now at the most important details of the Amazingly Hard Problem of mental causation, explicitly presented as a six-step argument:

(1) *The Causal Efficacy of the Mental* (CEM): Conscious, intentional $minds_{lo}$—in particular, conscious, intentional $minds_{lo}$ insofar as they are engaged in intentional actions—can cause physical events.
(2) *The Causal Closure of the Physical* (CCP): Only physical events can cause physical events.
(3) *The Causal Physicality of the Mental* (CPM): In order to cause physical events, conscious, intentional $minds_{lo}$ must be physical. [From (2)]
(4) *The Physical Irreducibility of the Mental* (PIM): Because mental properties are irreducible to physical properties, conscious, intentional $minds_{lo}$ are non-physical.
(5) *The Causal Failure of the Mental* (CFM): So conscious, intentional $minds_{lo}$ cannot cause physical events. [From (3) and (4)]
(6) Therefore conscious, intentional $minds_{lo}$ both can and cannot cause physical events. [From (1) and (5)] **Contradiction!**

Since CPM obviously follows validly from CCP, and since CFM obviously follows validly from CPM and PIM, the Amazingly Hard Problem rests ultimately on CEM, CCP, and PIM.

The Amazingly Hard Problem of mental causation is a genuine philosophical *paradox* and not just a philosophical *puzzle*, precisely because it is a logically valid argument leading to a contradictory conclusion, and each of its basic premises (CEM, CCP, and PIM) is independently strongly supported by good reasons, thereby making the conjunction of all its premises true, and the inference to the contradiction sound. But sound arguments cannot have even contingently false conclusions, much less contradictory ones! So the Amazingly Hard Problem is a genuine paradox. Let us now look at the good reasons behind the basic premises.

6.3 Good Reasons for Efficacy, Closure, Physicality, and Irreducibility

CEM is strongly supported by commonsense, neurophenomenological introspection, and deeply important practical considerations concerning

agency, autonomy, responsibility, and so on. Fodor captures that line of argument beautifully in the epigraph of this chapter. Indeed, as we noted in the Introduction, it is the starting point of this book that our basic intuitions about intentional action carry decisively greater rational force than any thesis in metaphysics that contradicts or undermines them. In particular, the thesis of *Epiphenomenalism*, which says that all mental properties and facts are caused by physical properties and facts—or at least are fully determined by causally efficacious physical properties and facts—but have no efficacious causal powers of their own, so that my conscious choices and doings, no matter how free they may seem, are never actually *up to me*, directly contradicts our basic intuitions about intentional agency. So any mind–body theory which entails the denial of Epiphenomenalism carries decisively greater rational force than any mind–body theory that is either consistent with Epiphenomenalism or entails it.

Indeed, if our basic intuitions about intentional agency were wrong, and if Epiphenomenalism were true, then not only would it be "the end of the world," but also it would seem to be self-stultifyingly impossible for us even to *believe* that these self-conceptions were wrong. This is because the psychological attitude of belief is a freely chosen and rationally defensible self-commitment to the truth of a proposition or to the conclusion of a valid argument. If I came to say "I believe *P*" only because I were *merely caused, compelled, or forced* to say this by something alien to myself inside or outside my body—that is, if I came to say "I believe *P*" only because of something that *merely happened to me* as opposed to something *I intentionally did*—that would entirely undermine its being the genuine expression of a belief. As a rational intentional animal and a sincere speaker, I choose to assert *only* those propositions that seem true to me, and *only because* they seem true to me. Hence *any* sort of logical reasoning presupposes our own intentional agency as conceived by us according to our basic intuitions about intentional agency,[20] and for this reason we cannot give up these basic self-conceptions about intentional agency without committing rational suicide.

Next, at a first pass, by CCP ("Only physical events can cause physical events") we mean:

[20] See Hanna, *Rationality and Logic*, ch. 7.

□ (∀x) (∃y) [(x is a simple or complex singular event & x is physical & y causes x) ⊃ (y is a simple or complex singular event & y is physical)].

In view of our working analysis of causation as nomological sufficiency between simple or complex singular events in simultaneous or successive spacetime, this slightly formalized version of CCP says that necessarily, for any singular event *X* that is physical and has some nomologically sufficient singular event-cause *Y*, *Y* is also a singular physical event. Assuming that the quantifiers range neutrally over spacetime events—whether quantum events, atomic events, molecular events, macro-physical events, chemical events, or biological events—this formulation allows that *any kind of physical singular event* can count as a cause of another singular event. It also leaves open the possibility, raised by quantum indeterminacy, that some physical events do not have singular event-causes. But in any case, as initially so formulated and understood, CCP is strongly supported by good reasons because at bottom it says that *no wholly non-physical items*—e.g., transcendent gods, angels or other ectoplasmic finite spiritual agencies, disembodied Cartesian souls, non-spatiotemporal agent-causes, platonic forms, etc.—*can ever be causes of physical events*. And since what we have primarily in mind is *efficacious* causation, that seems correct if anything does.

Nevertheless, there is a further and subtler issue about how we should interpret CCP. Kim's interpretation of CCP—which is not unique to Kim, and seems to be widely shared by materialists and dualists alike—assumes that fundamental physical properties *necessarily exclude* any inherent or intrinsic connections with fundamental mental properties.[21] In our terminology, anything whose fundamental physical properties necessarily exclude any inherent or intrinsic connections with fundamental mental properties is *fundamentally physical*, even if it happens to possess some accidental mental properties (e.g., *X* might be a fundamentally physical event that is the supervenience base of some mental properties) or to stand in some other sort of extrinsic relation to mental properties (e.g., *X* might be a fundamentally physical event that has the property of being thought about by me). So Kim must also assume that *necessarily whatever possesses a fundamental physical property is fundamentally physical*.

[21] See Kim, *Mind in a Physical World*; Kim, *Philosophy of Mind*; Kim *Physicalism, or Something Near Enough*; and Kim, *Supervenience and Mind*.

This assumption is what we call *Fundamentalism*. Hence Kim's fundamentalist interpretation of CCP, or CCP^F for short, says that *only fundamentally physical singular events can nomologically sufficiently cause singular physical events*, which when slightly formalized looks like this:

> $\Box$ ($\forall$x) ($\exists$y) [(x is a simple or complex singular event & x is physical & y causes x) $\supset$ (y is a simple or complex singular event & y is fundamentally physical)].

If CCP^F is true, then it seems to open up a royal road for a reductive materialist solution to the mental causation problem. But we shall raise some serious doubts about CCP^F in Chapter 7.

Finally, PIM is supported by at least eight well-known arguments for irreducibility: (1) the Anomalism of the Mental, (2) the Mutiple Realizability Argument (3) the Modal Argument, (4) the Explanatory Gap Argument, (5) the Knowledge Argument, (6) the Absent Qualia Argument, (7) the Inverted Qualia Argument, and (8) the Zombie Argument. (For a brief descriptions of the arguments, see two paragraphs below.) Now to say that mental properties are *reducible* to physical properties is to say that *mental properties are either identical with or logically strongly supervenient on certain physical properties*.[22] Again, the basic idea behind logical strong supervenience is that *A*-properties are logically strongly supervenient on *B*-properties if and only if anything's *B*-properties are logically sufficient for the existence of its *A*-properties, and logically necessarily there cannot be a change in its *A*-properties without a corresponding change in its B properties. (For the explicit definition of strong supervenience and its various sub-species, see Section 1.1 above.) The identity of mental properties with certain physical properties is *ontological reduction*, while the logical strong supervenience of mental properties on certain physical properties is *explanatory reduction*. Ontological reduction (according to which there is either an identity of mental properties with certain fundamental physical properties or an identity of mental properties with certain second-order physical properties) is a necessary but not sufficient condition of explanatory reduction. So explanatory reduction entails ontological reduction, but ontological reduction does not in and of itself entail explanatory reduction. In turn, explanatory reduction can be either

[22] See, e.g., Chalmers, "Consciousness and its Place in Nature"; and Kim, *Philosophy of Mind*, ch. 10.

(i) *type physicalist*, in which case mental properties would logically strongly supervene on physical properties because fundamental mental properties are identical to certain fundamental physical properties (e.g., properties of the human brain),

or else

(ii) *functionalist*, in which case mental properties would logically strongly supervene on physical properties because fundamental mental properties are identical to certain *second-order* physical properties (e.g., computational-functional properties, or causal-functional properties) which in turn logically strongly globally supervene on certain fundamental physical properties.

(For brief definitions and explications of the various types of reductive and non-reductive materialism, see the Introduction.) Two equivalent ways of talking about the explanatory reduction of mental properties to physical properties are to say that mental properties are either "nothing but" physical properties or "nothing over and above" physical properties. In either case, the core of what is meant by the notion of explanatory reduction is that if one were to know everything there is to know about the physical world, then one would thereby *also* have a priori inferential knowledge of everything there is to know about the mental, including the knowledge of any identities there might be between mental properties and physical properties, and also the knowledge of any specifically lawful relations running between mental properties and physical properties (a.k.a. "bridge laws").

Since according to the notion of explanatory reduction, the knowledge of all physical properties by means of physicalistic concepts automatically carries with it the a priori knowledge of all mentalistic concepts and mental properties, it follows that if this a priori physical knowledge of mentalistic concepts and mental properties *fails*, then explanatory reduction fails, and thus the logical strong supervenience of the mental on the physical also fails. That is the basic rationale behind the eight well-known argument arguments for irreducibility, which include

(1) *the Anomalism of the Mental* (Davidson), which says that there are no strict deterministic psychophysical laws because of the semantic

holism of intentional content, hence mental properties are not identical with physical properties;[23]

(2) *the Multiple Realizability Argument* (Putnam), which says that functional properties of the mind can possibly be realized in a great many different kinds of compositional physical stuff, hence mental properties are not identical with fundamental physical properties;[24]

(3) *the Modal Argument* (Kripke), which says that since identity statements are necessarily true if true at all, and it is conceivable and therefore logically possible for there to be pains without brains, then the mind–brain identity theory is false;[25]

(4) *the Explanatory Gap Argument* (Nagel), which says that first-person or consciousness-based mentalistic concepts are irreducible to impersonal physicalistic concepts;[26]

(5) *the Knowledge Argument* (Jackson), which says that since it is possible for someone to know everything there is to know about the physical world but still fail to know what it is like to subjectively experience colors, then qualia do not logically strongly supervene on (and therefore are also not identical with) physical properties;[27]

(6) *the Absent Qualia Argument* (Block), which says that the conscious mind cannot be merely a functional organization because it is conceivable and therefore logically possible to realize the functional organization of the mind in a living physical system that does not have consciousness;[28]

(7) *the Inverted Qualia Argument* (a cast of thousands), which says that mental properties do not logically strongly supervene on (and therefore are also not identical with) physical properties because it is conceivable and therefore logically possible for me to have all the same physical properties that I possess in the actual world and yet subjectively experience the complete spectrum of colors in a reversed way;

and last but not least,

(8) *the Zombie Argument* (Chalmers), which says that mental properties do not logically strongly supervene on (and therefore are also not

[23] Davidson, "Mental Events." [24] Putnam, "The Nature of Mental States."
[25] See Kripke, *Naming and Necessity*, 144–55. [26] Nagel, "What is it like to be a bat?"
[27] Jackson, "Epiphenomenal Qualia." [28] Block, "Troubles with Functionalism."

> identical with) physical properties because it is conceivable and therefore logically possible for me to have all the same physical properties that I possess in the actual world, and yet altogether lack consciousness.[29]

So, at least on the face of it, PIM too is strongly supported by good reasons.

To be sure, each of these well-known arguments has provoked critical worries. For example, with respect to the Anomalism of the Mental, there is the worry that it leads to Epiphenomenalism.[30] With respect to the Multiple Realizability Argument, there is a worry about the very idea of a "physical realization" and whether it is nomologically possible for there to be more than one realization of minds like ours in the actual world.[31] With respect to the Knowledge Argument, there is a worry about a failure to distinguish between knowing how and knowing that.[32] With respect to the Gap Argument, there is a worry about a failure to distinguish between explanatory non-reduction and ontological non-reduction. Because the former does not entail the latter, it is possible to hold that mentalistic *concepts* are irreducible to physicalistic *concepts,* while still holding that fundamental mental *properties* are identical to certain fundamental physical *properties*, and thus that type physicalism is still true, even if (e.g.) Reductive Functionalism is false.[33] And finally with respect to the last six arguments (Modal, Gap, Knowledge, Absent Qualia, Inverted Qualia, and Zombie) there are of course the familiar worries about the general validity of the inference from conceivability to logical possibility and the confusion of conceivability with imaginability.[34]

We endorse PIM insofar as we endorse the failures of both ontological reduction and explanatory reduction alike. Nevertheless, we ourselves have used the Epiphenomenalism objection against Davidson in Section 3.2. And our own Essential Embodiment Thesis entails that it is impossible for

[29] See Chalmers, *The Conscious Mind*, ch. 4.

[30] See, e.g., Davidson, "Thinking Causes"; Kim, "Can Supervenience and 'Non-Strict Laws' Save Anomalous Monism?"; McLaughlin, "On Davidson's Response to the Charge of Epiphenomenalism"; and Sosa, "Davidson's Thinking Causes."

[31] See, e.g., Shapiro, "Multiple Realizations"; and Shapiro, *The Mind Incarnate*.

[32] See Lewis, "What Experience Teaches"; and Nemirow, "Physicalism and the Cognitive Role of Acquaintance."

[33] See Levine, "Materialism and Qualia: The Explanatory Gap"; and Levine, "On Leaving Out What It's Like."

[34] See Chalmers, *The Conscious Mind*, ch. 4; and Jackson, "Conceptual Analysis and Reductive Explanation."

there to be more than one sort of embodiment of conscious, intentional $minds_{lo}$, once we individuate types of embodiment in terms of the causal powers of the vital systems and organs of a suitably neurobiologically complex living body (see Sections 1.1 and 7.1). So we have no basic disagreement with critics of Anomalous Monism and Multiple Realizability, in the sense that we agree that the Anomalist Monism and Multiple Realizability arguments do not sufficiently support their non-reductive conclusions. We also are prepared to agree with the critics of the Knowledge Argument that the failure to distinguish between knowing-that and knowing-how vitiates Jackson's conclusions about what Mary knows and what she does not know, and with critics of the Gap Argument that there is no direct entailment from conceptual non-reduction to ontological non-reduction. For all these reasons, we will concentrate instead on replying to the final pair of familiar worries in order to argue for irreducibility.

As we have seen in Section 5.4, it is possible to avoid any skepticism about the general validity of the inference from conceivability to possibility, and also to avoid any skepticism about confusing conceivability and imaginability, just by choosing the right background logic for a priori conceptual arguments. So since we have officially selected *the A Priori Argument logic*, or APA logic, in which liberal conceivability is identical with logical possibility, and which also contains an "Anti-End-of-the-World Principle" that prevents it ever being the case that every proposition or sentence is both true and false, those skeptical worries immediately disappear.

The APA logic is a highly open-minded deviant logic. But as we emphasized in Chapter 5, of course that does not mean that *any* putative a priori conceptual inference about the mental and the physical is thereby acceptable. Our Anti-End-of-the-World Principle entails that it is *not* the case that anything goes. Thus the burden of proof is still on us to provide some compelling a priori conceptual arguments in specific support of the irreducibility of mental properties to physical properties. Nevertheless our adoption of the APA logic does indeed guarantee that, because we *generally can* infer validly from liberal conceivability to possibility, our burden of proof is not *impossibly heavy*. In line with that lighter demand, and beyond the eight well-known arguments we mentioned—each of which by now has already generated a large critical literature, and so is to that extent

somewhat tarnished—here are three shiny new arguments collectively in support of PIM.

(I) *The Bladerunner Argument.*

(1) The "Nexus VI replicants" represented in the classic science fiction movie *Bladerunner*[35] are artificially constructed living humanoids who have conscious, intentional states exactly like ours.
(2) Nexus VI replicants are liberally conceivable and therefore logically possible (in the APA logic).
(3) So mental properties can be instantiated in different kinds of biological, chemical, and microphysical stuff, and are not identical to the fundamental physical properties with which our mental properties are co-instantiated in the actual world.
(4) Therefore type physicalist reduction is false, and PIM is true.

(II) *The Intrinsic Structural Properties Argument.*

(1) According to the usual metaphysical interpretations of contemporary physics, the physical world is determined by either the inherent or intrinsic non-relational properties of fundamental microphysical particles or else the extrinsic relational properties of those particles. And according to reductive functionalists, mental properties are identical to extrinsic relational second-order physical properties, whether computational or causal.
(2) But it is liberally conceivable and therefore logically possible (in the APA logic) that mental properties are inherent or intrinsic relational (and more specifically spatiotemporal, hence intrinsic structural) properties of living organisms of a suitable degree of neurobiological complexity.
(3) So fundamental mental properties and physical properties, whether fundamental physical or second-order physical (e.g., functional), are essentially different types of property.
(4) Therefore mental properties are explanatorily irreducible to physical properties, and PIM is true.

[35] Directed by Ridley Scott (1982).

(III) *The Necker Cube Argument*.[36]

(1) Our conscious visual perceptions of the two enantiomorphic, or mirror-image-reversed, representations of the Necker Cube—call them *the subjective experience of Necker aspect A* and *the subjective experience of Necker aspect B* respectively—occur spontaneously.
(2) Now suppose that in the actual world brain state α partially embodies the subjective experience of Necker aspect *A*. It is liberally conceivable and therefore logically possible (in the APA logic), assuming that all physical properties in the natural world, including functional and behavioral properties, are held fixed, that brain state α might have partially embodied the subjective experience of Necker aspect *B*.
(3) So mental properties do not logically strongly globally supervene on fundamental physical properties.
(4) Therefore both explanatory reduction and ontological reduction are false, and PIM is true.

Since each of the eight well-known arguments, if sound, would also show that Reductive Materialism is false in one way or another, then obviously the three new arguments we just offered will not be *radically* different from the others. Still, the three new arguments do indeed differ from the other eight in some very important ways. So let us consider each of them briefly in turn.

Re: The Bladerunner Argument

The Bladerunner Argument is superficially similar to the Multiple Realizability Argument. But Multiple Realizability arguments in their classic Putnamian form depend heavily on intuitions about what constitutes a given realization of a mental kind. *Realization* is a technical notion that requires mind–body strong supervenience and the token identity of mental states with physical states.[37] As Lawrence Shapiro has correctly pointed out, however, while it is quite true that if one holds that realizations are individuated by sheer differences in compositional stuff then it follows

[36] See Hanna and Thompson, "Neurophenomenology and the Spontaneity of Consciousness"; and Lee, "The Experience of Left and Right."
[37] See Kim, "Multiple Realizability and the Metaphysics of Reduction."

that there are a great many possible realizations of any given mental kind, nevertheless it is arguable that this criterion of individuation leads to highly implausible consequences. For example, it implies that a red-colored version and a blue-colored version of exactly the same model of corkscrew, which obviously differ only very trivially, are distinct realizations of the functional kind *corkscrew*. Now if one holds, far more plausibly, that realizations are individuated by differences in *causal powers*, then it appears that the number of possible realizations of any given mental kind is drastically reduced, perhaps even to only one realization in the actual world.[38]

The Bladerunner Argument avoids this good objection to the Multiple Realizability Argument by simply pointing out that it is conceivable and therefore possible—in the APA logic, of course—that the same conscious, intentional mental kind can be *instantiated* in different biological, chemical, and microphysical stuffs, and also that the same living biological kind can be *instantiated* in different fundamental physical stuffs. This liberally conceivable possibility of the multiple instantiability of mind-and-life is the whole premise of the movie *Bladerunner* and also of the brilliant science fiction novel, Philip K. Dick's *Do Androids Dream of Electric Sheep?*, on which it is based. Multiple instantiation does not itself entail multiple realization, since realization is a stronger relation than instantiation, in the sense that it entails instantiation but is not entailed by instantiation alone. Nevertheless, from the logical possibility of the weaker relation of the multiple instantiability of mind-and-life, it follows directly that mental properties cannot be identified with fundamental physical properties. So the Bladerunner Argument says, in effect, that if you watch Ridley Scott's movie or read Philip K. Dick's novel and it actually *makes sense* to you, then type physicalist reduction is false, and PIM is true.[39] The metaphysics of realization is entirely irrelevant to this minimalist line of argument.

Re: The Intrinsic Structural Properties Argument

It is standard fare for contemporary theorists of consciousness to hold that the phenomenal characters of subjective experience are intrinsic non-relational features of mental states, or *qualia*, and also that physical properties

[38] See note 31 above.

[39] It is true that even if type physicalist reduction is false, functionalist reduction could still be true. But the Intrinsic Structural Properties Argument and the Necker Cube Argument are each strong enough to entail the falsity of Reductive Functionalism.

are extrinsic relational features of the fundamental physical particles, forces, and processes. This dual assumption seems to be implicit in the Modal Argument, the Gap Argument, the Knowledge Argument, the Absent Qualia Argument, most versions of the Inverted Qualia Argument, and the Zombie Argument. If qualia and the extrinsic relationality of fundamental physical properties are assumed by hypothesis, and one also assumes that qualia and physical properties are co-instantiated in the actual world, then since it is obvious just as a matter of simple logic and metaphysics that *extrinsic relational features of X* can exist without corresponding *inherent or intrinsic non-relational features of X*, then the non-reductive conclusion follows trivially.[40] In fact, this appears to be the Master Argument lying behind most, if not all, of the well-known arguments against Reductive Materialism.

Now there are two obvious replies to the Master Argument. The first is that there are simply no such things as qualia, and that qualia should therefore be *eliminated* from our best metaphysical theory of the world.[41] The second is that if at least some physical properties are *also* inherent or intrinsic non-relational features of the fundamental particles, then nothing prevents mental properties (which in this context of course means *properties of qualia*) from being *identical* to the inherent or intrinsic non-relational physical features of the fundamental particles.[42] In either case, all arguments based on the Master Argument will be unsound.

Looked at comparatively and contrastively in this way, the Intrinsic Structural Properties Argument has two very important philosophical advantages over the eight well-known non-reductive arguments:

(1) it does not assume the existence of qualia and in fact is also consistent with qualia eliminativism (see Section 2.3),

and

(2) it allows for mental properties to be inherent or intrinsic *relational* (and more specifically spatiotemporal, hence inherent or intrinsic structural) properties of physical things.

This avoids both of the objections to the Master Argument. To be sure, our allowing for mental properties to be inherent or intrinsic relational

[40] See Montero, "Post-Physicalism." [41] See, e.g., Dennett, "Quining Qualia."
[42] See Perry, *Knowledge, Possibility, and Consciousness*.

properties of physical things is also strategic for our own purposes, since on our view fundamental mental properties are inherent or intrinsic structural properties of living organisms of a suitable degree of neurobiological complexity, which is what we call *neo-Aristotelian hylomophism* (see Section 7.1). But on the other hand it is perfectly legitimate for us to appeal to the logical possibility of a thesis *we take to be actually true*, and this is not circular since we have independent reasons for defending neo-Aristotelian hylomorphism.

Re: The Necker Cube Argument

It is quite illuminating to compare and contrast the Necker Cube Argument with the Inverted Qualia Argument. The Inverted Qualia Argument asks us to conceive of a possible world in which we hold every physical property of the actual world fixed, and also systematically invert some complete range of phenomenal characters or qualities (usually characters or qualities of the visual experience of color) for some conscious mental subject. As far as we know, the inverted qualia hypothesis is nomologically or physically impossible.[43] Unlike the possibility of inverted qualia, however, Necker reversal is in fact universal in normal human perceivers as a sub-species of the familiar phenomenon of perceptual multistability, and furthermore requires only one visual presentation or visual image.[44] So since the argument step from actuality to logical possibility is obviously much easier than the argument step from conceivability to logical possibility, even in the APA logic, then the Necker Cube Argument should easily convince anyone who is even *minimally* or *momentarily* inclined to be convinced by the Inverted Qualia Argument.

No doubt it will be somewhat difficult for card-carrying eliminative or reductive materialists to accept any of our three new arguments for PIM. But at the same time it is somewhat unlikely that *any* argument, no matter how good, will ever convince a *card-carrying* philosopher of the truth of an opposing doctrine. Presumably, the best we can do for card-carrying eliminative or reductive materialists is to get them to be bemusedly rationally interested in what follows from our theory about mental causation *if* they were to grant PIM as well as CEM and CCP. So it is really those who are currently *somewhat* open-minded about the physical reducibility

[43] See Chalmers, *The Conscious Mind*, 263–6.

[44] See Hanna and Thompson, "Neurophenomenology and the Spontaneity of Consciousness."

of the mental that we seek to convince. Taken all in all, then, it seems to us rationally incontrovertible that the *whole package* of arguments for PIM is cumulatively compelling. Surely the eleven non-reductive arguments taken together, even allowing for various worries about some of them, are rationally onto *something* of fundamental philosophical importance.

We conclude from our argument so far, then, that the Amazingly Hard Problem of mental causation is a genuine philosophical paradox and not merely a philosophical puzzle.

Now what are we to do? Not surprisingly, the standard solutions to the Amazingly Hard Problem involve denying one or more of the premises. According to *Eliminative Materialism* or *Reductive Materialism*, mental properties either do not exist, or are identical to physical properties, or are logically strongly supervenient on physical properties, so that all causal activity is just physical. According to *Substance Dualist Causal Interactionism*, the mental and the physical are essentially distinct kinds of substance that also interact causally. According to *Causal Overdeterminationism*, the self-same physical event can have two or more complete and independent causal explanations, and thereby have two or more complete and independent causes, one of which is non-physical, because it is either (a) immaterial and non-spatial (= Substance Dualist Causal Overdeterminationism) or else (b) merely strongly superveniently mental (= Non-Reductive Materialist Causal Overdeterminationism). According to *Substance Dualist Causal Parallelism*, the mental and the physical are essentially distinct kinds of substance that do not interact causally, but nevertheless operate in fully law-governed and coordinated ways in separate ontological realms. According to *Causal Anti-Realism*, causal relations are all mind-dependent facts that are either empirically, conventionally, or else innately constructed by us, and thereby merely imposed on our sensory experiences. And finally, according to *Epiphenomenalism*, as we mentioned above, the mental is caused by the physical—or at least the mental is metaphysically fully determined by the causally efficacious physical world in both its existence and its specific character—but lacks any efficacious causal powers of its own. Thus eliminative and reductive materialists will deny PIM; substance dualist causal interactionists, causal overdeterminationists, substance dualist causal parallelists, and causal anti-realists will all deny CCP; and epiphenomenalists and substance dualist causal parallelists will both deny CEM.

But quite obviously, since any one of these *denialist* approaches to the Amazingly Hard Problem always flies in the face of a body of robustly compelling reasons for at least one of the basic premises of the Problem, a more philosophically adequate solution would be to find a re-interpretation of one or more of the three premises such that they come out jointly consistent. We call this re-interpretationist strategy an *affirmationist* approach to solving the Problem.

6.4 The Causal Exclusion Problems

Another way of formulating the problem of mental causation has been developed by Kim, and is conventionally dubbed "the Causal Exclusion Problem." Actually, Kim formulates *two* Causal Exclusion problems.

The first is *the Explanatory Causal Exclusion Problem*.[45] This says that since two or more complete and independent causal explanations for the same event or phenomenon cannot exist, there can be only one complete and independent causal explanation of a given event or phenomenon. This is *the Explanatory Exclusion Principle*, or EEP. As we noted in Section 3.2, "complete" explanations are self-contained and require no other concepts to apply to the relevant event or phenomenon. By contrast, "independent" explanations are complete and also rule out certain other concepts from applying to the relevant event or phenomenon at the same time and in the same respects. To motivate our acceptance of EEP, Kim asks us to consider all the possible cases in which there might be two causal explanations respectively invoking C (the mental cause) and C^* (the physical cause) of the same event E:

(case 1) identity of C and C^* (= either reductive materialist type-type identity theory or non-reductive materialist token-token identity theory),
(case 2) strong supervenience of C on C^* (= either reductive functionalism or non-reductive materialism),
(case 3) C and C^* are distinct individually insufficient, individually necessary, and jointly sufficient causes of E (= the jointly sufficient mental-and-physical cause theory),

[45] See Kim, "Mechanism, Purpose, and Explanatory Exclusion," esp. at 250.

(case 4) C and C^* are different links in the same causal chain leading to E (= substance dualist causal interactionism),

and

(case 5) C and C^* are distinct individually sufficient causes of E (= causal overdeterminationism).

Kim persuasively argues that all the putative cases of dual explanation are either non-independent because the two causal explanations collapse into a single complete and independent causal explanation of E (cases 1 and 3), or else they are incomplete because either C violates CCP (case 4), or else C^* explanatorily excludes C (cases 2 and 5). Two further points about Kim's argument should be particularly noted, however.

First, as Kim explicitly points out, case 5, or Causal Overdeterminationism, is crucially ambiguous as between

(i) *standard* overdetermination cases—such as the two assassins example and the belt-and-suspenders example mentioned earlier—in which the same physical effect is brought about by two individually nomologically sufficient but distinct *physical* causes,

and

(ii) *non-standard* overdetermination cases in which the same physical effect is brought about by two individually nomologically sufficient causes, one of which is *non-physical*.

Standard overdetermination cases can be understood as involving only one single complex complete and independent cause. For example, the two assassins' bullets arriving simultaneously can be taken to constitute one single complex cause of the assassinated person's death, and presumably this is what would be cited in the Coroner's Report as the fatal event. Correspondingly, the belt and suspenders taken together could be taken to constitute one single complex Integrated Trousers-Upholding System, or ITUS—and it should also be noticed that this is another everyday case of simultaneous and continuous causation. A special feature of standard causal overdetermination is that while each overdetermining cause is individually sufficient for its effect, they are also *jointly necessary for that effect*. Hence standard overdetermination cases arguably satisfy EEP.

But what about *non*-standard overdetermination cases? Here everything turns on the precise *kind* of non-physical overdetermining cause that is in play. If the non-physical overdetermining cause is an immaterial, non-spatial mental substance (say, a Cartesian soul) or a purely non-spatiotemporal agent-cause (say, a Kantian noumenal subject, or even a divine cause), then these are obviously ruled out by EEP, and they also clearly violate CCP on any plausible interpretation of it. This is because these causes operate altogether *independently* of the physical cause, even if the causal operations of the two overdetermining causes happen to coincide at the spacetime location of the physical effect. But if, on the other hand, the non-physical overdetermining cause is *systematically* related to the physical overdetermining cause—say, by naturally or nomologically strongly supervening on that physical cause according to psychophysical bridge laws and by operating according to non-reductive *ceteris paribus* laws of the special sciences—then its violation of either EEP or CCP is not so obvious. This is primarily because the Anomalism of the Mental (to the effect that there are no strict deterministic psychophysical laws), which is one of the main critical sticking points in Davidson's account of mental causation, is thereby ruled out. Thus, it seems to be at least arguable that the *systematic* non-standard overdetermining cause is as likely to satisfy EEP as the *standard* overdetermining cause. We will come back to this point shortly.

Second, and crucially for our argument in the next chapter, we need to highlight two essential differences between

(a) case 3, or the jointly sufficient mental-and-physical cause theory,

and

(b) case 5, or the causal overdetermination theory.

The first essential difference is that because in a jointly sufficient cause each element of the dual cause is individually *insufficient*, then a jointly sufficient cause automatically rules out causal overdetermination of any sort. The second essential difference, as Kim points out, is that just like the identity theory or case 1, the jointly sufficient cause theory or case 3 *clearly satisfies EEP*, because it clearly provides a single causal explanation of the effect *E* that is both complete and independent. By sharp contrast, the causal overdetermination theory or case 5 is crucially ambiguous as between

standard and non-standard overdetermination, and thus runs a serious risk of violating either CCP or EEP.

Kim's second causal exclusion problem is the *Supervenience Causal Exclusion Problem*.[46] This zeroes in on Non-Reductive Materialism, and says the following:

(1) Suppose that mental properties are strongly supervenient on fundamental physical properties.
(2) Further suppose that mental properties are known to be non-identical to fundamental physical properties by the Multiple Realizability Argument (or some other non-reductive argument).
(3) Then the strongly supervenient instantiation of a mental property in a physical event must be causally inert or epiphenomenal, because all the *real* nomologically sufficient causal work is done by the fundamental physical properties of the supervenience-base of that physical event.
(4) Therefore the fundamental physical properties of that event metaphysically trump and *exclude* the causal powers of any mental properties of the same event

So the non-reductive materialist theory of mental causation fails by reducing to Epiphenomenalism.

There is obviously a sense in which both the Amazingly Hard Problem and the two Causal Exclusion Problems cover the same patch of logical and metaphysical ground—the problem of mental causation. The Amazingly Hard Problem, however, has the special dialectical value of posing the problem in a completely general way. Correspondingly, the special dialectical value of the Causal Exclusion Problems lies in their narrower focus. We believe that Kim's argument is sufficient to rule out the substance dualist interactionist solution, the substance dualist causal overdeterminationist solution, and most especially the non-reductive materialist solution[47]

[46] See Kim, "The Myth of Nonreductive Materialism"; Kim, "The Non-Reductivist's Troubles with Mental Causation."

[47] That in turn would leave unrefuted only the reductive materialist, substance dualist parallelist, and causal anti-realist solutions to the Amazingly Hard Problem. All of these are denialist solutions, however, and thus each flies in the face of one or more of the basic, well-supported premises of the problem. Reductive Materialism violates PIM. Substance Dualist Parallelism violates CEM *and* CPM. And Causal Anti-Realism violates CPM, not to mention its also contradicting the causal realism we have adopted.

to the problem of mental causation. Indeed, Kim's argument is largely responsible for the serious comeback that Reductive Materialism has been making in the 2000s. But because Non-Reductive Materialism has been the favored theory—indeed, the *default* theory—for most philosophers of mind since the 1980s and 90s, we will briefly consider how non-reductive materialists have tried to reply to Kim, before moving on to our own solution to the Amazingly Hard Problem and Causal Exclusions Problems alike.

Not surprisingly, there has been a large amount of work done with the aim of saving Non-Reductive Materialism from the Causal Exclusion Problems. The most important and interesting of these attempted rescues have been made by Fodor,[48] Stephen Yablo,[49] Robert Van Gulick,[50] Derk Pereboom,[51] and the Macdonalds (Cynthia and Graham).[52] In our opinion however, all that this work, as interesting as it is, has been able to show is the following:

(1) If mental properties are epiphenomenal or causally inert then by the same token so are *macrophysical* properties, which is absurd (Fodor).
(2) If we construe the logical strong supervenience relation in terms of the determinable-determinate relation, then mental properties are sometimes causally relevant and figure directly in intentional explanations (Yablo).
(3) Systematic non-standard causal overdetermination (i.e., non-standard causal overdetermination that is combined with the nomological supervenience of mental events on the physical overdetermining causes according to psychophysical bridge laws, and the conformity of those mental events with non-reductive *ceteris paribus* laws of the special sciences) is as explanatorily acceptable as standard causal overdetermination (Van Gulick).
(4) If we assume that constitution is not identity, then we can reject the identity of mental types and tokens with physical types and tokens, and also metaphysically ground the causal powers of the mental in the causal powers of the physical (Pereboom).

[48] See Fodor, "Making Mind Matter More." [49] See Yablo, "Mental Causation."
[50] See Van Gulick, "Who's in Charge Here? And Who's Doing All the Work?"
[51] See Pereboom, "Robust Nonreductive Materialism."
[52] See Macdonald and Macdonald, "The Metaphysics of Mental Causation."

(5) Extrinsic or co-instantiated mental properties of causally efficacious physical events can be causally relevant to the effects of those events (the Macdonalds).

In counter-reply to (1), as Kim has persuasively argued, the epiphenomenality of macrophysical properties can be equally well interpreted as a metaphysical *causal inheritance* from fundamental physical properties, which in fact metaphysically *vindicates* those higher-level properties by conferring upon them the causal powers of their microphysical basis.[53] So at best Fodor can get only a draw on this point.

The non-reductive materialist replies in (2) through (5) can be considered as a single package. That single package says that *systematically* causally overdetermining nomologically strongly supervenient mental events that are upwardly determined according to psychophysical bridge laws and also operate according to non-reductive *ceteris paribus* laws of the special sciences, are *robustly causally relevant*, and therefore that those mental events are neither explanatorily nor metaphysically excluded by the causal powers of the physical events that are the supervenience-bases of those mental events. Hence Kim's claims that

(A) causally overdetermining supervenient mental events are explanatorily excluded,

and

(B) causally overdetermining supervenient mental events are epiphenomenal,

can both be rejected. In counter-reply to this single package of claims we want to insist, as Jackson has persuasively argued, that causal *relevance*, as nice as it is, and as robust as it might be, is just not the same thing as good old-fashioned causal *efficacy*.[54] Earlier in this section, we defined causal efficacy and causal relevance as follows:

Causal Efficacy: A singular event e_1 is causally efficacious if and only if either (i) e_1 is itself a nomologically sufficient simple singular event cause of some physical event e_2 or (ii) e_1 is a necessary proper part of e_3,

[53] See Kim, "Epiphenomenal and Supervenient Causation"; Kim, *Mind in a Physical World*; and Kim, "The Non-Reductivist's Troubles with Mental Causation."

[54] See Jackson, "Mental Causation," 397.

which itself is a nomologically sufficient complex singular event cause of e_2; a property P is causally efficacious if and only if P is instantiated as an inherent or intrinsic property by events that are causally efficacious; and a physical substance S is causally efficacious if and only if S is constituted by causally efficacious events and properties.

Causal Relevance: An event e_1 is causally relevant if and only if either (i*) e_1 is a necessary condition for some event e_3's being a nomologically sufficient cause of some physical event e_2 or (ii*) some correct description of e_1 enters directly into an informative characterization of e_3's being a nomologically sufficient cause of e_2; a property P is causally relevant if and only if some of P's instantiations are causally relevant; and a physical substance S is causally relevant if and only if S is constituted by causally relevant events and properties.

Now Jackson himself does not require that all mental events be causally efficacious. Indeed, he believes that phenomenal consciousness$_{lo}$ is thoroughly epiphenomenal.[55] But for us, the problem of mental causation will be solved *only if* the causal efficacy of conscious, intentional minds$_{lo}$ has been demonstrated. Demonstrating causal relevance is just not good enough. The popular formulation of the mental causation problem as "How can the mind make a difference in a physical world?" makes it seem that demonstrating causal relevance alone would actually do the required metaphysical job. But this is seriously misleading philosophical advertising, and a bit like saying that the problem of learning how to play a flute is to be solved by blowing in the thin end and running one's fingers very quickly up and down the side that has holes in it. Blowing in the thin end and running one's fingers very quickly up and down the side that has holes in it, even if it is obviously *relevant* to playing the flute, is perfectly consistent with being utterly unable to play the flute. So too, causal relevance is perfectly consistent with Epiphenomenalism.

Here is a more precisely formulated way of making the same critical point. Let us assume that the explanatory and metaphysical situation is exactly as the non-reductive materialist causal overdeterminationists—and in particular, Yablo, Van Gulick, Pereboom, and the Macdonalds—say it is. This can be made clear by a simple diagram that is very familiar

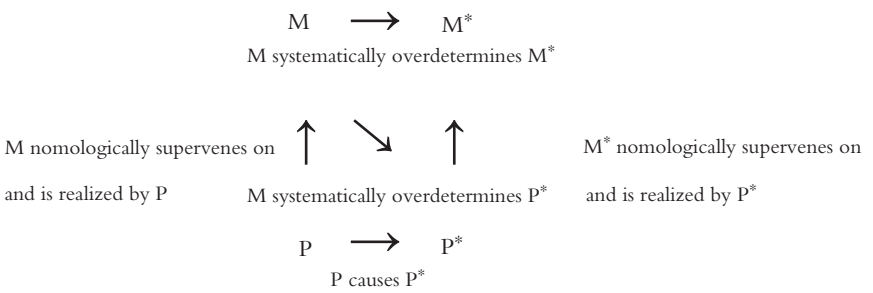
[55] See Jackson, "Epiphenomenal Qualia."

in the philosophy of mind literature since the appearance of Kim's causal exclusion problems.

In the diagram, we have adopted the following conventions:

- M = an event instantiating the fundamental mental property of my consciously willing to raise my right arm at time t_1
- P = an event instantiating the fundamental physical property of being the total state of my brain and body as I will to raise my right arm at t_1
- M* = an event instantiating the fundamental mental property of my consciously experiencing the raising of my right arm at time t_2
- P* = an event instantiating the fundamental physical property of being the total state of my brain and body as my right arm is raised at t_2

Mental Causation According to Non-Reductive Materialist Causal Overdeterminationism:



Now according to non-reductive materialist causal overderminationism, P and M together constitute a complete and independent causal explanation of P* and M*. Furthermore, P* and M* are each causally overdetermined by M. Also M and M* are strongly supervenient on P and P* respectively according to psychophysical bridge laws. And, finally, the relation between M and M* is governed by a *ceteris paribus* law of the special sciences (e.g., cognitive psychology). But it seems to us that given this explanatory and metaphysical situation, it is perfectly conceivable and (according to the APA Logic) thereby possible that in another world that is a "minimal physical duplicate" of the actual world[56]—where a minimal physical duplicate of the actual world is any logically possible world that is guaranteed to have all the same fundamental physical properties and all the same fundamental

[56] See, e.g., Braddon-Mitchell and Jackson, *Philosophy of Mind and Cognition*, 23–4.

physical laws as the actual world, but not guaranteed to have anything *else* from the actual world—the following situation obtains:

P *on its own* causes P*

Minimal physical duplicate worlds need not have either the same *psychophysical* laws or the same *ceteris paribus* laws of the special sciences as the actual world, since these do *not* strongly supervene on the fundamental physical properties and fundamental physical laws alone. For this reason, the psychophysical laws that support M's nomological strong supervenience on P (and, of course, also support M*'s strong supervenience on P*), can be just *missing* in such worlds, and thus *M too* (and, of course, also M*) can be missing in such worlds. Then, because P on its own causes P* in this minimal physical duplicate of the actual world, it follows that M cannot be doing *any* causally efficacious work at all in the actual world. At the very most, it could be argued that P and M together cause P* and M* *in an informatively different way* in the actual world, i.e., in a way that is also missing in the minimal physical duplicate world in which P alone causes P*, and both M and M* are missing. But that at most shows causal relevance, not causal efficacy.

For example, as Pereboom himself admits, there is no reason whatsoever to think that the material constitution relation confers any causal efficacy on mental types or mental tokens *as mental*. Hence M has no efficacious causal powers. In order for M to be causally efficacious it has to be true, at the very least, that M's mental properties are inherently or intrinsically related to the physical properties of P. But that cannot be true if M merely nomologically strongly supervenes on P, for if M merely nomologically strongly supervenes on P, then M's mental properties are at best *accidental* or *extrinsic* properties of P. Therefore, given the explanatory and metaphysical situation as described by non-reductive materialist causal overdeterminationism, M *cannot* be causally efficacious.

We conclude that, in view of the Amazingly Hard Problem, together with the robustness of the Causal Exclusion Problems, a new solution to the problem of mental causation is urgently required. And that is precisely what we will attempt to do in the next chapter.

7

The Metaphysics of Agency II: And How to Solve It

> The fundamental problem of mental causation for us [materialists], then, is to answer this question: How is it possible for the mind to exercise its causal powers in a world that is fundamentally physical?
>
> Jaegwon Kim[1]

> Understanding human action must begin from the assumption that people are dynamical entities whose behavior reflects their complexity.
>
> Alicia Juarrero[2]

> Where there is life there is mind, and mind in its most complex forms belongs to life. Life and mind share a core set of formal or organizational properties, and the formal and organizational properties distinctive of mind are an enriched version of those fundamental to life. More precisely, the *self-organizing* features of mind are an enriched version of the self-organizing features of life.
>
> Evan Thompson[3]

7.0 Introduction

In this chapter we describe and defend our new solution to the problem of mental causation, including both the Amazingly Hard Problem and the two Causal Exclusion Problems. The crux of our new solution is this:

> Nature basically includes complex dynamic organismic life, and essentially embodied $minds_{lo}$ are alive. So because organismic life is basically

[1] Kim, *Mind in a Physical World*, 30.
[2] Juarrero, *Dynamics in Action*, p. 221.
[3] Thompson, *Mind in Life*, p. ix.

> causally efficacious, then essentially embodied $minds_{lo}$ are basically causally efficacious too. In order to solve the problem of mental causation, you just *find* $minds_{lo}$ *in life,* from which it immediately follows that some essentially mental-and-physical complex singular events are jointly sufficient causes of other physical events.

The only possible ways to reject this line of thought would be to claim either that the principle of the Causal Closure of the Physical or CCP rules out organismic living events as physical events that efficaciously cause other physical events, or that the Causal Exclusion Problems show that organismic life is epiphenomenal. And neither of these options seems in any way philosophically viable, even for the most hardnosed materialist.

Although we fully agree with Kim's critiques of Dualism (whether Substance Dualism, or Property-Dualism-Without-Substance-Dualism) and Non-Reductive Materialism, we also think that he unfairly stacks the deck in favor of solutions based on Reductive Materialism. He does this by adopting a certain interpretation of the highly plausible principle CCP, which says that only physical events can nomologically sufficiently cause physical events. Kim's interpretation is also the *standard* interpretation. The standard interpretation of CCP assumes that the fundamental physical properties of the natural world necessarily exclude intrinsic connections with fundamental mental properties. In our terminology, for something *X* to be physical and also such that its fundamental physical properties necessarily exclude intrinsic connections with fundamental mental properties, is for *X* to be *fundamentally physical*. So in other words, the standard interpretation of CCP assumes that necessarily whatever possesses a fundamental physical property is fundamentally physical. This is the assumption of *Fundamentalism*.[4]

But Fundamentalism is seriously questionable, because it also assumes that contemporary natural science yields a knowledge of the real nature of the physical world that we have no good reason to think we actually possess, and in fact several good reasons to think that we do *not* possess, given the actual history of science. On the contrary, and consistently with reasonable scruples about the limits of contemporary natural

[4] It is an unfortunate accident of contemporary English that 'fundamentalism' carries various cultural and political connotations. But the word is so appropriate and handy that it would be a shame to have to avoid using it for purely non-philosophical reasons. And even if you don't like the connotations, it's not an obscenity! So needless to say, our use of it explicitly eschews the cultural and political connotations.

scientific knowledge about the physical world, we deny that fundamental physical properties necessarily exclude inherent or intrinsic connections with fundamental mental properties. Moreover, we also think that it is both metaphysically possible and also actually the case that fundamental physical properties *include* intrinsic connections with fundamental mental properties. We call this alternative metaphysical doctrine *Post-Fundamentalism*.

In Section 7.1, we propose an interpretation of CCP that incorporates Post-Fundamentalism (abbreviated as CCP^{PF}). Now Kim's fundamentalist interpretation of CCP (abbreviated as CCP^{F}) when made fully explicit, says:

(i) that only physical events can nomologically sufficiently cause physical events,

and

(ii) that the fundamental physical properties of the natural world necessarily exclude inherent or intrinsic connections with fundamental mental properties.

By contrast, our proposed post-fundamentalist interpretation of CCP, or CCP^{PF}, says:

(i) that only physical events can nomologically sufficiently cause physical events,

(ii*) that the fundamental physical properties of the natural world do not necessarily exclude inherent or intrinsic connections with fundamental mental properties,

and

(iii*) that it is both metaphysically possible and also actually the case that fundamental physical properties include inherent or intrinsic connections with fundamental mental properties.

The crucial thing about CCP^{PF} is that it allows for the metaphysical possibility and actual existence of what we call *mental-physical property fusion*. Taking CCP^{PF} together with mental-physical property fusion, and then combining it with the further notion of *a jointly sufficient essentially mental-and-physical cause*, enables us to avoid both of Kim's Causal Exclusion Problems and

also provide an adequate, *affirmationist* solution to the Amazingly Hard Problerm of mental causation. We call this adequate, affirmationist solution the Essentially Embodied Agency Theory *of mental causation*. This is because the three notions of CCP^{PF}, property fusion, and jointly sufficient mental-and-physical causation yield a natural metaphysical interpretation of the Essentially Embodied Agency Theory *of action* that we presented in Chapters 3–5.

In Section 7.2, we sketch a big metaphysical picture of a post-fundamentalist natural world in which mental-physical property fusion and jointly sufficient essentially mental-and-physical causation both actually exist—namely, *our* world, the complete natural world in which all minded animals, including ourselves, actually exist, and in which Thompson's $minds_{lo}$-in-life thesis is true—which we call *the Dynamic World*. Then in Sections 7.3 and 7.4 we further elaborate two of the central elements in the dynamic world picture: *dynamic systems theory* (DST), and *non-logical or strong metaphysical a priori necessity*. According to DST, intentional agents essentially are, as Juarrero aptly puts it in the second epigraph of this chapter, "dynamical entities whose behavior reflects their complexity." And according to the notion of non-logical or strong metaphysical a priori necessity, the mental properties and physical properties of intentional agents are bound together with a non-logical or strong metaphysical necessity that cannot be known by empirical means alone. Recondite as this may seem, it is also a surprisingly controversial idea. This is due to its close connections with the classical pre-Quinean and post-Quinean debate about the intelligibility and tenability of the analytic-synthetic distinction, and also with the more recent post-Kripkean debate about the existence and implications of the necessary a posteriori. So we conclude the chapter by defending the very idea of non-logical or strong metaphysical a priori necessity against the most important objections to it.

7.1 From Causal Exclusion to Property Fusion

As we pointed out in Section 6.3, the special dialectical value of the Causal Exclusion Problems consists in their collective ability to refute the substance

dualist interactionist, substance dualist causal overdeterminationist, and non-reductive materialist approaches to mental causation. Nevertheless, we also think that there is good reason to question the standard interpretation of CCP, and correspondingly, to close down the royal road it appears to open up for the reductive materialist solution to the problem of mental causation. According to the standard interpretation of CCP,

(i) only physical events can cause physical events,
(ii) a physical event is any real occupant of spacetime that possesses some fundamental physical properties,

and

(iii) fundamental physical properties necessarily exclude inherent or intrinsic connections with fundamental mental properties.

Now to say that something has a fundamental physical property, and thereby necessarily excludes inherent or intrinsic connections with fundamental mental properties, is to say that this thing is *fundamentally physical*. So taking (ii) and (iii) together entails what we have dubbed *Fundamentalism*, which says that necessarily if something *X* possesses a fundamental physical property then *X* is fundamentally physical. Since the standard interpretation of CCP strictly implies Fundamentalism, we have dubbed this interpretation CCP^F.

Ironically enough, and significantly, a very close relative of Fundamentalism also is defended by *Cartesian substance dualists*, who hold that necessarily if something *X* possesses a fundamental mental property, then *X* is fundamentally mental, in the sense that *X*'s fundamental mental properties necessarily exclude intrinsic connections with fundamental physical properties. Just to give this doctrine a name, we will call it "Funda-*Mentalism*." In fact, Fundamentalism and Funda-Mentalism alike are built implicitly into Cartesian Substance Dualism in the form of "the real distinction between mind and body" defended in the sixth *Meditation*. For they each assert one conjunct of what the "real distinction" argument directly entails, namely that fundamental physical properties and fundamental mental properties cannot be intrinsically connected. So Fundamentalism and Funda-Mentalism are, at bottom, really just two different sides

of the same explanatorily and ontologically exclusionary Cartesian coin. Indeed, Descartes explicitly states the conjunction of Funda-mentalism and Fundamentalism as a self-evident thesis in the *Principles*:

> To each substance there belongs one principal attribute; in the case of mind, this is thought, and in the case of body it is extension.[5]

In other words, no substance can have two principal attributes, and no substance can be essentially mental-and-physical. But what is the *argument* for this?

We think that both (i) and (ii) above are acceptable and true, but that (iii), and therefore also Fundamentalism, are unacceptable and false. The direct denial of Fundamentalism, which says that

> (iii*) fundamental physical properties do not necessarily exclude any inherent or intrinsic connections with fundamental mental properties, and it is both metaphysically possible and also actually the case that fundamental physical properties include inherent or intrinsic connections with fundamental mental properties,

is what we call *Post-Fundamentalism*. To say that a fundamental physical property includes an inherent or intrinsic connection with a fundamental mental property is to say that any such pair of properties exemplifies what we call *mental-physical property fusion*, which we will spell out in detail in a few paragraphs.

It should be noted here, however, that Post-Fundamentalism is consistent with at least three distinct positive metaphysical theses:

(a) the natural world is composed of a neutral or undifferentiated kind of thing that instantiates both fundamental physical properties and fundamental mental properties, but is itself neither fundamentally physical nor fundamentally mental,
(b) all fundamental physical properties in all of their natural-world instantiations necessarily include inherent or intrinsic connections with fundamental mental properties,

[5] Descartes, *Principles of Philosophy*, part I, §53, 210 (underlining added). Many thanks to Nathan Smith for calling this passage to our attention.

and

(c) some but not all fundamental physical properties, in some but not all of their natural-world instantiations, necessarily include inherent or intrinsic connections with fundamental mental properties.

Thesis (a) expresses the idea of a *Neutral Monism* of the sort defended by Spinoza, early Russell, and others, according to which nature is at bottom one undifferentiated kind of thing with two irreducibly distinct aspects.[6] Thesis (b) says that every part of the natural world at every time is actively or dispositionally conscious, and it is therefore a *Pan-Experientialism* of the sort explicitly defended by Alfred North Whitehead in the 1920s, and more recently proposed by Nagel, Chalmers, and Gregg Rosenberg.[7] But thesis (c) says that *only some parts* of the natural world at *only some times* are conscious—e.g., living organisms of a suitable degree of neurobiological complexity. This is what we will call *Emergent Experientialism*, because the notion of emergence captures the idea of irreducible inherent or intrinsic mental properties that are *naturally novel* in that they are instantiated in physical nature only under certain conditions and at certain times. In Section 7.2, we will argue explicitly in favor of options (a) and (c)—Neutral Monism and Emergent Experientialism—and against option (b) or Pan-Experientialism. And in Section 8.2, we will explicitly work out the salient metaphysical details of our theory of emergence. But for the time being, and as a necessary preliminary to that, here are two arguments against Fundamentalism and in favor of Post-Fundamentalism.

First, since we do not currently *know* what the nature of the physical world is, then we are in no position to assert with sufficient justification that the physical world is fundamentally physical. As Noam Chomsky points out:

[Materialism] will be a coherent position if its advocates tell us what counts as "physical" or "material." Until that is done, we cannot comprehend the doctrine, let alone such derivative positions as "eliminative materialism" and the like.[8]

[6] See Spinoza, *Ethics*; and Russell, *The Analysis of Matter*.

[7] See Whitehead, *Process and Reality*; Rosenberg, *A Place for Consciousness*; Nagel, "The Psychophysical Nexus"; Nagel, "Panpsychism"; and Chalmers, *The Conscious Mind*, ch. 8.

[8] Chomsky, *New Horizons in the Study of Language and Mind*, 85.

Chomsky's excellent point is closely connected with what Barbara Montero aptly calls "the Body Problem."[9] If on the one hand we look at the history of physics, and note the procession of failed theories, then it seems very likely that our current best physical theories are basically misguided and wrong. But if on the other hand we assert that the final physics will look pretty much like our current best physical theories, then we are either begging the question or else merely betting on future science without ruling out the possibility of a regress of knowledge. Thus far, we have what is sometimes called "Hempel's Dilemma." But because the materialist cannot justifiably assert that he *knows* what the real nature of the physical world is, it follows that Materialism is epistemically undersupported. That is the Body Problem.[10] What we would then add to the Body Problem is that if the materialist cannot justifiably assert that he knows what the nature of the physical world is, then of course he cannot justifiably assert the truth of *Fundamentalism* either, since sufficiently justifiying either thesis would require knowledge of the real nature of the physical world. So Fundamentalism is just as epistemically undersupported as Materialism itself.

Second—and here's the rub—we also think that the following situation is logically and metaphysically possible:

(1) Some event or physical substance *X* has some fundamental mental properties *M1, M2, M3*, etc.
(2) *X* also has some non-identical or distinct fundamental physical properties *P1, P2, P3*, etc.
(3) For every *Mi* there is a one-to-one correlation with a corresponding *Pi*.
(4) The members of each 1–1 correlated *Mi–Pi* pair are necessarily co-extensive.
(5) The members of each 1–1 correlated *Mi–Pi* pair are not logically necessarily co-extensive.
(6) The members of each 1–1 correlated *Mi–Pi* pair are mutually inherent or intrinsic structural properties of *X*.
(7) *X* is a suitably complex living organism.

We will call what is described by this seven-part description *mental-physical property fusion*. Mental-physical property fusion, of course, is simply an

[9] Montero, "The Body Problem."
[10] See also Crane and Mellor, "There Is No Question of Physicalism."

abstract characterization of the minds$_{lo}$-in-life thesis. But it exposes part of the latter's internal structure. We borrow the very apt term "property fusion" from Paul Humphreys's important work on the metaphysics of emergence. Humphreys's notion of property fusion is relevantly similar to ours, in that both notions entail a *dynamic* conception of property emergence—see Section 8.2. But Humphreys's notion is also crucially different in that it is based on nomological necessity, which presupposes *modal monism*, and also *the Layered World picture*—see Section 7.3.

In any case, *our* notion of mental-physical property fusion says that some suitably complex living organisms are *essentially mental-and-physical*. Needless to say, we are talking about essentially embodied conscious, intentional minds$_{lo}$. If mental-physical property fusion is logically and metaphysically possible, that is, if suitably complex living organisms with essentially embodied consciousness$_{lo}$ are possible, precisely because they are already actualized in minded animals, then Fundamentalism is false and Post-Fundamentalism is true. Fundamentalism entails that mental-physical property fusion is logically and metaphysically *impossible*. But in direct refutation of that, Post-Fundamentalism says that fundamental physical properties do not necessarily exclude inherent or intrinsic connections with fundamental mental properties and that it is metaphysically *possible* for fundamental physical properties to include inherent or intrinsic connections with fundamental mental properties—and this is immediately entailed by the logical and metaphysical possibility of mental-physical property fusion.

It seems clear, moreover, that mental-physical property fusion *really is* both logically and metaphysically possible, precisely because it is already *actualized* in the minds$_{lo}$-in-life relation. But the metaphysical relation of property fusion is not restricted to the minds$_{lo}$-in-life relation, and in fact *already actually exists in nature in other basic forms as well*. Property fusion in general says of two properties *P1* and *P2* that:

(i) *P1* and *P2* are are necessarily co-extensive,
(ii) *P1* and *P2* are not logically necessarily co-extensive,
(iii) *P1* and *P2* are mutually inherent or intrinsic structural properties of anything *X* that co-instantiates them,

and

(iv) *X* is a spatiotemporal entity or fact.

The basic idea behind property fusion, then, is that it captures the natural modal metaphysical phenomenon of *the complementarity of properties*. As we are using this notion, complementary properties are non-identical properties that are nevertheless necessarily reciprocally inherently or intrinsically structurally correlated in every actual and possible member of some domain of spatiotemporal entities or facts.

A good example of complementary properties taken from applied geometry is provided by the relationship between concavity and convexity in the domain of finite material curved figures. A necessary condition of the exact identity of any two properties is that they are logically necessarily co-extensive. But there is no *logical* impossibility in the thought that *X* is concave but not convex. It is certainly liberally conceivable in the APA logic that there could be a world that consists entirely and intrinsically in a single infinite concave surface—call this a *hyperbolic* or *Lobachevskian world*—and nothing else. So concavity and convexity are not logically necessarily co-extensive, and thus they are not identical properties. Nevertheless, concavity and convexity are obviously necessarily reciprocally intrinsically structurally correlated in the domain of finite material curved figures. By its very nature, no finite material curved figure in our actual world or in any possible world containing the material stuff of our actual world can instantiate concavity without also instantiating convexity, and conversely. Therefore concavity and convexity are *non-logically* necessarily reciprocally intrinsically structurally correlated in the domain of finite material curved figures.

Another example of complementary properties, this time from physics, is the non-logically necessary reciprocal inherent or intrinsic structural relationship between the particle-position and particle-momentum in *quantum entanglement*:

> What Schrödinger showed was that if two particles are prepared in a quantum state such that there is a matching correlation between two 'canonically conjugate' dynamical quantities—quantities like position and momentum whose values suffice to specify all the properties of a classical system—then there are infinitely many dynamical quantities of the two particles for which there exist similar matching correlations: every function of the canonically conjugate pair of the first particle matches with the same function of the canonically conjugate pair of the second particle.[11]

[11] Bub, "Quantum Entanglement and Information."

And a third example of complementary properties, this time from biology, and therefore closer to the $minds_{lo}$-in-life relation, is the non-logically necessary reciprocal inherent or intrinsic structural relationship between DNA-structure and organismic structure in *cellular life*.[12]

So in this way the metaphysics of property fusion has a significantly broad application beyond our proposed use of it in the philosophy of mind, and clearly establishes both the real logical and metaphysical possibility of mental-physical property fusion, and its actualization in the natural world.

It should be particularly noted that mental-physical property fusion modally binds distinct mental and physical properties together as closely as possible but still short of type-type identity. So the relation of mental-physical property fusion is *modally the next strongest mental-physical relation to type-type identity*. This is very important because it means that mental-physical property fusion captures a version of the mental-physical relation that has all the important modal-metaphysical advantages of the type-type identity theory, and especially the strong modal reciprocity and symmetry (or "two-wayness") of necessarily co-extensive properties, but without any of the well-known disadvantages of Reductive Materialism—as demonstrated, e.g., by the eleven arguments against Reductive Materialism that we canvassed in Section 6.3.

It should also be particularly noted that property fusion is *not* the same as the bilateral strong supervenience of properties, according to which property *P1* strongly supervenes on property *P2* and conversely. According to property fusion, two fused properties are not only necessarily co-extensive—and thereby both necessarily co-variant and necessarily co-dependent—but also *mutually inherent or intrinsic structural properties of* whatever instantiates them, which is to say that they are mutually necessary relational spatiotemporal features of that sort of thing. In other words, they belong to the *natural essence* of that sort of thing. By contrast, since strong supervenience is merely a strongly modal *co-variation* and *dependency* relation, and not itself a relation of *natural essence*, two properties can be bilaterally strongly supervenient without being necessarily co-extensive mutually inherent or intrinsic structural properties of whatever instantiates them. So while property fusion entails bilateral supervenience, property

[12] See, e.g., Weber, "Life."

fusion is nevertheless an inherently more *intimate* modal metaphysical relation.

Furthermore, the inherently greater "modal intimacy" of property fusion is significantly increased by its being a relation of non-logical or *strong* metaphysical a priori necessity, and not logical or *weak* metaphysical a priori necessity. The a priority of something is its underdetermination by all empirical facts and sensory experiences. Logical necessity, or truth of a proposition in all logically possible worlds, obviously is a priori. By contrast, non-logical or strong metaphysical a priori necessity means that a proposition or property-connection holds in a restricted class or "space" of logically possible worlds, not in *all* logically possible worlds, and that it cannot be known by empirical means alone. Restricted classes or spaces of logically possible worlds follow from special universal intrinsic structural constraints being placed on the constitution of possible worlds, over and above logical consistency. Such constraints reflect the global spatiotemporal, causal, and mathematical architecture of the actual world, whatever that happens to be. The relation between such constraints and the class or space of possible worlds is one of inverse proportionality. A richer set of constraints entails a more restricted class or space of possible worlds, and conversely a more restricted class or space of possible worlds entails a richer set of constraints. For our purposes here, what this means is that the relation of property fusion, as a relation involving non-logical or strong metaphysical a priori necessity, is *inherently richer* than the relation of exact property identity, although it belongs to the same general family of "two way" strong modal relations to which exact property identity also belongs. The other modal relations in this general family include the numerical identity of individuals, the necessary equivalence of propositions, the intensional equivalence or analytic identity of properties or propositions, the synonymy of predicates or sentences, and so on. Furthermore, because the constraints are universal, intrinsic, and structural, they cannot be known by empirical means alone and thus are a priori. We will come back again to the crucial idea of non-logical or strong metaphysical a priori necessity in Section 7.3.

In any case, if Fundamentalism is highly dubitable because of the extended Body Problem, and if Fundamentalism is also arguably false because mental-physical property fusion is both really possible and actualized in the natural world, then obviously adding the assumption of Fundamentalism to CCP,

as the standard interpretation of it does, renders CCP^F arguably false. But when CCP is considered *apart* from Fundamentalism it seems to be obviously true. Therefore the *correct* interpretation of CCP must be the Post-Fundamentalist interpretation CCP^{PF} and *not* CCP^F. Or in other words, the *correct* reading of the Principle of the Causal Closure of the Physical must be this one:

(i) only simple or complex singular physical events can nomologically sufficiently cause simple or complex physical events,
(ii) a singular physical event, as the real occupant of some spacetime extension, possesses some fundamental physical properties,

and

(iii) fundamental physical properties do not necessarily exclude inherent or intrinsic connections with fundamental mental properties and it is metaphysically possible and also actually the case that fundamental physical properties include inherent or intrinsic connections with fundamental mental properties.

Let us now suppose that CCP^F is false and that CCP^{PF} is true, and reconsider the Causal Exclusion Problems. What CCP^{PF} entails is that it is metaphysically possible and also actually the case that something that is not only physical, but also *essentially mental-and-physical*, by virtue of mental-physical property fusion, is able to cause something else that is physical, without in any way violating CCP. Again, of course, this essentially mental-physical thing is a suitably neurobiologically complex living organism with an essentially embodied conscious, intentional $mind_{lo}$. Organismic life is basically causally efficacious, and essentially embodied $minds_{lo}$ are alive. Therefore $minds_{lo}$ are basically causally efficacious. So let us now postulate that causally efficacious mental-physical property fusion is not just metaphysically possible but also actualized in nature in *minded animals*. This allows us to reinterpret the crucial case (3) in Kim's formulation of the explanatory exclusion problem, the case of a jointly sufficient cause, as *a jointly sufficient essentially mental-and-physical cause (i.e., a suitably neurobiologically complex living organism with an essentially embodied conscious, intentional* $mind_{lo}$*) that is fully consistent with the correct interpretation of CCP.*

Now a jointly sufficient essentially mental-and-physical cause that is fully consistent with the correct interpretation of CCP clearly avoids both

of the Causal Exclusion Problems. First, as we noted above, the jointly sufficient cause theory rules out causal overdetermination. Second, as we also noted above, and as Kim himself points out, the jointly sufficient cause theory satisfies the Explanatory Exclusion Principle or EEP, so there is no explanatory causal exclusion worry for jointly sufficient causes. Third, a jointly sufficient essentially mental-and-physical cause that is fully consistent with the *correct* interpretation of CCP is automatically also fully consistent with CCP. Fourth and finally, since in a jointly sufficient essentially mental-and-physical cause there is a fully "two-way" or symmetric non-logical or strong metaphysical a priori necessitation relation between the mental and the physical, then the mental is *neither* asymmetrically dependent—that is, either logically or nomologically strongly supervenient—on the physical, as in Non-Reductive Materialism, *nor* is it reducible to bilateral strong supervenience, since bilateral strong supervenience is not a modal relation of natural essence. Hence there is no supervenience causal exclusion worry for a jointly sufficient essentially mental-and-physical cause either.

As we noted in Section 7.0, the only possible ways to reject the crux of our solution to the problem of mental causation would be to claim either that CCP rules out organismic living events as physical events that efficaciously cause other physical events, or that the Causal Exclusion Problems show that organismic life is epiphenomenal. But both options are philosophical non-starters. How could basic physics be in causal competition with basic chemistry and basic biology?

It should be more than obvious by now that *essentially embodied agency* in the sense we spelled out in Chapters 1 to 5—whereby a motile, egocentrically-centered and spatially oriented, thermodynamically irreversible suitably neurobiologically complex living organism with an essentially embodied conscious, intentional mind_{lo} is the emotive cause of its own basic intentional actions by means of synchronous effortless trying and its active guidance—is the jointly sufficient essentially mental-and-physical cause of intentional body movements, and is fully consistent with the correct and post-fundamentalist interpretation of CCP, i.e., CCP^{PF}. For convenience, we will call the four-fold conjunction that consists of the Essentially Embodied Agency theory of action, plus CCP^{PF}, plus mental-physical

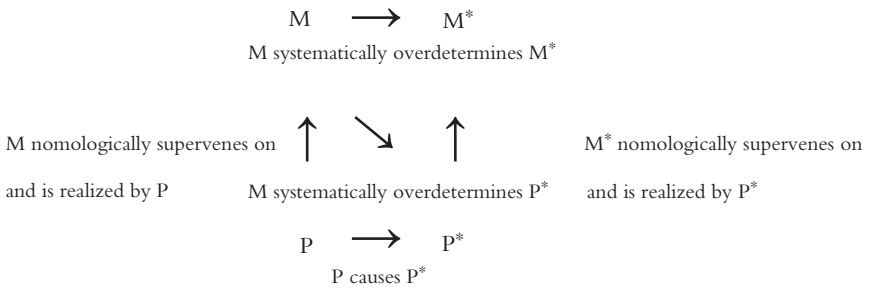
property fusion, plus jointly sufficient causes, *the Essentially Embodied Agency Theory of mental causation*. Then we can easily see that the Embodied Agency Theory of mental causation *solves* both of the Causal Exclusion Problems.

Even if one concedes that our argument for a new solution to the Causal Exclusion Problems is sound, it may at first glance seem ad hoc and "Scholastic." It is true that the contemporary debate about mental causation occasionally carries a whiff of ad-hockery and bad Scholasticism. But our new solution is *not* driven by the need to find some ingenious way of solving a philosophical brain-teaser. Rather, it is driven by the much deeper need to bring our metaphysics of mental causation into line with our basic intuitions about intentional agency, and our neurophenomenology. The philosophy of mind is *not* an intellectual game. It is about *the nature of minds*$_{lo}$ and thus is about *our own nature*. We want to *know ourselves*— to know what we really are and who we really are. So our new solution to the mental causation problem fully keeps faith with our deepest metaphysical and neurophenomenological commitments. Again, subtle metaphysical details apart, the bottom line of our new solution to the problem of mental causation is this: *Find minds*$_{lo}$ *in life*.

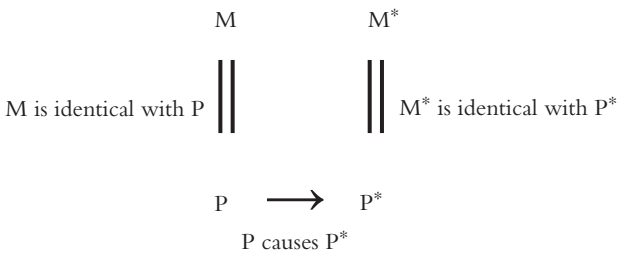
Moreover, we think that the theoretical advantages of our Essentially Embodied Agency Theory of mental causation become virtually self-evident when we diagramatically compare and contrast it with the two favored contemporary solutions to the problem of mental causation: Non-Reductive Materialist Causal Overdeterminationism, and Reductive Materialism. In the following three diagrams, we have again adopted the following conventions:

- M = an event instantiating the fundamental mental property of my consciously willing to raise my right arm at time t_1
- P = an event instantiating the fundamental physical property of being the total state of my brain and body as I will to raise my right arm at t_1
- M* = an event instantiating the fundamental mental property of my consciously experiencing the raising of my right arm at time t_2
- P* = an event instantiating the fundamental physical property of being the total state of my brain and body as my right arm is raised at t_2

1) Mental Causation According to Non-Reductive Materialist Causal Overdeterminationism:



2) Mental Causation According to Reductive Materialism:



3) Mental Causation According to the Essentially Embodied Agency Theory:

M is fused with P M/P ⟶ M*/P* M* is fused with P*

M/P jointly sufficiently causes M*/P*

According to Non-Reductive Materialist Causal Overdeterminationism, the mental event or property M is completely causally efficaciously otiose, *even if* we grant that it systematically causally overdetermines a corresponding mental or physical effect M* or P*, and is causally relevant. The mental event or property M is merely *along for the causal ride*, while its fundamental physical supervenience-base P does all the causally efficacious work.

According to Reductive Materialism, by contrast, the causal powers of the mental event or property M are *nothing but* the causal powers of the physical property P with which it is identical. Thus Physicalism can indeed "save" the efficacious causal power of the mental, but *only* in the same sense in which, according to the notorious remark of a US general during

the Vietnam War, sometimes the only way to save a village is to destroy it. In other words, physicalism can "save" the efficacious causal power of the mental only by *reducing* it to the efficacious causal power of the physical. With philosophical friends like that, who needs enemies?

Therefore the *only* one of the three theories of mental causation that gives mental properties efficacious causal powers *as mental* is the Essentially Embodied Agency Theory.

But there is even more. We now also have in hand an *affirmationist* solution to the Amazingly Hard Problem that we sketched in Section 6.2. Here, again, is the Problem.

(1) *The Causal Efficacy of the Mental* (CEM): Conscious, intentional minds_{lo}—in particular, conscious, intentional minds_{lo} insofar as they are engaged in intentional actions—can cause physical events.
(2) *The Causal Closure of the Physical* (CCP): Only physical events can cause physical events.
(3) *The Causal Physicality of the Mental* (CPM): In order to cause physical events, conscious, intentional minds_{lo} must be physical. [From (2)]
(4) *The Physical Irreducibility of the Mental* (PIM): Because mental properties are irreducible to physical properties, conscious, intentional minds_{lo} are non-physical.
(5) *The Causal Failure of the Mental* (CFM): So conscious, intentional minds_{lo} cannot cause physical events. [From (3) and (4)]
(6) Therefore conscious, intentional minds_{lo} both can and cannot cause physical events. [From (1) and (5)] **Contradiction!**

And here is the solution to the Problem. Suppose that the Essentially Embodied Agency Theory of mental causation is true. Then obviously the Causal Efficacy of the Mental or CEM is true too, because essentially embodied conscious, intentional minds_{lo} can cause intentional body movements by means of synchronous effortless trying and its active guidance. By hypothesis, the correct interpretation of the principle of the Causal Closure of the Physical or CCP—i.e., the post-fundamentalist interpretation CCP^{PF}—is also true. Now the Causal Physicality of the Mental or CPM follows directly from CCP, and thus CPM automatically becomes a post-fundamentalist version of CPM, or CPM^{PF} for short. Then CPM^{PF} is true too, precisely because essentially embodied conscious, intentional minds like ours can efficaciously cause physical things

by being inherent or intrinsic proper parts of jointly sufficient essentially mental-and-physical complex singular event causes of physical things. And finally the Physical Irreducibility of the Mental or PIM is also true, because according to mental-physical property fusion, mental properties are neither identical to nor logically supervenient on physical properties, and thus are irreducible to physical properties, even if mental and physical properties are non-logically or strongly metaphysically a priori necessarily co-extensive and intrinsically related. So if the Essentially Embodied Agency Theory of mental causation is true, and if Post-Fundamentalism is true, then there is *no ultimate inconsistency* between CEM, CCP, and PIM, and we can now safely reject the Causal Failure of the Mental or CFM, hence the Amazingly Hard Problem is thereby adequately and affirmatively solved.

Stepping back now for a moment and taking a larger view of things, we can clearly see that what was covertly generating the paradox in the Amazingly Hard Problem was the standard, but incorrect interpretation of CCP as CCP^{F}. Or in other words, the hidden source of the philosophical paradox was the widely shared false assumption that fundamental physical properties necessarily exclude intrinsic connections with fundamental mental properties—i.e., *Fundamentalism*. It is equally clear that CCP in and of itself was *not* the problem, for it nicely survives the solution of the Amazingly Hard Problem in the form of CCP^{PF}. In this way, since Fundamentalism was covertly *vitiating* CCP, once we dropped Fundamentalism we were able to save CCP from a fate worse than death—i.e., obvious falsity—reject CFM, and solve the paradox.

Now anyone who explicitly accepts the logical and metaphysical intelligibility and actual existence of mental-physical property fusion, and thereby implicitly or explicitly rejects both Fundamentalism and Funda-Mentalism alike, and also implicitly or explicitly rejects the fundamentalist interpretation of CCP, is a post-fundamentalist. In turn, Post-Fundamentalism, together with the assertion of the actual truth of mental-physical property fusion, entails what Gregg Rosenberg aptly calls *Liberal Naturalism*,[13] which says

(1) that nothing exists over and above the natural world,

[13] Rosenberg, *A Place for Consciousness*, 8–10.

and also

(2) that the natural world contains mental properties, and also basic laws governing the causal powers of essentially mental-and-physical events, as necessary proper parts of its basic ontology.

Organismic life belongs to the basic ontology of nature, and $minds_{lo}$ are alive; therefore $minds_{lo}$ belong to the basic ontology of nature. In this way, we are post-fundamentalists and also liberal naturalists.

We are also Essentially Embodied Agency theorists. So to summarize what we have argued in this chapter so far, we believe that in order to avoid both of the Causal Exclusion Problems and also to get a fully adequate and affirmationist solution to the Amazingly Hard Problem of mental causation, one should do two things. First, one should reject the fundamentalist interpretation of CCP, replace it with the post-fundamentalist interpretation of CCP. And second, one should also assert four further doctrines:

(1) the Essentially Embodied Agency Theory of action,
(2) the $minds_{lo}$-in-life thesis,
(3) the existence of mental-physical property fusion in essentially embodied agency,

and

(4) the existence of jointly sufficient essentially mental-and-physical causes in essentially embodied agency.

This two-step strategy will not only solve the problem of mental causation but also entail the truth of Liberal Naturalism.

7.2 The Dynamic World

Of course we are quite aware that dualists and materialists are not likely to be *fully* convinced by our arguments in the last section, to put it somewhat optimistically. Nevertheless, and on the one hand, we do think that—unlike Dualism and Materialism—the Essentially Embodied Agency Theory of mental causation offers a truly radical solution to the problem of mental causation, and thereby makes some much

needed conceptual progress in this area. As Nagel aptly remarks in this connection:

> My reading of the situation is that our inability to come up with an intelligible conception of the relation between mind and body is a sign of the inadequacy of our present concepts, and that some development is needed.[14]

Yes, *with bells on*: the Essentially Embodied Agency Theory is *precisely* that needed development.

But on the other hand, precisely because our theory of mental causation is *neither* dualist *nor* materialist, then it is quite likely it will seem somewhat disorienting to dualists and materialists. It may also seem somewhat disorienting even to those cautiously or skeptically uncommitted philosophers who regard Dualism and Materialism as the only viable options but also cannot find any way of accepting either of the classical alternatives, see no way out of this theoretical cul de sac, and are as it were just hopelessly sitting around and waiting for a philosophical Godot.

Any significant deviance from classical or standard norms, whether in personal life, social practices, natural science, or philosophy can induce its own peculiar sort of vertigo. Just think of the iconically famous zoom-in, track-out shot as James Stewart attempts to climb that fatal tower in Alfred Hitchcock's stunning 1958 psychodrama *Vertigo*. To avoid all this, you have to be able to feel the ground under your feet and see the sky over your head, feel your own living body in impulsive intentional movement, and know exactly where you are, and what you are, and who you are. Easier said than done—but one must try.

So there is a real need for us to try to provide more cognitive orientation, both by saying something about the big metaphysical picture we are offering of the natural world, which we call *the Dynamic World*, and also by providing an elaboration of two central elements of this picture:

(1) dynamic systems theory or DST,

and

(2) non-logical or strong metaphysical a priori necessity.

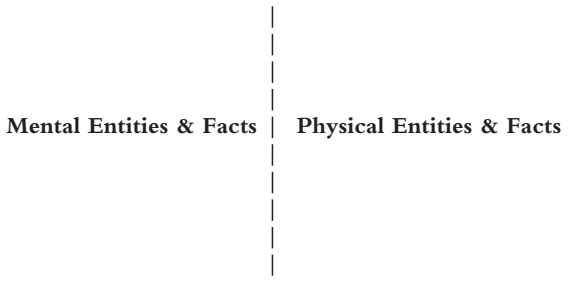
If these two elements can be made to seem more familiar and intuitively plausible, then the way will be open to a genuine three-way philosophical

[14] Nagel, "Conceiving the Impossible and the Mind-Body Problem," 338.

debate between dualists, materialists, and essentially embodied agency theorists (a.k.a. "mind-body animalists"—see Section 8.1).

According to Cartesian Interactionist Substance Dualism, the world consists of two essentially distinct kinds of substance (*mind* and *matter*) and correspondingly of two essentially different kinds of property (*mental* and *physical*), each of which constitutes a domain of logically and metaphysically distinct substantial particulars (*minds* and *bodies*) under that kind and instantiating those properties. Then those two kinds of substances, properties, and substantial particulars are by some entirely unexplained means—perhaps as a result of God's incomprehensible and all-powerful will—supposed to interact causally, despite their splendid mutual logical and metaphysical isolation. This is of course the classical early-modern metaphysical picture of the Bifurcated World:

THE BIFURCATED WORLD



Historically speaking, the Cartesian Bifurcated World picture did not survive the rise of modern natural science. As Kim has correctly observed, since the seventeenth century

> the Cartesian model of a bifurcated world has been replaced by that of a layered world, a hierarchically stratified structure of "levels" or "orders" of entities and their characteristic properties. It is generally thought that there is a bottom level, one consisting of whatever microphysics is going to tell us are the most basic physical entities out of which all matter is composed (electrons, neutrons, quarks, or whatever). And these objects, whatever they are, are characterized by certain fundamental physical properties and relations (mass, spin, charm, or whatever). As we ascend to higher levels, we find structures that are made up of entities belonging to the lower levels, and, moreover, the entities at any given level are thought to be characterized by a set of properties distinctive of that level.[15]

[15] Kim, "The Non-Reductivist's Troubles with Mental Causation," 190.

The Layered World picture began to emerge in Boyle's seventeenth-century "corpuscularian" theory of matter, and took its final shape in the early twentieth-century Rutherford-Bohr atomic theory of matter. More generally, the Layered World picture is intimately bound up with the parallel developments of particle physics and microscopy.[16] The Layered World is a world of *increasingly small microphysical compositions*, apparently all the way down, such that each lower level or stratum of reality is populated by a different sort of smaller material particle, out of which all the entities at higher levels are constructed.

Just as the Bifurcated World picture belongs to Substance Dualism, so too the Layered World picture belongs to *Materialism*. This is because in the Layered World the relation between the layers is one of asymmetric, non-reciprocal or one-way "upwards" modal dependence based on the part-whole relation: the higher levels are all ultimately either identical with or (logically or nomologically) strongly mereologically supervenient on the lower levels, in the sense that higher levels are entirely built out of smaller and smaller items occurring at the lower levels:

THE LAYERED WORLD

Mental facts

mereological ↑ supervenience

Biological facts

mereological ↑ supervenience

Chemical facts

mereological ↑ supervenience

Molecular, atomic, and quantum facts

The fatal metaphysical flaw in the Bifurcated World picture was the incomprehensibility of the causal relationship between the two essentially distinct domains of mental and physical facts. But now there seem to be *two* fatal metaphysical flaws in the Layered World picture.

The first flaw is the great difficulty of reconciling *inert particles* with *active forces*, which leads to the several equally difficult sub-problems of understanding action-at-a-distance, the aether, relativity, gravity, electromagnetic fields, waves, "wavicles," quantum phenomena, and so on.

[16] See, e.g., Galison, *Image and Logic: A Material Culture of Microphysics*; and Wilson, *The Invisible World: Early Modern Philosophy and the Invention of the Microscope*.

Neither relativity theory nor quantum mechanics conforms especially well to the Layered World picture.

The second flaw in the Layered World picture is the great difficulty of understanding the nature of the conceptual, ontological, and causal *gaps or transitions* between levels, which is the same as the problem of reconciling *the continuity of downward decomposition* with *the discontinuity of upward evolution*, especially at the levels of biological and mental facts. The possibility of a downward decomposition of all entities and facts at any given level into mereological sums occurring at lower levels in the hierarchy strongly suggests that all the higher levels should explanatorily, ontologically, or at least causally collapse down onto the bottom level. But upward evolution of the levels over physical time strongly suggests, contrariwise, that each new higher level has its own conceptual, ontic, or causal integrity and thereby resists any such downward collapse. This downward vs. upward tension in the Layered World picture provided by Materialism seems in the end to be every bit as theoretically vitiating as the bilateral dichotomy in the Bifurcated World picture provided by Substance Dualism.

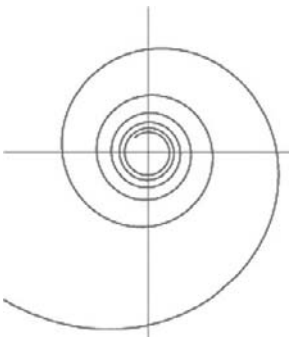
Despite their obvious differences, both the Bifurcated World and Layered World pictures also share a few enabling assumptions. According to both the Bifurcated World picture and also the Layered World picture, the physical world is fundamentally physical and the mental world is fundamentally mental, in that neither fundamental physical properties nor fundamental mental properties can have intrinsic connections to one another. So the Bifurcated World and Layered World pictures alike are committed to Fundamentalism and Funda-Mentalism.

But according to the metaphysical picture of the *Dynamic World* that lies behind the Essentially Embodied Agency Theory of mental causation, the physical world is *not* fundamentally physical, and the mental world is *not* fundamentally mental. In the Dynamic World, in fact, there are *no such things* as explanatorily or ontologically distinct physical and mental worlds, nor are there any such things as distinct explanatory or ontological levels of microphysical composition. The essential features of the Dynamic World are *action* and *mutual interaction, energy*, and *force*. Molecules, atoms, and quantum phenomena are just different ways in which different kinds of inherently active and interactive, energetic, and force-driven phenomena operate according to different sets of laws of varying scope. So there

is one and only one natural world, which is essentially a law-governed spatiotemporal totality of processes in various kinds of patterned change, motion, and evolution (with limiting cases of dispersal, entropy, permanent equilibrium, heat-death, and stasis), some of which are the intentional body movements of motile, situated, forward flowing suitably neurobiologically complex living organisms with essentially embodied conscious, intentional $minds_{lo}$.

Therefore, in sharp opposition to the static binary oppositional world picture provided by Substance Dualism, and also in equally sharp opposition to the static hierarchical upwards-dependency picture provided by Materialism, the Dynamic World picture seems best captured by the simple image of a hyperbolic spiral superimposed on a rectilinear grid:

THE DYNAMIC WORLD



Mental Facts/Biological facts/Chemical facts/Molecular, atomic, and quantum facts

Think of the rectilinear grid, somewhat like Wittgenstein's notion of "logical space" in the *Tractatus*,[17] as the totality of all possible natural facts. Then think of the hyperbolic spiral as the trajectory or unfolding of all the *actual* natural events in *actual* space and time. Some of these natural events are chemical facts but not biological facts, although all of the biological facts are also chemical facts. Some of these natural events are chemical and biological facts but not mental facts, although all of the mental facts are also biological facts and chemical facts. So some of these natural events are

[17] See Wittgenstein, *Tractatus Logico-Philosophicus*, props. 1.13, 2.013–2.0131, 2.11, 2.202, and 3.4.

mental, biological, and chemical facts, and *all* of these natural events are also molecular, atomic, and quantum facts. The natural mental events and facts occur on the outermost edge of the infinitely unfolding spiral, and thereby necessarily link together all of the other kinds of events and facts. In this way, the mental facts, biological facts, chemical facts, molecular facts, atomic facts, and quantum facts are all unevenly but still systematically distributed *throughout* the natural world-spiral.

In the Dynamic World, individual physical substances really exist, but they are themselves really nothing but differently inherently or intrinsically structured sets of inherently active and interactive, energetic, and force-driven physical events operating under causal laws—*dynamic systems*. Everything in nature is either a dynamic system itself or else a necessary proper part of some dynamic system. For example, the weather on a certain day is a dynamic system, and a certain cloud formation is a necessary proper part of it. Likewise, that cloud formation is itself a dynamic system, and a certain water droplet is a necessary proper part of it. It is also possible for the same thing to be a necessary proper part of many different dynamic systems: the water droplet is a necessary proper part of both the cloud formation and the weather system alike. Necessary proper parthood in a dynamic system means playing a certain efficacious causal role within that system, and contributing in some definite way to the system's efficacious causal powers. So the natural world is nothing but causally-empowered dynamic systems and their necessary proper parts, all the way around and all the way through.

This is not, however, to say that each dynamic system is the *same* system. On the contrary, each dynamic system has its own inherent or intrinsic structural causal-nomological profile such that it is irreducibly the individual system that it is, and not some other one. And there are irreducibly different natural kinds of dynamic systems, not to mention irreducibly different classes of dynamic systems under various shared properties. In this way the ontology of dynamic systems is *monistic*, but non-reductive. The natural world is composed of a single kind of thing, dynamic systems, out of whose dynamics emerge an infinite variety of different properties.

All of the dynamic systems exemplify fundamental molecular, atomic, and quantum physical properties that are instantianted spatiotemporally. And so according to the Dynamic World picture there are no *fundamentally mental, or essentially non-physical* entities in the natural world. No dynamic

system is fundamentally mental in that it cannot instantiate an inherent or intrinsic physical property. So too, according to the Dynamic World picture, there are no *fundamentally physical, or essentially non-mental* entities in the natural world. No dynamic system is fundamentally physical in that it *cannot* instantiate an inherent or intrinsic mental property. But at the same time, however, *many* dynamic systems are *predominantly physical* in that they *do not* instantiate inherent or intrinsic mental properties (e.g., rivers, mountains, and weather systems). Similarly, *many* dynamic systems are *predominantly mechanical* since they *do not* instantiate inherent or intrinsic biological properties (e.g. automobiles, Coke machines, and laptop computers). But *not all* dynamic systems are predominantly physical, just as *not all* dynamic systems are predominantly mechanical and unliving. Some dynamic systems not only *can* but in fact also *actually do* instantiate inherent or intrinsic biological properties but not inherent or intrinsic mental properties (e.g., plants), and some dynamic systems not only *can* but in fact also *actually do* instantiate inherent or intrinsic mental properties as well as inherent or intrinsic biological properties (e.g., animals of a suitable degree of neurobiological complexity). And there may be real borderline cases between non-living and living dynamic systems (e.g., viruses), and also between non-conscious and conscious living dynamic systems (e.g., insects). The crucial metaphysical point is that an infinite multiplicity of real non-living or mechanical, living or biological, and conscious, intentional dynamic systems *compatibly co-exist* in the dynamic natural world.

Otherwise put, and now to come back again to the three contrasting philosophical pictures for a moment, the hyperbolic spiral image of the Dynamic World picture obviously contrasts very sharply with both the binary plane image of the Bifurcated World picture and also the stratified plane image of the Layered World picture. In the Dynamic World there is at once an indissoluble holistic blending and an inevitable pluralistic scattering of quantum facts, atomic facts, molecular facts, chemical facts, facts about living organisms, facts about essentially embodied consciousness$_{lo}$ and intentionality$_{lo}$, and facts about rational minded animals or persons, over the infinitely many dynamic systems. To put a twist on Josiah Royce's pithy definition of Idealism ("the world and the heavens, and the stars are all *real*, but not so *damned* real"[18]), according to the Dynamic

[18] Royce, *The Letters of Josiah Royce,* 217.

World picture, the natural world of dynamic systems is everywhere and everywhen physical, but not always so *damned* physical. Thus the Dynamic World picture presents a *dynamic* Neutral Monism. The single kind of thing that composes the natural world is neither fundamentally mental nor fundamentally physical, but instead is inherently active and interactive, energetic, and force-driven—like the spinning Saul Bass spiral graphic in the opening titles of *Vertigo*.

The point made in the last paragraph about the holistic blending and pluralistic scattering of different types of properties over different dynamic systems requires a little more elaboration. As we noted in Section 7.1, it is important to recognize that there are at least three different possible ways in which Post-Fundamentalism can be true of the natural world:

(a) the natural world is composed of a neutral or undifferentiated kind of thing that instantiates both fundamental physical properties and fundamental mental properties but is itself neither fundamentally physical nor fundamentally mental,
(b) all fundamental physical properties, in all of their natural-world instantiations, necessarily include inherent or intrinsic connections with fundamental mental properties,

and

(c) some but not all fundamental physical properties, in some but not all of their natural-world instantiations, necessarily include inherent or intrinsic connections with fundamental mental properties.

Thesis (a) yields Neutral Monism. Thesis (b) yields Pan-Experientialism. And thesis (c) yields Emergent Experientialism. We have already implied that we should accept thesis (a), because the Dynamic World picture entails a non-reductive Neutral Monism of dynamic systems. But while Neutral Monism is consistent with both Pan-Experientalism and Emergent Experientialism alike, Pan-Experientialism and Emergent Experientialism are mutually inconsistent. So which, if either, should we accept?

Emergent Experientialism, especially when it is understood to be combined with the dynamic world picture, and thereby understood to be a *Dynamically Emergent Experientialism*, seems not only clearly theoretically preferable to Pan-Experientialism but also independently acceptable, for two reasons.

First, Pan-Experientialism implies that everything whatsoever in nature, at every time, including beer, BMWs, dust, gases, planets, rocks, stars, trees, and viruses, actually has a characteristically beer-ish, BMW-ish, dusty, gaseous, planetary, rock-ish, tree-ish, or viral sort of proto-consciousness or proto-what-it-is-like-to-be. But that seems false or at the very least highly unlikely. If, following Nagel, we agree that we cannot know what it is like to be a bat, how can we ever have *any* conception *whatsoever* of what it is like to be *a beer*? Similarly, the thesis that everything in nature at every time is actually alive—*Animism*—seems false. It is true that like Animism, Pan-Experientialism is liberally conceivable in the APA logic—as, e.g., in the living, thinking planet of Andrei Tarkovsky's brilliant 1972 sci-fi film *Solaris*—and thus it is not *logically* impossible. But surely the actual presence of either life or subjective experience in everything at every time is not *well* supported by empirical evidence.

By contrast, Dynamically Emergent Experientialism says only that everything in nature at every time is or belongs to a dynamic system of some kind. And that seems true. Furthermore it also seems true that some but not all dynamic systems at some but not all times are alive, and that some but not all living systems at some but not all times are conscious. And surely that is empirically well-supported by contemporary physics, biology, and chemistry, not to mention cognitive science. Hence Dynamically Emergent Experientialism is not only a better theory than Pan-Experientialism when they are compared pairwise, but also quite acceptable on its own merits.

Second, obviously the basic explanatory motivation for Pan-Experientialism is to solve the mind–body problem, and in particular the problem of mental causation, in a way that is closed to both dualists and materialists. But prima facie, any metaphysical solution to the mind–body problem and the problem of mental causation that *resists* postulating the actual existence of consciousness in *everything* in nature at *every* time is to be preferred to one that postulates Pan-Experientialism, since that is an excessively strong hypothesis. Using Pan-Experientialism to solve the problem of mental causation is like screwing a lightbulb into its socket by having one person hold the bulb still and another one hundred people spin the room around. All you really need, of course, is one person who knows how to make finegrained and hyper-finegrained intentional body

movements involving his arm, wrist, and hand. Therefore Dynamically Emergent Experientialism, which says only that consciousness$_{lo}$ and intentionality$_{lo}$ actually exist in some but not all dynamic systems at some but not all times, is to be preferred to Pan-Experientialism. Furthermore, Dynamically Emergent Experientialism can be combined smoothly with the Essentially Embodied Agency Theory of mental causation in order to provide a fully adequate solution to the mind–body problem in general and also the Amazingly Hard Problem of mental causation in particular. So, again, Dynamically Emergent Experientialism is not only theoretically preferable to Pan-Experientialism, but also independently acceptable.

7.3 Dynamic Systems Theory

What we argued in the last section was this: *The world is the totality of dynamic systems, not static things*. But what, more precisely, *are* dynamic systems? Here is a very brief primer of contemporary dynamic systems theory or DST.[19] DST is the mathematical theory of sets of physical elements—where each such set is perceived by us as a single entity—whose states change over time in ways that depend on their current states according to rules. The Dynamic World picture entails that dynamic systems are not merely *perceived* unities, but also *real* unities in nature. So as we interpret DST, dynamic systems are real, unified physical processes whose collective behaviors, effects, and outputs occur in some ordered pattern that can be mathematically described in relation to their present conditions.

This is not to say, however, that every dynamic system operates like two billiard balls colliding on a flat surface, like mechanical clockwork, or like a digital computer. Many dynamic systems—including the roiling movements of boiling water, traffic patterns, the weather, ecosystems,

[19] See, e.g., Haken, *Principles of Brain Functioning: A Synergetic Approach to Brain Activity, Behavior, and Cognition*; Juarrero, *Dynamics in Action*; Kelso, *Dynamic Patterns*; Port and Van Gelder (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*; Nicolis and Prigogine, *Self-Organization in Nonequilibrium Systems*; Thelen and Smith, *A Dynamic Systems Approach to the Development of Cognition and Action*; Varela, *Principles of Biological Autonomy*; and Weber and Varela, "Life After Kant: Natural Purposes and the Autopoietic Foundations of Biological Individuality."

planets, solar systems, stars, star systems, and the movements of living organisms—are *complex*. Complexity includes two essential features:

(i) being *non-equilibrium* or *far-from-equilibrium*,

and

(ii) being *non-linear*.

Being non-equilibrium or far-from-equilibrium means that a dynamic system is such that its energy sources, energy expenditures, information levels, and material constituents are not constant in value—this phenomenon is also known as "fluctuation"—due to direct exchanges of energy, information, and matter with the environment. For example, frozen water at temperatures approaching absolute zero is in thermodynamic equilibrium, and boiling water is far-from-equilibrium. On the other hand, being non-linear means that a dynamic system is such that its outputs, effects, or collective behaviors

(a) are not a mere recursive or digitally computable function of their inputs,
(b) are not a posteriori predictable from our knowledge of the system's initial conditions, which include its individual elements and facts about their past dynamic history, the currently existing relations between those elements, the currently existing relations between those elements and other things, and the current laws of nature,

and

(c) are not a priori derivable from all the facts about the system's initial conditions.

Non-linear dynamic systems are describable by non-linear functions, while linear dynamic systems are describable by linear functions. For example, the movements of colliding billiard balls on a flat surface are describable by linear functions, while the movements of billiard balls on a curved surface are describable by non-linear functions.

The most interesting dynamic systems have what is called *dissipative structure* and are *self-organizing*. The notion of being "dissipative" here means that the energy-loss or entropy of a system is absorbed and dispersed (hence "dissipated") by the systematic re-introduction of energy and matter

into the system. Thus a dissipative structure is one that maintains a non-static causal balance between the inner states of the system and its surrounding natural environment:

> With the help of this energy and matter exchange with the environment, the system maintains its inner non-equilibrium, and the non-equilibrium in turn maintains the exchange process. . . . A dissipative structure continuously renews itself and maintains a particular dynamic regime, a globally stable space-time structure.[20]

Self-organization is how a non-equilibrium, non-linear dynamic system with dissipative structure internally generates forms or patterns of order that determine its own causal powers, and in turn place constraints (''demands'' or ''needs'') on the later collective behaviors, effects, and outputs of the whole system, in order to maintain itself. Or in other words, self-organization is *natural purposiveness* or *natural teleology*. The prime example of self-organizing systems is of course living organisms, although non-living complex systems like the roiling movements of boiling water, traffic patterns, the weather, ecosystems, the Earth, solar systems, stars, and star systems are all also self-organizing in the comprehensive sense of DST.

The fact that DST is a *mathematical* theory is important. Its descriptive formalism specifically includes the following seven elements:

(1) a *state space*, which is the set of points whose coordinates completely specify the range of possible collective behaviors of the system,
(2) a *phase space*, which is the state space insofar as its points can be considered as functions of time,
(3) a *trajectory*, which is a particular path taken by the system through the state space over time, i.e., a particular temporal sequence of collective behaviors of the system,
(4) a *control parameter*, which is a constant that can be manipulated externally to the system and given different values to produce systems with varying behaviors,
(5) an *order parameter*, which is a collective variable that determines the behavior of the individual elements of the system,

[20] See Jantsch, *The Self-Organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution*; Judson, *The Eighth Day of Creation: Makers of the Revolution in Biology;* Kauffman, *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*; and Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution.*

(6) *attractors*, which are subsets or regions of the state space, specifying a certain repertoire of collective behaviors, towards which the whole system moves and in which the system temporarily or permanently lives, as time passes,

and finally

(7) the *capacity for chaos*, which is a form of non-linear, non-stochastic instability in which small changes in initial conditions can lead to large changes in the behavior of the system in computationally intractable and unpredictable ways.

Unlike other mathematical formalisms, DST essentially includes the actual or brute fact of *the passage of time* in its equations, functions, and graphs. So the value of DST as a mathematical tool is that its formalism captures patterned material change, process, and evolution over elapsed time in a finegrained, systematic, and intuitive way that cannot be captured by other formalisms. Considered purely as a mathematical theory, DST is *metaphysically neutral*. So at least in principle, the mathematics of DST could be usefully deployed by Dualism, Materialism, and of course also by a theory like ours that is neither dualist nor materialist.

But even if the *mathematics* of DST are metaphysically neutral, DST *itself* is not exhausted by its mathematical tools and is not a metaphysically neutral theory. This is because it commits itself crucially to the notion of *circular or reciprocal causality*, which is how the "local" properties of the individual material proper parts or elements of the system on the one hand, and the "global" or system-wide properties of the system considered as an overall unity on the other hand, synchronously mutually determine the causal powers and the causal efficacy of the whole system. In Section 8.2 we will analyze this circular or reciprocal causality in terms of *dynamic emergence*. The causal contribution of the local properties of the individual material proper parts of the dynamic system is what we will call *efficient material causation*, and the contribution of the global or system-wide macro-properties of the system considered as an overall unity is what we will call *structuring causation*.

Sometimes DST theorists call these two types of causation "bottom-up" causation and "top-down" causation, but this imagery is seriously misleading because it is too much in the metaphysical grip of the Layered World picture based on mereological strong supervenience. On our view,

the complementary local and global *parts* of a given dynamical system are not distinct explanatory or ontological *levels* of that system.

Another crucial metaphysical commitment of DST is its orientation towards the life sciences, especially organismic biology. The concept of *the living organism* is absolutely central to DST in particular and to the Dynamic World picture more generally. In this picture, the facts about conscious, intentional minds_{lo} are *strongly continuous* with the facts about organismic life. As Peter Godfrey-Smith puts it, according to the strong continuity view:

> Life and mind have a common abstract pattern or set of basic organizational properties. The . . . properties characteristic of mind are an enriched version of the . . . properties that are fundamental to life in general. Mind is literally life-*like*.[21]

In other words, biological life has everything that is metaphysically and naturally required for conscious, intentional minds_{lo}, but is not always organized in a suitably complex way. Conscious, intentional minds_{lo} are inherent or intrinsic structural properties of living organisms that dynamically emerge when and only when those biological systems reach a certain suitable level of complexity. Thus the strong continuity of mind and life does *not* mean that every organism has a conscious, intentional mind_{lo}, but it *does* mean that every creature with a conscious, intentional mind_{lo} is necessarily also a living organism.

Moreover, it is *not* true on the strong continuity view that biological life is somehow a form of "unconscious mind." The metaphysical connection, instead, goes precisely the other way. As Thompson aptly puts it, conscious, intentional mind_{lo} is *in* life[22]—or as we put it, conscious, intentional mind_{lo} is a specific *structural kind* of organismic life. So too organismic life is a specific structural kind of molecular, atomic, and quantum fact. In the world described by DST, conscious, intentional minds_{lo} are strongly continuous with organismic life, and in turn, organismic life is strongly continuous with molecular, atomic, and quantum thermodynamics.[23] All the basic facts in the natural world are strongly continuous with *each other*. The natural world is an ontological *spiral*, not an ontological *bifurcated plane*, and also not an ontological *hierarchy of levels*. In other words, and now emphasizing

[21] See Godfrey-Smith, *Complexity and the Function of Mind in Nature*, p. 320.
[22] See Thompson, *Mind in Life*; and also Matthews, "Consciousness and Life."
[23] See Schrödinger, *What is Life?: The Physical Aspect of the Living Cell*.

the local and global aspects of dynamic systems: In the world described by DST, conscious, intentional $mind_{lo}$ *dynamically emerges* from organismic life; in turn, organismic life *dynamically emerges* from molecular, atomic, and quantum thermodynamics; and all three domains of facts *dynamically continuously intertwine* with each other.

7.4 Strong Metaphysical A Priori Necessity

We now turn from DST to the modal metaphysics of the Dynamic World, which will require a brief foray into modal semantics and modal epistemology. The notion of non-logical or strong metaphysical a priori necessity plays a crucial role in the Dynamic World picture because it precisely specifies the kind of strong modal connection that holds between fundamental mental properties and certain fundamental physical properties in the relation of mental-physical property fusion. Since the strong modal connection between mental properties and physical properties is one of symmetric *non-logical* necessitation, then both exact property identity, which requires *logically* necessary co-extension of properties, as well as logical supervenience, which requires *logical* sufficiency relations, are ruled out from the start. This in turn entails that both *explanatory reduction* and *ontological reduction* are ruled out from the start (see Section 6.3). Non-logical or strong metaphysical a priori necessity, as we are understanding it, is inherently also *non-reductive necessity*.

At the same time, however, there is significant contemporary philosophical controversy about the very idea of non-logical or strong metaphysical a priori necessity, deriving from two different sources. First, there are significant doubts about the existence of a genuine distinction between logical or weak metaphysical necessity on the one hand, and non-logical or strong metaphysical necessity on the other. Indeed, there are three different sub-doubts here:

(1) that logical or weak metaphysical necessities cannot be replaced by "brute" non-logical or strong metaphysical necessities since these do not suffice to account for logical a priori truth,
(2) that postulating two irreducibly different kinds of necessity (a.k.a. modal dualism) is a violation of Ockham's Razor,

and

(3) that unlike logical or weak metaphysical necessity, which can be known a priori by conceivability, we have no cognitive capacity for knowing non-logical or strong metaphysical a priori necessity.[24]

Let us call these, collectively, *the Modal Monist worry*. Modal monists also have a domestic disagreement about whether logical necessity has only one or in fact two modal "dimensions," that is, two different ways of dividing up the total space of logically possible worlds.[25] According to Two-Dimensional Modal Semantics, as we mentioned in Section 1.3, one way of dividing up the space of the world is a priori, purely conceptual, and based on a certain kind of intension, sometimes called the *primary intension* or *1-intension*. (The semantic function of an intension is to map sentences, predicates, and referring terms to possible world extensions.) And the other way of dividing up the space of worlds is a posteriori, natural scientific, and based on another kind of intension, sometimes called the *secondary intension* or *2-intension*. But whatever they think about Two-Dimensionalism, modal monists still agree that logical necessity is the only *basic* type of necessity.

Second, even amongst the defenders of non-logical or strong metaphysical necessities, there are significant doubts that there can be anything but non-logical or strong metaphysical *a posteriori* necessities. The worry here is that the only intelligible kind of non-logical or strong metaphysical necessities are all identities based on empirical natural scientific knowledge—as, e.g., in Kripke's famous argument to the effect that natural scientists know a posteriori that water is necessarily identical to H_2O. Let us call this *the Scientific Essentialist Worry*.[26]

The upshot, in any case, is that we will need to respond directly to both the modal monist and scientific essentialist worries in order to show that our metaphysical theory is both intelligible and plausible.

Unfortunately for us, however, those are not even the *only* relevant doubts. Since logical necessity correlates directly with the classical notion of *analytic* necessity and since non-logical necessity correlates directly with the classical notion of *synthetic* necessity, the very idea of non-logical

[24] See, e.g., Chalmers, *The Conscious Mind*, 136–8.

[25] See, e.g., Chalmers, "The Foundations of Two-Dimensional Semantics."

[26] See, e.g., Yablo, "Concepts and Consciousness"; and Yablo, "Is Conceivability a Guide to Possibility?"

or strong metaphysical a priori necessity presupposes the classical analytic–synthetic distinction. Consequently there will also be doubts about our theory arising from the old and huge controversy, running from Kant to Quine and beyond, over the very idea of an analytic–synthetic distinction. Given the oldness and hugeness of the analytic–synthetic debate, we could not say anything adequately convincing about the analytic–synthetic distinction here without launching into a much longer discussion than the scope of this book permits. But we do need to say *something*.

So for the purposes of keeping our present discussion somewhat manageable in size and scope, we will say for the record that we do *accept* the analytic–synthetic distinction, and do thereby *reject* Quine's skepticism about it, and that one of us has argued for these claims at some length elsewhere.[27] We will also say for the record that we understand the analytic–synthetic distinction as the absolute *semantic* (as opposed to epistemic) distinction between:

(i) propositions that are true by virtue of inherent or intrinsic conceptual connections alone, as determined by the inconceivability, and therefore logical impossibility, of the denial of those propositions either in the APA logic or in some more familiar classical, extended, or deviant logic (= analytic truths),

and

(ii) propositions that are true by virtue of whatever concepts may occur in them *together with* some essentially indexical referential relations to the actual world that are grounded in the egocentrically-centered perspective of the thinker's living body in global orientable space and thermodynamically irreversible time within a mathematical structure rich enough to guarantee the elementary arithmetic of the natural numbers (i.e., Peano arithmetic) (= synthetic truths, whether necessary or contingent).

And mutatis mutandis for analytic falsehoods. The basic idea, in a nutshell, is this. Analytic truths are *purely conceptual truths* whose denials entail logical contradictions and whose truth-conditions are semantically *insensitive* to

[27] See Hanna, *Kant and the Foundations of Analytic Philosophy*, chs. 3–5.

whatever entities or kinds of things may happen to exist in different possible worlds or in the actual world. By contrast, synthetic truths are *essentially indexical actual-world-dependent truths* whose denials are logically consistent and whose truth-conditions are inherently and directly semantically *sensitive* to world-structures (e.g., global orientable space, thermodynamically irreversible time, causation, and basic mathematics), entities, and kinds of things that exist in the actual world and other possible worlds directly related to the actual world. For example, "Not every proposition is both true and false" and "Red is a color" are analytic truths, while "$7 + 5 = 12$" and "Roses are red" are synthetic truths, respectively a priori and a posteriori.

Now provisionally granting to us that version of the analytic–synthetic distinction, at least for the purposes of being able to present our argument clearly, then we can see that the very idea of strong metaphysical a priori necessity has three basic features:

(i) its being non-logical, strong, or *synthetic* necessity,
(ii) its being *metaphysical* necessity,

and

(iii) its being *a priori* necessity.

In unpacking these, we will start with the second modal feature (i.e., its being metaphysical), which should be the most familiar to contemporary philosophers in a post-Kripkean context, and then work backwards towards the first modal feature (i.e., its being non-logical, strong, or synthetic), which will be the least familiar to contemporary philosophers, by way of the third modal feature (i.e., its being a priori), which is at least somewhat familiar to contemporary philosophers.

The notion of *metaphysical* necessity means that a proposition holds in every member of a class of logically possible worlds. This is as opposed to a purely *linguistic* necessity, which concerns only our dispositions to assert certain sentences, and to make certain inferential moves in our language-using practices, in the face of all sorts of behavioral and experiential inputs to the speaker. Nowadays, post-Kripke, very few modal metaphysicians think that all necessity should be regarded as purely linguistic. Indeed, most contemporary modal metaphysicians are also *essentialists*, who think that individual things and kinds can and do have intrinsic properties and natures.

We fully accept the metaphysical and essentialist conceptions of necessity, and so do not differ from most contemporary modal metaphysicians in these respects. Furthermore, we also fully accept Fine's distinction between *essence* and *modality*, that is, a distinction between (i) a substantive, non-logical, and synthetic necessity that flows from the actual existence and specific natures of things in the world, and (ii) a purely logical and analytic necessity that holds generally for all things, without implying the actual existence or specific nature of any kind of things.[28]

The notion of *a priori* necessity means that a proposition cannot be known by empirical means alone. This does *not* mean that it can be somehow known *without* empirical inputs or sensory evidence, but rather only that any and all empirical inputs and sensory evidence actually associated with understanding that proposition *strictly underdetermine* the meaning and truth of that proposition, and also the justifiability of belief in it. It is probably correct to say that fewer contemporary modal metaphysicians accept the existence of the a priori, than accept metaphysical necessity and essentialism. But it does still seem to be true that *many* contemporary modal metaphysicians—e.g., George Bealer, David Chalmers, Frank Jackson, and of course, Kripke—do accept the existence of the a priori.[29] We are on this team as well, and so again do not differ from many other contemporary modal metaphysicians in this respect.

Finally, the notion of strong, non-logical, or synthetic necessity means the following:

> A proposition *P* is non-logically, strongly, or synthetically necessarily true if and only if
>
> (1) *P* is true by virtue of whatever concepts may occur in *P* together with some essentially indexical referential relations to the actual world, which are grounded in the egocentrically-centered perspective of the thinker's living body in global orientable space and thermodynamically irreversible time within a mathematical structure rich enough to guarantee the elementary arithmetic of the natural numbers,

[28] See Fine, "Essence and Modality."

[29] See, e.g., Boghossian and Peacocke (eds.), *New Essays on the A Priori*; and Gendler and Hawthorne (eds.), *Conceivability and Possibility*.

(2) *P* is true in all and only the logically possible worlds that have the same set of universal intrinsic structural constraints on the nature of space, time, causation, and mathematics as our actual world (whatever those actual-world constraints turn out to be), and

(3) *P* is never false in *any* logically possible world, because *P* is a truth-value gap in every logically possible world falling *outside* the restricted class of worlds that have the same set of universal intrinsic structural constraints on the nature of space, time, causation, and mathematics as our actual world.

And mutatis mutandis for non-logically, strongly, or synthetically necessary falsehoods.

In short, non-logical, strong, or synthetic necessity constrains necessary truth, and also the corresponding necessary connections of properties picked out by the concept-terms in necessary truths, to a specially semantically and metaphysically delimited or restricted space of logically possible worlds that is focused on the universal spatiotemporal, causal, and mathematical nature of our actual world, and ultimately fixed by the egocentrically-centered standpoint of the living body of the thinker in global orientable space and thermodynamically irreversible time within a mathematical structure rich enough to guarantee the elementary arithmetic of the natural numbers. In other words, according to our view non-logical, strong or synthetic necessity is *non-reductive or liberal a priori* ***natural*** *necessity*. By contrast, according to our view, logical, weak, or analytic necessity is *non-reductive or liberal a priori* ***logical*** *necessity*.

For the purposes of showing that our account is intelligible and plausible, luckily we do not have to be able to articulate and demonstrate specifically just *what* the complete set of universal intrinsic structural constraints on the specific character of space, time, causation, and mathematics in our actual natural world actually is. But suppose, e.g., that the set includes one or more of these:

(i) that actual natural world space must always be a global orientable space of *variable* curvature including some regions of almost zero curvature (hence that actual world space must be such that it is almost Euclidean in some places and more or less non-Euclidean elsewhere),

(ii) that actual natural world time must always be not only thermodynamically asymmetric or irreversible but also include real temporal *duration* or *passage*,

and

(iii) that actual natural world causation must always be *simultaneous and continuous* even if it is sometimes also *relatively sequential*.

As a matter of fact, we *do* believe that these three requirements belong to the set of universal structural constraints on the spatiotemporal and causal specific character of our actual natural world. But even if we were *wrong* about that, it is easy enough to see what *sort* of constraints they would have to be. The crucial point, in any case, is that with those characterizations in hand we can now offer some explicit responses to the Modal Monist and Scientific Essentialist Worries about non-logical or strong metaphysical a priori necessity.

The modal monist worry has three distinct sub-doubts, and so requires three sub-responses. The first sub-doubt is that logical or weak necessities cannot all be replaced by "brute" non-logical or strong necessities, since these do not suffice to account for logical a priori truth. But since we are modal *dualists* and not modal monists, we have no intention whatsoever of trying to *replace* logical necessity with non-logical or strong necessity. So the first sub-doubt is beside the point. The only leftover comment we would need to make is that non-logical or strong necessities are tendentiously misdescribed by our modal monist critics as "brute," because that seems to imply that they have, unlike logical necessities, no underlying structure. But on *our* conception of non-logical or strong necessities, they inherently reflect the universal inherent or intrinsic spatiotemporal, causal, and mathematical structure of the actual world, whatever that turns out to be—and perhaps it includes (i) to (iii) described just above. Whatever it turns out to be, this is clearly not *purely conceptual* structure, but still clearly structure *enough* to be distinctly *non-brute*.

The second sub-doubt is that our postulating two irreducibly different kinds of necessity—a.k.a. Modal Dualism—is a violation of Ockham's Razor. But Ockham's Razor says that entities are not to be multiplied *without necessity* (in one of its classical formulations it says: *entia non sunt multiplicanda praeter necessitatem*), not that entities are not to be multiplied,

period. In other words, the Razor does *not* say that a theory about *X* has to be ontologically minimalist or reductive, *or bust*. What it says, which seems entirely correct, is that a theory should allow for exactly as many kinds of entities as it takes to give the best overall explanation of *X*—no more and no less. So if, as we believe, the best overall explanation of the mind–body relation, mental causation, and intentional action requires exactly *two* irreducibly different types of necessity, no more and no less, then Ockham's Razor not only permits Modal Dualism but in fact *requires* it. For our modal monist critic to say that Modal Dualism is an a priori violation of the Razor, in advance of considering the overall explanatory value of our theory, is just to beg the question.

The third and final sub-doubt under the collective modal monist worry is that unlike logical or weak necessity, which can be known a priori by our capacity for conceivability, by contrast non-logical, strong metaphysical, or synthetic a priori necessity corresponds to no cognitive capacity possessed by us. Now since we are modal dualists, and since we have explicitly accepted the APA logic, of course we fully accept the existence of a cognitive capacity for a priori knowledge, namely via conceivability. All that we are arguing for, then, is the existence of *another* cognitive capacity for a priori knowledge that is not itself a capacity for cognition via conceivability *alone*.

Notice that it is perfectly open to us to hold that this second cognitive capacity for a priori knowledge always operates *in conjunction* with the capacity for cognition via conceivability. And in fact, that is precisely what we are suggesting: that some kinds of a priori knowledge require *both* an essentially conceptual capacity for cognition via conceivability, and *also* an essentially *non-conceptual* capacity for cognition that can operate in conjunction with conceivability to produce a priori knowledge, but introduces mental representations with semantic structure and psychological function that are categorically distinct from those of concepts.[30]

One plausible candidate for the capacity for non-conceptual a priori knowledge, with a highly respectable empirical track record in contemporary cognitive psychology, is the capacity for generating and manipulating what Philip Johnson-Laird calls *mental models*.[31] A good example of this would be to ask yourself whether your right and left hands could ever

[30] See Hanna, "Kantian Non-Conceptualism."
[31] See Johnson-Laird, *Mental Models*; and Johnson-Laird, *How We Reason*.

occupy exactly the same volume of space in our actual world. Of course you quickly generate the self-evident belief that it is impossible. Now ask yourself again, and then phenomenologically introspect how you generated that belief. You did it by forming and manipulating a mental model. Of course this particular mental model might be inadequate to the real facts, and thus our capacity for cognizing non-logical, strong metaphysical, or synthetic a priori necessary truths is *fallible*. Our substantive proposal is simply that, to the extent that *we really do sometimes have synthetic a priori knowledge*, then our cognitive capacity for knowing non-logical, strong metaphysical, or synthetic a priori necessities is none other than the capacity for *mental modelling*, operating in conjunction with the capacity for conceivability. And what we further hypothesize is that the appeal to our capacity for mental modelling, operating together with our capacity for conceivability, gives a much better overall explanation of our a priori knowledge of both *mathematics* and *the metaphysics of nature*, than does appealing to the capacity for conceivability alone.

In this connection, it is also arguable that our capacity for mentally modelling parts of the structure of time, as immediately given to us in temporal consciousness, gives us direct non-conceptual, non-platonic a priori cognitive access to the intended model of Peano arithmetic, namely the system of natural numbers,[32] and thus explains our plainly manifest synthetic a priori knowledge of the necessary fact that $7 + 5 = 12$. Correspondingly, in the case of the metaphysics of nature, it is arguable that our capacity for mentally modelling parts of the structure of time, parts of the structure of space, and parts of the structure of causation, as immediately given to us in temporal consciousness, spatial consciousness, and primitive bodily awareness, gives us direct non-conceptual, non-platonic access to the complete set of universal inherent or intrinsic structural constraints on the nature of space, time, causation, and mathematics in our actual world and all other possible worlds that share the same set of inherent or intrinsic structures, and thus explains our plainly manifest synthetic a priori knowledge of natural metaphysical truths.

The general thought here is of course a broadly Kantian one, so that mental modelling corresponds to what Kant rather unhelpfully calls the

[32] See Hanna, "Mathematics for Humans: Kant's Philosophy of Arithmetic Revisited": and Hanna, "Mathematical Truth and Knowledge Regained: A Positive Solution to Benacerraf's Dilemma."

"transcendental synthesis of the imagination," "transcendental schematism," and "construction in pure intuition," and not a Fregean one, which is based solely on "grasping" (*greifen*) concepts (*Begriffe*) that in turn seem to be platonic abstract entities.[33] But surely that does not render our proposal automatically *suspect*. In fact, historically speaking, surely that gives the mental-modelling-capacity proposal a prima facie *equal* place in the classical debate about the nature of a priori mathematical and metaphysical knowledge. For Kant is no *second-rater* in the history of a priori epistemology. Indeed, and on the contrary, he is one of the original and, as it were, founding members of a priori epistemology's *Big Three*—Plato, Descartes, and Kant.

We conclude, then, that there is every bit as good reason to hold that we have a dedicated or innate cognitive capacity for knowing non-logical, strong metaphysical, or synthetic a priori necessities via mental modelling together with conceivability, as there is to hold that we have a dedicated or innate cognitive capacity for knowing logical, weak metaphysical, or analytic a priori necessities via conceivability.

This brings us finally to the Scientific Essentialist Worry, which says that the only intelligible kind of strong metaphysical necessities are identities based on empirical natural scientific knowledge. The prime example of this is Kripke's famous argument to the effect that natural scientists know a posteriori that water is necessarily identical to H_2O. Now Kripke himself is a property dualist, and not a type-type identity theorist, in the debate about the mind–body problem.[34] And the orthodox contemporary interpretation of Kripke's necessary a posteriori necessity says that it is a dimension of logical or *weak* metaphysical necessity,[35] not non-logical or strong metaphysical necessity. Nevertheless there is also a significant minority who hold that Two-Dimensional modal semantics is false and that non-logical or strong necessity is its own unique basic kind of necessity—indeed, that non-logical or strong necessity is the *only* basic kind of necessity.[36] So the members of this significant minority are modal monists, although not modal monists about *logical necessity*. In any case,

[33] See Frege, "Concept and Object"; Frege, "Function and Concept"; Frege, "Logic [1897]"; Frege, "Thoughts"; Hanna, "How Do We Know Necessary Truths? Kant's Answer"; and Hanna, *Kant, Science, and Human Nature*, ch. 7.

[34] See note 27 above. See also Chalmers, *The Conscious Mind*, 146–9.

[35] See Chalmers, *The Conscious Mind*, ch. 2, and pp. 137 and 149.

[36] Ibid., 136–8.

a direct but non-orthodox application of Kripke's Scientific Essentialism to strong metaphysical necessities about mind–body relations would yield the striking view, recently proposed by Nagel, that mental properties are identical to certain physical properties according to strong a posteriori metaphysical necessity:

> It seems to me that post-Kripke, the most promising line of attack on the mind-body problem is to see whether any sense can be made of the idea that mental processes might be physical processes necessarily but not analytically.[37]

It is possible for Nagel to hold this ontologically reductive view consistently with his famous Gap argument, if he interprets the Gap argument as a demonstration merely of the *explanatory* irreducibility of first-person mentalistic concepts to impersonal physicalistic concepts. As we mentioned earlier in Section 6.3, the explanatory irreducibility of concepts is perfectly consistent with the ontological reducibility of the properties corresponding to those concepts.[38]

But there is a crucial problem with the very idea of non-logically or strongly metaphysically necessary a posteriori identities. How can it be shown that a proposition is known to be a non-logically or strongly metaphysically necessary truth of *identity* without directly appealing to liberal conceivability in the APA logic, which yields *a priori* conceptual knowledge only? Everyone concedes, following Kripke again, that identity propositions are such that they are necessarily true if true at all.[39] Indeed, this is a definitional feature of an identity proposition that discriminates it semantically from other kinds of propositions, and also a central tenet of Scientific Essentialism. But as Kripke himself points out, the proposition that every identity proposition is necessarily true if true at all, is an a priori conceptual truth,[40] established by—as we would put it—liberal conceivability in the APA logic. So in order to be able to *justify* my claim to know the truth of a given *identity* proposition *P*, as opposed to any other kind of proposition, I must also know that *P* is necessarily true if true at all, which is an a priori conceptual truth. But this entails that my *knowledge* of *P* is a priori conceptual knowledge, even if I *learned* and *came to*

[37] Nagel, "The Psychophysical Nexus," 134.
[38] See also note 36 above. [39] See Kripke, "Identity and Necessity."
[40] See Kripke, *Naming and Necessity*, 159.

believe P in an a posteriori way, assuming that knowledge requires justified true belief.

Analogously, even though it was through empirical means—say, by reading about it in Manfred Kuehn's excellent *Kant: A Biography*—that I originally learned and came to believe the true proposition

M: Kant is a male,

which indeed suffices for the truth of the conditional proposition

S: If Kant is a bachelor, then Kant is male,

nevertheless my *knowledge* of *S* requires that I infer *M* analytically from *S*'s antecedent, namely, the proposition

B: Kant is a bachelor,

which entails that both my knowledge of *S* and also of *M*'s following analytically from *B* are cases of a priori conceptual knowledge. Thus I can adequately justify my belief in *S* and *M* alike only by appealing to my knowledge that *S* is analytic. My knowledge of *S*, in other words, is a priori conceptual even though I originally learned and came to believe *S* through experience.

Therefore the scientific essentialist's appeal to strongly metaphysically necessary identity propositions that we learn or come to believe a posteriori does not suffice to show that our *knowledge* of them is a posteriori. Indeed, it is far more plausible to hold that we know all non-logically or strongly metaphysically necessary identity propositions a priori, even when we do learn or come to believe them a posteriori. So the scientific essentialist worry about non-logical or strong metaphysical a priori necessity fails.

We conclude, then, that it is perfectly legitimate for us to deploy the concept of non-logical or strong metaphysical a priori necessity in our metaphysics of the Dynamic World. The crucial pay-off of this conclusion, moreover, is the further fact that our solution to the Amazingly Hard Problem and Causal Exclusion Problems—the Essentially Embodied Agency theory of mental causation—thereby continues to hold up well under close critical scrutiny. For now it can be clearly seen that it is perfectly legitimate for us to claim that mental-physical property fusion is both metaphysically possible and also actual, and that the jointly sufficient mental causation of intentional

body movements is both metaphysically possible and also actual. We do this by claiming that minds$_{lo}$ are alive, and also that the irreducible mental properties and physical properties of essentially embodied intentional agents are bound together in a two-way or reciprocal relation of natural essence involving a non-logical or strong metaphysical necessity which cannot be known by empirical means alone. Or, in other words, we arguably know a priori how it is both metaphysically possible and also actually the case that animals with consciousness$_{lo}$ and intentionality$_{lo}$, under the right inner and outer conditions, can move our own bodies when we want to. The innate mental modelling ability that we use in order to cognize this strongly metaphysically necessary a priori truth, in turn, is ultimately grounded on our pre-reflectively conscious, essentially non-conceptual, primitive bodily awareness of our own egocentrically-centered, spatially oriented, and thermodynamically irreversible living organismic bodies.

In the next and final chapter we will explicitly apply the Essentially Embodied Agency theory of mental causation to the metaphysics of intentional agency.

8

The Metaphysics of Agency III: Where the Action Is

> The soul is the first actuality of a natural body which has life potentially.
>
> Aristotle[1]

> No part of an animal is either purely material or purely immaterial.
>
> Aristotle[2]

> Mind itself is a spatiotemporal pattern that molds the . . . dynamic patterns of the brain.
>
> J.A. Scott Kelso[3]

> Throw your hands in the A-yer,
> And wave them like you just don't KA-yer.
>
> Andre 3000 and Big Boi[4]

8.0 Introduction

In this chapter we complete our argument for the Essential Embodiment Theory of the mind–body relation, mental causation, and intentional action. In a wordbite, what we want to demonstrate is *that essentially embodied minds are where the action is*.

In Section 8.1, we argue that the mental–physical property fusion relation at the basis of essentially embodied minds is best understood, in a neo-Aristotelian way, as a hylomorphic—i.e., a matter/form or stuffing/structure—relation of *joint constitution*, according to which a conscious, intentional mind_{lo} is the irreducible and truly global intrinsic structure of a suitably neurobiologically complex living animal body. Now the Latin word for 'mind' or 'soul' is *anima*, and this beautifully captures the

[1] Aristotle, *De Anima*, II.1.412a22.
[2] Aristotle, *On the Parts of Animals*, I.3.643A24–26.
[3] Kelso, *Dynamic Patterns*, 288.
[4] A.k.a. OutKast, from *ATLiens* (1996).

sense in which a conscious, intentional $mind_{lo}$ is that which *animates* a suitably neurobiologically complex living organismic body. To animate something in this sense is to channel its natural forces and causal powers by providing its otherwise unstable dynamic processes and disparate moving parts with an *inherently dominating* organization or pattern. Such an organization or pattern *purposively guides* the entire system. Insofar as this truly global or inherently dominating organization or pattern gradually comes into existence and establishes a new dynamic regime for that entire system by purposively guiding it, then that living body is not merely alive, but also *has a life of its own*. So my conscious, intentional mind animates my own neurobiologically complex living organismic body by *intrinsically structuring* it in this sense. In turn, a suitably neurobiologically complex living organismic body that is animated by its truly global or inherently dominating intrinsic structure is nothing more and nothing less than a minded animal. For this reason, the conjunction of mental–physical property fusion and neo-Aristotelian hylomorphism is what we call *Mind–Body Animalism*.

In Section 8.2, we unpack the general concept of *emergence* and distinguish between three crucially different types of emergence:

(i) epistemic emergence,
(ii) supervenient emergence,

and

(iii) dynamic emergence.

We then argue that fundamental mental properties, and in particular those fundamental mental properties whose instantiations are necessary proper parts of jointly sufficient essentially mental-and-physical causes—i.e., episodes or events of pre-reflectively conscious desire-based emotion and effortless trying and its active guidance of intentional body movements—are dynamically emergent properties of suitably neurobiologically complex animals.

To anticipate very briefly, dynamic emergence is how complex thermodynamic systems come to have novel causally efficacious self-organizing truly global or inherently dominating intrinsic structures. Dynamic emergence is *natural creativity*. Star systems have it. Planetary ecosystems have it.

Weather systems have it. Oceans, lakes, rivers, and streams have it. Birds have it, bees have it, and even educated fleas have it. And minded animals, including us, have it too. So according to the Essential Embodiment Theory, not only conscious, intentional $minds_{lo}$, but also all biological life, and all thermodynamic complexity—including conscious, intentional $minds_{lo}$ insofar as they are ineluctably *in* biological life[5] and are thermodynamically complex systems—are just special cases of a basic natural creativity that is pervasive in the dynamic world.

Finally, in Section 8.3 we argue that the pre-reflectively conscious desire-based emotive activity of synchronous effortless trying and its active guidance of intentional body movements is a species of dynamically emergent *structuring causation*, which necessarily implies but is also irreducible to the *efficient material causation* of all dynamic processes. The distinction between trying-based structuring causation and efficient material causation, in turn, adequately explains the difference between an *arm-raising*, which is something I do, and an *arm-rising*, which is something that merely happens to me.

8.1 Mind-Body Animalism

According to the Essential Embodiment Thesis, every conscious, intentional $mind_{lo}$ is necessarily and completely embodied in a suitably neurobiologically complex living organism. Therefore every conscious, intentional $mind_{lo}$ is also alive. We also argued in Chapters 1, 2, and 5 that a $consciousness_{lo}$ is primarily manifest as desire-based emotion. So the conscious, intentional life of a minded animal is neither the dualist's *ghost in the machine*, nor the materialist's *machine in the machine*, but instead *the life of a desiring animal*. The life of a desiring animal, in turn, is necessarily interdependent with the fundamental physical properties of that animal. This doctrine is what we call Mind–Body Animalism. Or in other words, to run a slight variation on Aristotle's very deep thoughts in the first two epigraphs of this chapter: Minded animals are animals whose minds actualize their

[5] See Thompson, *Mind in Life*; and Sections 7.1–7.3 above.

living bodies by providing them with truly global or inherently dominating dynamic structures, and whose necessary proper parts are as much mental as they are physical.

Mind–Body Animalism can also be equivalently and more explicitly defined as the conjunction of two non-synonymous but still necessarily connected theses:

(i) the fundamental mental properties of conscious, intentional $minds_{lo}$ are non-logically or strongly metaphysically a priori necessarily reciprocally intrinsically related to corresponding fundamental physical properties in a living animal's body (the thesis of *mental-physical property fusion*),

and

(ii) the fundamental mental properties of conscious, intentional $minds_{lo}$ are irreducible truly global or inherently dominating intrinsic structures of motile, egocentrically-centered and spatially oriented, thermodynamically irreversible living organisms of a suitable degree of neurobiological complexity (the thesis of *neo-Aristotelian hylomorphism*).

We have already explicated and defended the notions of mental–physical property fusion and non-logical or strong metaphysical a priori necessity in Sections 7.1 and 7.3. So in this section we will concentrate on neo-Aristotelian hylomorphism.

If the mental–physical property fusion thesis is correct, then the mind–body relation is *not* a relation of identity between mental properties and certain physical properties. For that would entail Reductive Materialism, which we rejected for eleven different reasons in Section 6.3. Instead, we want to claim that the mind–body relation is one of neo-Aristotelian hylomorphic joint constitution. Our thesis that the mind–body relation is a species of hylomorphic joint constitution has historical precedents in Aristotle's metaphysics, and more specifically in his striking doctrines, quoted in the first two epigraphs of the chapter and which we paraphrased just above, that the soul or mind of an animal is the actualization of its living body, and that no part of an animal is either purely material or purely immaterial. These doctrines are directly reflected in our commitments to

the thesis that there is a strong continuity between conscious, intentional $minds_{lo}$ and biological life (the $minds_{lo}$-in-life thesis) and also to the thesis that animals of a suitable degree of neurobiological complexity are essentially mental-and-physical (the mental-physical property fusion thesis).

Aristotle's metaphysical hylomorphism, as found in his *Metaphysics, Physics, De Anima*, and his books on animals and other natural facts, says that all things in nature are the joint result of combining form with matter, or of combining *structure* with *stuffing*. A given form/matter combination yields a *whole*, of which the form and the matter are complementary proper parts, or mutually necessary aspects. Aristotle also thinks that a given form or structure necessarily confers three special properties on the whole to which it belongs:

(i) *activating actualization* (as opposed to the relatively inertial and potential character of the material component alone),
(ii) *essential individuation* (as opposed to the relatively indeterminate and unspecified character of the material component alone),

and

(iii) *natural purposiveness* (as opposed to the relatively non-purposive and mechanical character of the material component alone).

We emphatically endorse these classical Aristotelian ideas, and also assert that a conscious, intentional $mind_{lo}$ confers activating actualization, essential individuation, and natural purposiveness on its *entire* living organismic animal body, as defined by the causal powers of all its vital organs, vital systems, and vital processes. What makes our theory of hylomorphic joint constitution a *neo*-Aristotelian one, however, and not strictly speaking an Aristotelian one, are our three further theses that:

(iv) conscious, intentional $minds_{lo}$ are *essentially embodied* (since Aristotle explicitly commits himself to the doctrine of separable *noûs* in the *De Anima*),
(v) that the modal connection between mental and physical properties in minded animals is a *reciprocal relation of non-logical or strong metaphysical a priori necessitation* (since this is basically a Kantian idea),

and

(vi) that the conscious, intentional animals that are jointly hylomorphically constituted by mental and physical properties are essentially *dynamic systems* (since this requires the modern mathematical theory of complex systems dynamics).

In any case, the crucial elective affinity between the Essential Embodiment Theory and Aristotle's metaphysical hylomorphism lies in our shared idea of a robustly *biologically-oriented* and *teleological* metaphysics of the mind–body relation.

We also need to say more about hylomorphic joint constitution. The nowadays familiar *material constitution thesis* says that *X* is materially constituted by *Y* if and only if

(a) *X* and *Y* materially coincide, in the sense that they have completely overlapping spacetime volumes or extensions,
(b) *X* and *Y* are non-identical,

and

(c) *X* locally strongly supervenes on *Y*.[6]

(Again, for the explicit definition of strong supervenience and its various sub-species, see Section 1.1 above.) Thus, the famous statue of David Hume in Edinburgh materially coincides with the hunk of matter from which it is made, but it is nevertheless non-identical with this hunk of matter because the same statue could have been made of different stuff and the same hunk of matter could have been differently formed. What makes it specifically a relation of material *constitution* rather than just material coincidence, however, is the further thesis that the statue locally strongly supervenes on that actual hunk of matter. The idea is that when you fix all the actual fundamental physical properties of that hunk of matter, then that is sufficient to yield all the physical properties of the statue of Hume, and there cannot be a change in the statue's physical properties without a corresponding change in the physical properties of its constituting hunk of matter.

One important feature of material constitution is that the local strong supervience relation holds only for physical properties of the statue, and not

[6] See Baker, "Why Constitution is Not Identity." See also Section 1.1.

for the statue's value properties (e.g., its being beautiful or expensive) and its converse intentional properties (e.g., its being perceived or remembered by me). For these properties obviously can vary independently of the physical properties of the hunk of matter that composes the statue.

To be sure, one might argue that the statue's value properties and its intentional properties both *globally* strongly supervene on fundamental physical properties.[7] But this seems very unlikely to be true, for two reasons. First, for us value properties will always include *non-instrumentally* or *categorically normative* properties of the sort implied by our Desire-Overriding Internalism about reasons in Section 3.4. And second, for us intentional properties always include not only irreducible properties of *consciousness*$_{lo}$ (via our "Phenomenology$_{lo}$ of Intentionality$_{lo}$Thesis" or $P_{lo}I_{lo}$ Thesis—see Section 1.2) but also irreducible *logical* properties that are inherent in all conceptual content.[8] So we would want to reject the global strong supervenience of value properties and intentional properties on fundamental physical properties.

Be that as it may, however, another and even more important feature of the material constitution relation is that it *also* fails for living organisms. This is because the complex evolutionary and teleological-functional properties of a living organism are not determined by the fundamental physical properties of its composing hunk of matter alone. Instead, these biological properties *dynamically emerge* from those physical properties *together with* a full set of causal relations over time between the organism and its environment, *together with* whatever the organism itself formally or structurally contributes to the patterning and organizing of this complex thermodynamic process. In Section 8.2 we will spell out the notion of dynamic emergence in detail. Amongst the most important consequences of dynamic emergence is the fact that dynamically emergent biological properties, just like dynamically emergent mental properties, are *not* globally strongly supervenient on fundamental physical properties.

At the moment, however, the crucial thing is that we recognize that hylomorphic constitution is sharply different from material constitution precisely because hylomorphic constitution is *not* a strong supervenience relation of any sort, whether logical or nomological, whether local or

[7] This was proposed by one of the anonymous readers for OUP.
[8] See Hanna, *Rationality and Logic*, esp. chs. 1 and 7.

global, and whether it be one of asymmetric or one-way "upwards" modal dependency or of bilateral supervenience. This is because hylomorphic constitution entails not only that material or stuff-like physical properties and formal or structural properties are *reciprocally* non-logical or strong metaphysical a priori necessitated, hence ruling out *asymmetric* modal dependence, but also that these corresponding properties are *mutually inherently or intrinsically structurally* related to each other, hence ruling out a merely *extrinsic bilateral* modal dependency.

In a neo-Aristotelian form/matter composite, the form cannot be precisely the form that it is without *just that (kind of) matter*, in the sense that that matter is a set of causally efficacious physical substances which play all the same causal-dynamic roles. Likewise, the matter cannot be precisely the causally-defined matter it is without *just that (kind of) form*, in the sense that that form imposes upon that matter a necessarily spatially oriented and thermodynamically irreversible relational structure. Thus according to neo-Aristotelian hylomorphism, which implies both modal dualism and Dynamic Systems Theory, the form of any material object and its matter are just as tightly connected as Yeats's dancer and her dance. In this way, hylomorphic joint constitution is a species of *property fusion*, and thereby entails *neither* identity *nor* any sort of supervenience.

There is, moreover, something else very significant to be learned from the fact that the living organism does not locally strongly supervene on its composing hunk of fundamental physical matter. This is the fact that, since a minded animal is *also* a living organism, the hylomorphic joint constitution of that minded animal *also* cannot be metaphysically detached from its full set of environmental causal relations over time. The minded animal necessarily occurs in *its surrounding world* or (to borrow the shorter German term) what we will call its *Umwelt*. More precisely then, this full set of environmental causal relations over time for a given minded animal, which we will call its *Umwelt-relations*, is a necessary condition of the hylomorphic joint constitution of the minded animal, just as the Umwelt-relations of an apple (including, e.g., its having grown to ripeness in that particular way on that particular branch on that particular tree) are a collectively necessary condition of its joint hylomorphic constitution.

But it does not follow from this fact that conscious, intentional minds$_{lo}$ are *essentially embodied* by their Umwelt-relations. Otherwise put, it is a fallacy to think that every *causally necessary condition* of essential embodiment

is or can be literally a *part* of essential embodiment.[9] Let us call this *the Embodiment Fallacy*.

The Embodiment Fallacy *is* fallacious, for the following reasons. Suppose it were true that every causally necessary condition of essential embodiment is or can be literally a part of essential embodiment. Then conscious intentional minds_{lo} could be embodied not just in their brains and other vital organs and vital systems right out to the skin, but also well out *beyond* the skin into their local and distal causally-implicated environments, both past and present. And then my conscious, intentional mind might not be where and when my living body actually *is*, but instead be where and when my body actually is *not*—e.g., my conscious, intentional mind might be *at* the tree I am visually perceiving now, and also *at* the pub I remember visiting yesterday. But how can a mind_{lo} ever be embodied at locations and times at which its own living body does not actually exist? If it were true that every causally necessary condition of embodiment is or can be literally a part of essential embodiment, then it would follow that when I actually died and when my body is actually destroyed *my essential embodiment* might nevertheless survive in all the present Umwelt-relations of my essentially embodied mind.

But leaving aside subtle theological questions about the Resurrection of the Body, how can my essential embodiment ever survive the actual death and destruction of my living body? That seems absurd. Furthermore, if it were true that every causally necessary condition of embodiment is or could be literally a part of essential embodiment, there would then be no principled reason why my conscious, intentional mind could not be essentially embodied even well beyond its Umwelt-relations, *in the past and present natural world as a whole*, since the past and present natural world as a whole is a causally necessary condition of my Umwelt-relations. But that is the fast track to *Pan-Experientialism*, which we have rejected in Section 6.3.

Contemporary theorists committed to the Embodied Cognition program often speak of a "brain-body-world nexus" that is the locus of the embodiment of $\text{consciousness}_{lo}$ or $\text{intentionality}_{lo}$.[10] We think that this is a perfectly acceptable way of speaking, as long as it is recognized that the worldly relations of essential embodiment are not and cannot be literally

[9] See, e.g., Rockwell, *Neither Brain Nor Ghost*; and Hanna and Ivy, "Review of Rockwell's *Neither Brain Nor Ghost*."

[10] See, e.g., Clark, *Being There*; Noë, *Action in Perception*; and Rowlands, *Body Language*.

parts of essential embodiment just because they are causally necessary conditions of it. It also seems to us quite possible that the following two theses, each of which is independently quite plausible, have individually or jointly motivated Embodied Cognition theorists to commit, or at least to be on the verge of committing, the Embodiment Fallacy:

(1) *Weak Content Externalism*,[11] which says that all representational mental contents are at least partially determined or individuated by factors in the external causal or social environment of the cognizer,

and

(2) *the Weak Extended Mind Thesis*,[12] which says that at least some items in the external environment of a cognizer—say laptop computers, Blackberrys, typewriters, pocket notebooks, index cards, pens, pencils, and other cognitive or practical tools—are *ancillary* or *peripheral* causally efficacious vehicles of her mental contents.

But Weak Content Externalism and the Weak Extended Mind Thesis are clearly *logically independent* of the claim that the worldly relations of essential embodiment are literally or possibly parts of essential embodiment just because they are causally necessary conditions of it. It is entirely possible for Weak Content Externalism and the Weak Extended Mind thesis to be both true—respectively, say, for my blindsighted visual experience of a certain tree *T*, and my use of a laptop computer to record some interesting facts about *T* (say, that it is a larch) that I have temporarily forgotten—but also false that my conscious, intentional mind can ever literally be embodied *at* the larch tree *T* that I blindsightedly see or extendedly remember, or *at* the laptop computer on which I have recorded some interesting facts. Indeed, although we do think that the Essential Embodiment Thesis, Weak Content Externalism, and the Weak Extended Mind thesis are all true, nevertheless the Embodiment Fallacy remains a fallacy. For us

[11] See, e.g., McGinn, *Mental Content*; McCulloch, *The Mind and its World*; and Rowlands, *Externalism*. Strong Content Externalism says that all representational mental contents are solely and wholly determined or individuated by their external causal-environmental or social relations. This seems too strong, and for various other reasons we prefer the *conjunction* of Weak Content Externalism and also Weak Content Individualism—which says that all representational mental contents are at least partially determined or individuated by factors internal to the cognizer.

[12] See Clark and Chalmers, "The Extended Mind"; Clark, *Mindware*, ch. 8; and Rowlands, *Body Language*, ch. 3. The Strong Extended Mind Thesis says that some items in the external environment of a cognizer are the primary causally efficacious vehicles of mental content. This also seems too strong.

the conscious, intentional mind is not and *cannot* be embodied in the Umwelt-relations required by its essential embodiment. The limits of the existence of my conscious, intentional mind are the same as the limits of my living organismic body. Therefore conscious, intentional $minds_{lo}$ are embodied *only* wherever and whenever their living organismic bodies exist.

Now let us try to be even more precise about the specific nature of the relation between the mental and physical aspects of the hylomorphically jointly constituted minded animal. As we have said, according to our neo-Aristotelian hylomorphism, a conscious, intentional $mind_{lo}$ is nothing more and nothing less than *the truly global or inherently dominating intrinsic structure of the motile, egocentrically-centered and spatially oriented, thermodynamically irreversible living organismic body of a suitably neurobiologiucally complex animal.* This involves six different factors.

First, a conscious, intentional $mind_{lo}$ is a *truly global* structure, which is to say that it *inherently dominates* or *perversely drives* the entire living animal, understood to be identified and individuated by the causal powers of *all* its vital organs, vital systems, and vital processes insofar as they are working together as a single, unified dynamic system. This corresponds directly to the complete neurobiological embodiment of conscious, intentional $minds_{lo}$.

Second, a conscious intentional $mind_{lo}$ is an *intrinsic structure*, which is to say that it is a non-logically or strongly metaphysically a priori necessary, internal spatiotemporal relational property of the motile, egocentrically-centered and spatially oriented, thermodynamically irreversible suitably neurobiologically complex living organismic body of the animal.

Third, a conscious intentional $mind_{lo}$ is an intrinsic structure of a specifically *motile, egocentrically-centered and spatially oriented, thermodynamically irreversible* living organism. In short, the very nature of a conscious, intentional $mind_{lo}$ entails causal, spatial, and temporal *asymmetry*.

Fourth, a conscious, intentional $mind_{lo}$, as *the* truly global or inherently dominating intrinsic structure of the living organismic body of an animal, is essence-conferring. It thereby uniquely individuates that living animal, and makes it the very animal, within its species, that it is. A conscious, intentional $mind_{lo}$ exists when a complex dynamic system is not merely *alive* but also, as we commonly say about ourselves and other minded animals, *has a life of its own*.

Fifth, because a conscious, intentional mind$_{lo}$ is the essence-conferring and uniquely individuating truly global or inherently dominating intrinsic structure of a motile, situated, forward flowing suitably neurobiologically complex living organismic body of an animal, and because the animal is itself a non-equilibrium, non-linear, self-organizing thermodynamic system, it follows that a conscious, intentional mind$_{lo}$ confers *natural purposiveness* on the animal by molding the precise dissipative structure that must be maintained or ramified in order to preserve that animal's own life over time.

And sixth and finally, because a conscious, intentional mind$_{lo}$ confers natural purposiveness on the animal by molding the precise dissipative structure that preserves the animal's life over time, and because the animal is a complex dynamic system with *circular causality*, it follows that a conscious, intentional mind$_{lo}$ is the truly global or inherently dominating intrinsic structure that specifically molds the causal powers, causal operations, and nomological essence of the individual animal. In other words, the causal efficacy of a conscious, intentional mind$_{lo}$ is the result of *structuring causation*. Conscious, intentional minds$_{lo}$ cause covert or overt intentional body movements by specifically or hyper-specifically structuring the neurobiological processes of our own living organismic bodies according to finegrained or hyper-finegrained natural laws of intentional body movement.

Drawing on Chapters 1 to 5, we can now say that intentional agents carry out this structuring causation by means of essentially embodied pre-reflectively desire-based emotive effortless trying and its active guidance. Drawing on Chapters 6 and 7, we can also now say that this structuring causation is the same as the jointly sufficient essentially mental-and-physical efficacious causation of overt intentional body movements. And that is how I can raise my arm—and even more specifically, how I can throw it in the air, and wave it like I just don't care. Of course this is precisely the philosophical punchline of our book. But unlike other kinds of punchline, philosophical ones usually need to be repeated in order to secure their full effect, so we will come back to this same set of basic points again in Section 8.3.

Mind–Body Animalism also includes a biologically-oriented *functionalist* component. As we have mentioned before, the doctrine of Reductive Functionalism about the mental says that fundamental mental properties (whether of consciousness$_{lo}$ or of intentionality$_{lo}$) are identical to multiply realizable, logically strongly supervening, second-order physical properties

that consist in various configurations of computational or otherwise causal transitions between different first-order (i.e., fundamental) physical properties of some animal or machine.[13] Or in other words, according to Reductive Functionalism, a given mental property is nothing but the multiply realizable second-order physical property of being a first-order physical property that plays a certain computational or causal role, and it either locally logically strongly supervenes on the first-order physical properties of its causal role-players, or else regionally or globally logically strongly supervenes on some or all fundamental physical properties.

We have already rejected Reductive Functionalism about the mental for three different reasons. First, we reject the Multiple Realizability Argument—although we do in fact also accept a subtly different variation on it, the Blade Runner Argument (see Section 6.3). Second, Reductive Functionalism cannot adequately solve the problem of mental causation without implausibly denying the Physical Irreducibility of the Mental or PIM. Third and most importantly, the eleven a priori arguments against Reductive Materialism provide a collectively compelling sufficient reason to reject any version of it (see again Section 6.3).

But as George Bealer once remarked to one of us in conversation, even if Functionalism were not a correct theory of the *mind,* in some sense it could still be a correct theory of the *body*. We strongly agree with that idea, at least as far as the living organismic body is concerned, and if functional properties are allowed to include naturally purposive or teleological properties along with computational and causal-role properties. The ability of something to play a certain causally efficacious role *in* some dynamic system, or to play a certain causally efficacious role *as* a dynamic system, is just what we mean by a *causal power* of that thing. More precisely then, according to our Mind–Body Animalism, necessarily anything that has the same *naturally purposive self-organization* as a living animal body and also has all the same *causal powers* as a living animal body, is a living animal body. Or, in other words, necessarily anything that is the same kind of dynamic system as a living organismic animal body, is a living organismic animal body. Let us call this thesis *Living Body Functionalism*.

Now it is liberally conceivable (and therefore logically possible in the APA logic) that the same kind of dynamic system as a living organismic

[13] See Putnam, "The Nature of Mental States"; and Block, "Troubles with Functionalism."

animal body could be instantiated in different types of compositional stuff. So, e.g., it is logically possible in the APA logic for human bodies, owl bodies, snake bodies, etc., to be made out of an artificially created stuff, and not out of naturally created stuff. Indeed, this is directly entailed by the notion of a *replicant* in the Blade Runner Argument we spelled out in Section 6.3. If so, then at least in principle the fundamental physical properties of the very same kind of living organismic animal body can *change* as it is instantiated in different possible compositional stuffs. Let us call this modal fact *multiple embodiability*.

Living Body Functionalism and multiple embodiability together imply a refined version of our initial formulation of mental-physical property fusion in Section 7.2. This refined version relativizes the reciprocal intrinsic necessitation relation between mental and physical properties to all and only the merely compositionally different instances of a suitably neurobiologically complex living organismic animal body. That in turn allows for the multiple embodiability of that living organismic body, and correspondingly for changes in its fundamental physical properties as it is instantiated in different possible compositional stuffs. Let us call a complete class of physical instantiations *I* of a given type of living organismic animal bodies *B* (say, human bodies or cat bodies) in a single type of compositional stuff *an embodiment E*. Then the refined version of mental–physical property fusion says this:

(1) Under an embodiment *E*, an event or physical substance *X* has some fundamental mental properties *M1, M2, M3*, etc.
(2) Under the same embodiment *E*, *X* also has some non-identical or distinct fundamental physical properties *P1, P2, P3*, etc.
(3) For every *Mi* there is a one-to-one correlation with a corresponding *Pi*.
(4) The members of each 1–1 correlated *Mi–Pi* pair are necessarily co-extensive.
(5) The members of each 1–1 correlated *Mi–Pi* pair are not logically necessarily co-extensive.
(6) The members of each 1–1 correlated *Mi–Pi* pair are mutually inherent or intrinsic structural properties of *X*.
(7) *X* is a suitably complex living organism.

The thesis of Living Body Functionalism is supported by some striking empirical facts.

First, empirical facts about *successful organ transplants and artificial organs* obviously support Living Body Functionalism. The material composition of the body changes, but the same body lives on.

Second, in the case of *successful prosthesis*, by means of a pre-reflectively conscious desire-based emotive act of acceptance and appropriation, someone can literally *incorporate* a rubber hand or mechanical leg, or some other artificial body part, and add it to her own living organismic body, thereby restoring some specific causal power of her body in relation to the basic causal powers of all her vital organs and vital systems.[14] So again, the material composition of the body changes, but the same body lives on. But here it is to be remembered that we are individuating living bodies by their naturally purposive organizations and causal powers, so it does not matter whether a body part is literally *inside* the body or not. Of course it needs to be *directly attached* to the body, and to occupy part of the same egocentrically-centered location as the rest of the body, for the purposes of cognition and intentional action. But what matters is how the body part contributes to the vital operations of the *whole* body.

Third, empirical facts about *neural plasticity* also support Living Body Functionalism.[15] Neural plasticity is when there are changes in the neural structure and composition of the brain itself in response to global activities of the minded animal—or in other words, neural plasticity occurs when the intentional causal powers of the essentially embodied agent are directly reflected in the brain's *rewiring*. Indeed, as Alva Noë and Susan Hurley have pointed out, blind users of Tactile Visual Substitution Systems (TVSS)—a device worn next to the skin that translates video camera pictures into patterns of skin stimulation—begin to manifest new patterns of neuronal activity in the previously inert *visual* cortex and also report the existence of new conscious *visual* sensations.[16] If Noë and Hurley are correct, then the empirical data about TVSS not only support Living Body Functionalism,

[14] See Blakesee and Ramachandran, *Phantoms in the Brain*, chs. 2–3; Gallagher, *How the Body Shapes the Mind;* and Merleau-Ponty, *The Phenomenology of Perception*, 76–89.

[15] See, e.g., Hurley and Noë, "Neural Plasticity and Consciousness."

[16] Ibid.

but also the *structuring causation* thesis of our neo-Aristotelian hylomorphism. This is because it seems clear that the changing dynamic patterns of brain activity and the existence of new visual experiences are the direct, efficaciously causal results of conscious, intentional agents who are effortlessly or effortfully *trying* to use TVSS.

8.2 Dynamic Emergence

As we have emphasized, the fact of mental–physical property fusion does not happen everywhere or everywhen in nature. Organismic life appeared only a very short time ago in cosmological terms, and conscious, intentional $minds_{lo}$ appeared in organismic life even more recently than that. So conscious, intentional $minds_{lo}$ are relative newcomers to natural history, and therefore Mind–Body Animalism is most decidedly not a version of Pan-Experientialism. The important fact that fundamental mental properties and certain fundamental physical properties are fused (relative to an embodiment) *only* over certain time spans and *only* in organisms of a suitable level of neurobiological complexity, brings us to the concept of *emergence*. Emergence is the last basic element in our metaphysics of essentially embodied agency.

The concept of emergence has its historical source in early twentieth-century debates about scientific reductionism, and in particular the mechanism vs. vitalism controversy in the philosophy of biology.[17] For some reason, that controversy withered away—possibly because by the end of the 1950s Scientific or Reductive Naturalism was the conventional wisdom in Anglo-American philosophy.[18] In the late 1990s and early 2000s however, the concept of emergence has (as it were) re-emerged in the context of the mind–body problem and more specifically in the context of the problem of mental causation.[19] The aim of this section is to unpack the general concept of emergence, distinguish sharply between three fundamentally different types of emergence, and then to extend our metaphysics of essentially

[17] See Stephan, "Emergence: A Systematic View of its Historical Facets"; and McLaughlin, "The Rise and Fall of British Emergentism."

[18] See Hanna, *Kant, Science, and Human Nature*, 8–17.

[19] See, e.g., Beckermann et al., (eds.), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*; and O'Connor and Wong, "Emergent Properties."

embodied agency by adding to it the notion of a non-supervenient and robustly causal type of emergence that is characteristic of the truly global or inherently dominating intrinsic structural properties of all self-organizing thermodynamic systems, especially including conscious, intentional systems like us. This is what we call *dynamic* emergence.

Very simply put, the notion of emergence in general is the idea *that new properties of a dynamic system can come out of old properties of that dynamic system*. So emergence is *natural creativity*. It should be immediately noticed, however, that the contrast between the *newness* versus the *oldness* of the properties of a dynamic system is ambiguous as to whether it should be understood as the contrast between, on the one hand,

(a) the less ontologically basic properties of a dynamic system vs. the more ontologically basic properties of that system (e.g., the system's temperature vs. its mean molecular motion, or the system's being water vs. its being H_2O),

or whether on the other hand it should be understood as the contrast between:

(b) those properties of a dynamic system whose instances exist earlier in time vs. those properties of that system whose instances exist only later in time (e.g., the system's being a build-up of towering cumulus clouds vs. its later being a thunderstorm, or the system's acorn-ness vs. its later oak tree-ness).

Correspondingly, the notion of *coming out of* is ambiguous as to whether it should be understood as, on the one hand,

(a*) the simultaneous strong supervenience of various global properties of the system on the non-relational or relational properties of its local proper parts (e.g., the strong supervenience of the system's temperature at any given time on its mean molecular motion at that same time, or the strong supervenience of the system's being water on its being H_2O),

or whether on the other hand it should be understood as

(b*) the generation of various global properties of the system only over time (e.g., the growth of a thunderstorm out of a build-up of

towering cumulus clouds, or the growth of an oak tree out of an acorn).

Noting these conceptual ambiguities is extremely important, because the pair consisting of (a) and (b), and the pair consisting of (a*) and (b*), while they may seem superficially consistent with each other, are in fact *inconsistent*. This becomes clearer when we formulate the notions corresponding to each pair more explicitly. The conceptual pair consisting of (a) and (a*) is what we will call *essentially synchronic emergence*, and the conceptual pair consisting of (b) and (b*) is what we will call *essentially diachronic emergence*.

Essentially Synchronic Emergence

A dynamic system has essentially synchronically emergent properties if and only if new properties of that system come out of old properties of that system such that

(i) necessarily the new properties of that system occur at a less ontologically basic level than the old properties (e.g., the system's temperature vs. its mean molecular motion),

and

(ii) necessarily the new properties are global properties of the system that simultaneously locally strongly supervene on the non-relational or relational properties of the system's local proper parts (e.g., the local strong supervenience of the system's temperature at any given time on its mean molecular motion at that same time, or the local strong supervenience of the system's being water at any given time on its being H_2O at that same time).

Essentially Diachronic Emergence

A dynamic system has essentially diachronically emergent properties if and only if new properties of that system come out of old properties of that system such that

(i) necessarily the new properties of that system are instantiated later than its old properties and do not exist in that system at any time prior to the existence of its old properties (e.g., the system exemplifies being

a thunderstorm later than it exemplifies being a build-up of towering cumulus clouds, and can never exemplify being a thunderstorm before it has exemplified being a build-up of towering cumulus clouds; or the system exemplifies oak tree-ness later than it exemplifies acorn-ness, and can never exemplify oak-treeness before it has exemplified acorn-ness),

and

(ii) necessarily the new properties of the system are truly global or inherently dominating intrinsic structural properties of the system generated by its local proper parts together with various causal interactions between the system and its environment over thermodynamically irreversible time (e.g., the growth of a thunderstorm out of a build-up of towering cumulus clouds, or the growth of an oak tree out of an acorn).

Now a dynamic system with essentially synchronic emergent properties *upwardly determines* those emergent properties, while a dynamic system with essentially diachronic emergent properties *generates* those emergent properties. The paradigm of essentially synchronic emergence is the fact of what Kim aptly calls *mereological supervenience*,[20] whereby the global properties of a system locally strongly supervene on its compositional atoms, whereas the paradigm of essentially diachronic emergence is the fact of *natural growth*. No dynamic system can be both essentially synchronically emergent and also essentially *diachronically* emergent, precisely because the diachronic emergence of a dynamic system necessarily involves both the thermodynamically irreversible passage of time (i.e., "time's arrow," and irreversible process) and the Umwelt-relations of that system, whereas essentially synchronic emergence does *not* necessarily involve either of these. So essentially synchronic emergence is always consistent with *time-reversible* or time-symmetric physical processes (i.e., physical processes that can flow backwards in time without changing their fundamental physical properties[21]) and a narrowly *local* strong supervenience of a system's global properties on the intrinsic non-relational properties of the proper

[20] Kim, "Multiple Realizability and the Metaphysics of Reduction."

[21] See, e.g., Brading and Castellani, "Symmetry and Symmetry Breaking"; and Savitt, "Introduction to *Time's Arrows Today*."

parts of the system together with their accidental modes of combination with one another. Moreover, whereas essentially synchronic emergence presupposes *the Layered World* conception of nature (see Section 7.3), essentially diachronic emergence does *not* presuppose it, and in fact fits smoothly with the *Dynamic World* conception of nature instead. We will come back to these important points later.

Perhaps the best-known contemporary argument in favor of emergentism is the one developed by Timothy O'Connor.[22] According to O'Connor, mental properties are strongly supervenient on physical microproperties, irreducible to those physical microproperties, and causally efficacious. Like digital computational functional properties, emergent properties strongly supervene on fundamental physical properties. But unlike digital computational functional properties, emergent properties exhibit a novel causal influence that is irreducible to the microproperties on which they supervene. According to O'Connor, the causal powers of an emergent property go beyond those of the supervenience base properties. More precisely, he says that an emergent property has four individually necessary and jointly defining features:[23]

(1) An emergent property is potentially had only by objects of some complexity.
(2) An emergent property is not had by any of the object's parts.
(3) An emergent property is distinct from any "structural" property of the object.
(4) An emergent property has direct, "downward" determinative influence on the pattern of behavior involving the object's parts.

We will now briefly unpack each of these features, and then use O'Connor's theory of emergence as a comparative and contrastive segue to our theory.

First, O'Connor's idea that an emergent property is had only by complex objects is linked to the notion that an object's emergent properties strongly supervene on the properties of its parts. If we think of these parts as compositional atoms, the general idea is that higher-level properties of complex objects are necessitated by atomistic properties. O'Connor claims that this local strong supervenience relation is nomologically necessary, and also believes that this relation captures the sense in which emergent properties

[22] O'Connor, "Emergent Properties." [23] Ibid., 95.

of complex objects arise from their fundamental or first-order and thereby lowest-level physical properties. This eliminates the unattractive possibility that mental properties mysteriously float free of physical properties.

Second, an emergent property is not had by any of the object's parts. Because emergent properties arise only at a certain level of complexity, any basic part of the object cannot have an emergent property. Therefore, one must think of an emergent property as a *systemic* property.[24]

Third, an emergent property is always distinct from what O'Connor calls a "structural" or "resultant" property of the object. In O'Connor's account, structural or resultant properties of an object are those that locally strongly supervene on atomistic properties, are multiply realizable, and are strictly a matter of extrinsic relations between the object's parts. If emergent properties were this kind of structural or resultant property, their causal powers would be novel only in the sense that there are applicable higher-level laws and concepts that abstract away from the details of the atomistic properties. O'Connor wants to reject this view of emergent properties, for it would amount to a theory of emergence only in an epistemological sense. We would still be left with ontological reductionism.

Fourth, an emergent property exhibits "downward" determinative causal influence. Note that in the case of structural or resultant properties, the causal powers bestowed are simply a summation of the causal powers bestowed by the atomistic properties on which those structural properties supervene. The legs of the table cause a square imprint on the floor in virtue of their atomistic properties, without there being any novel, downward causal influence. Emergent properties, on the other hand, are supposed to exert novel causal influence that goes beyond the causal powers bestowed by the atomistic properties on which they supervene. On O'Connor's view, these properties can, in fact, influence the dynamic trajectory of underlying atoms. Note that this fourth defining feature of downward causal influence is crucial to emergentism and its potential solution to the problem of mental causation, for without it we will be left with Epiphenomenalism.

It seems clear that contemporary theories of emergence like O'Connor's are based on some questionable assumptions. The thesis of physical monism upon which these theories rely is a thesis about the nature of systems

[24] See also Stephan, "Emergentism, Irreducibility, and Downward Causation," 80.

that instantiate emergent properties. This thesis says that the bearers of emergent properties are made up only of physical parts, the rationale being that there are no supernatural components responsible for a system's having emergent properties.[25] Living organisms, moreover, consist of the same basic components that make up inanimate nature. At first glance, this thesis may seem plausible enough. However, we should look closer at its two parts. The first part of the thesis says that existing entities consist solely of physical parts, while the second part states that emergent properties are instantiated by systems consisting exclusively of inanimate fundamentally physical entities. But it seems that if the notion of the physical is interpreted in a sufficiently philosophically careful sense, we can plausibly accept the first part of the thesis without accepting the second part. As we saw in Section 7.2, to say that systems consist exclusively of physical entities is not yet to establish that these systems are *fundamentally physical*, i.e., that they are such that their fundamental physical properties necessarily exclude inherent or intrinsic relations to fundamental mental properties. By accepting that the object or system's constituent parts necessarily exclude intrinsic relations to mental properties, the emergentist has conceded too much. He has accepted not only physical monism, but in fact a *fundamentalist* physical monism. He has also accepted the *layered* picture of the world (see Section 7.3). Indeed, the Layered World picture is obviously implicit in the very idea of "downward" causation.

In turn, Kim has persuasively argued that the very idea of downward causation looks to be highly suspicious, and that ultimately the theory of emergentism faces the same worries about explanatory and supervenient causal exclusion that we discussed in Section 6.4.[26] He considers two types of downward or "reflexive" (that is, system) causation, *synchronic* and *diachronic*, and argues that neither type can do the work that theorists want it to do. According to synchronic reflexive causation, wholes have causal influence on their own micro-constituents, and downward causation is simultaneous with upward determination. But how, asks Kim, could higher-level properties causally influence and alter the physical conditions from which they arise? Kim finds it incoherent to suppose that while the

[25] Ibid., 79. [26] Kim, "Making Sense of Downward Causation," 305.

presence of *X* is entirely responsible for *Y*, *Y* somehow exerts simultaneous causal influence on *X*. For him, this implies an absurd sort of reflexive causation that is "viciously circular."[27] He believes that upward determination must take place *before* the higher-level properties can exert downward causal influence. Diachronic reflexive causation makes more sense insofar as it removes the circularity found in synchronic downward causation and establishes the needed time delay. Mental properties emerge from certain basal physical conditions over time and *then* exert downward causal influence. Kim correctly points out, however, that this leads directly to a version of the Causal Exclusion Problems that is specific to emergence. If a mental property *M* emerges from basal physical conditions *P* over time, then *P* displaces *M* as a cause of any of *M*'s putative effects. Because *M*'s emergence base *P* is nomologically sufficient for *M*, then *P* is also nomologically sufficient for any alleged effects of *M*.[28]

If the emergentist accepts the picture that Kim sketches, then there is no hope for "downward" causation. But must emergentism necessarily involve a commitment to the sort of "upward" determination that Kim envisions? More generally, must the emergentist accept *any* sort of "upwards-downwards" picture—i.e., a *levels* picture? According to Kim and the usual interpretations of emergentism, emergent properties are physically grounded and strongly supervene on fundamental microphysical properties. Properties of higher-level entities arise out of intrinsic non-relational properties and extrinsic relations that characterize their constituent parts, so that emergent properties are completely nomologically determined by items at the bottom physical level. This is why Kim speaks of emergent properties as if they obey a bottom-up, essentially synchronic supervenient determination.

However, according to the *essentially diachronic* form of emergence that we want to advocate, it is *not* true that *M*'s physical emergence base *P* is nomologically sufficient for *M*, *nor* is it true that emergent features obey the relation of bottom-up, essentially synchronic locally strongly supervenient determination. The failure of "downward" causation depends essentially on the fact of "upward" determination, and both are predicated on the Layered World picture. Thus O'Connor's conception of emergence, while it quite correctly appeals to diachronic emergence, is from our standpoint

[27] Ibid., 316. [28] Ibid., 318.

incompletely emergentist precisely because it accepts Fundamentalism and the layered world picture. He has therefore accepted a theory of emergence that is essentially *synchronic* and only *accidentally* diachronic. But at the same time he has very usefully pointed us in the right philosophical direction.

In all fairness, it should also be noted that O'Connor has changed his views since his important 1994 article. In the last few years, he has adopted a non-supervening, dynamical, and diachronic conception of emergence that is *quite* close to our notion of dynamic emergence.[29] And we are *very* happy to have him on board. In any case, on our view—which is explicitly *post*-fundamentalist, committed to the minds$_{lo}$-in-life thesis, committed to the thesis of mental–physical property fusion, committed to the idea of synchronous causation, committed to the joint sufficient cause theory of mental causation, committed to neo-Aristotelian hylomorphism, and also committed to the larger Dynamic World picture—emergent properties are *essentially* diachronic global properties of dynamic systems that are underdetermined by all the intrinsic non-relational or extrinsic relational properties of the system's local proper physical parts. This is why we call it *dynamic emergence*. In dynamic emergence, again, emergent properties are *the truly global or inherently dominating intrinsic structural properties* of dynamic systems and necessarily require both the thermodynamically irreversible passage of time and the Umwelt-relations of the system, i.e., the full set of its environmental causal relations over time. Moreover, these emergent properties have new efficacious causal powers in addition to the old causal powers of the system's constituent parts. This special set of causal characteristics, as we mentioned earlier, is the *circular causality* of self-organizing thermodynamic systems (see Section 7.3).

The non-organismic growth of thunderstorms out of build-ups of towering cumulus clouds, and the organismic growth of oak trees out of acorns, are good everyday examples of dynamic emergence. But there is also evidence for dynamic emergence in entangled quantum systems and quantum field theory.[30] In entangled quantum systems, the newly resulting compound determines the original constituents (particles) rather than the

[29] See O'Connor and Wong, "Emergent Properties," esp. section 3.2.1; and O'Connor and Wong, "The Metaphysics of Emergence."

[30] See Silberstein and McGeever, "The Search For Ontological Emergence," 187; and Stapp, *Mind, Matter, and Quantum Mechanics*.

other way around, as mereological supervenience would suggest. Quantum field theory strongly suggests that there is no ultimate level of "real" particles on which everything else is supervenient.[31] Rather, quantum fields are patterns of process over time that exist in many different types of complexity. The phenomenon of spontaneous symmetry breaking[32] likewise points to dynamic emergence. And there is also David Bohm's idea of quantum potential, a kind of pilot-wave governing the behavior of the particle.[33] Thus, our best current physical theory of the fundamental properties of matter, on at least some interpretations of it, strongly suggests that insofar as properties such as particle mass, charge, and spin are intrinsic structural, causally efficacious properties that necessarily require both the thermodynamically irreversible passage of time and their Umwelt-relations for their existence, they are dynamically emergent.[34]

We can also formulate the same points we have just been making in a slightly different way. The general concept of emergence, or natural creativity, is the conjunction of a basic positive idea and a basic negative idea. The basic positive idea behind emergence is that

> nature contains dynamic physical wholes or systems whose local proper parts relationally interact in a way that yields novel global properties of these systems.

And the basic negative idea behind emergence is that

> the novel global properties of these dynamic systems are irreducible to the intrinsic non-relational properties of their local proper parts.

Combining both the positive and negative ideas of emergence, then the general concept of emergence has three logically distinct versions:

(i) epistemic emergence,
(ii) supervenient emergence,

and

(iii) dynamic emergence.

[31] See Bickhard and Campbell, "Emergence," 331.
[32] See Brading and Castellani, "Symmetry and Symmetry Breaking."
[33] See Bruntrup, "Is Psycho-Physical Emergentism Committed to Dualism? The Causal Efficacy of Emergent Mental Properties," 147.
[34] See Silberstein, "Converging On Emergence: Consciousness, Causation and Explanation," 75–7.

These in turn can be formulated as follows:

(i) *Epistemic emergence*: Nature contains dynamic physical wholes or systems whose proper parts relationally interact in a way that yields novel global properties of these systems. These global features cannot be predicted by us from scientific knowledge of the proper parts alone.

(ii) *Supervenient emergence*: Nature contains dynamic physical wholes or systems whose local proper parts relationally interact over time in a way that yields novel global properties of these systems. These novel global properties cannot be predicted from scientific knowledge of the proper parts alone. Nevertheless these novel global properties do locally strongly supervene on the local intrinsic, non-relational fundamental physical properties of their proper parts, together with the extrinsic relational properties of those proper parts, and are not identical with any of those properties. Supervenient emergence is accidentally diachronic but essentially synchronic.

(iii) *Dynamic emergence*: Nature contains dynamic physical wholes or systems, namely self-organizing thermodynamic systems, whose local proper parts relationally interact in a way that necessarily requires the thermodynamically irreversible passage of time and the system's Umwelt-relations and thereby generates novel truly global or inherently dominating properties of these systems. These novel truly global properties cannot be predicted from scientific knowledge of the proper parts alone. Moreover these novel truly global properties do *not* locally strongly supervene on the local intrinsic non-relational fundamental physical properties of all their proper parts, together with the extrinsic relational properties of those proper parts, and are not identical with any of those properties. Nor do these novel truly global properties globally strongly supervene on fundamental physical properties. On the contrary, these novel truly global properties, together with the fundamental physical properties of the proper parts of the system, hylomorphically jointly constitute the whole system and fix its causal powers and operations. As a consequence, these novel truly global intrinsic structural properties are causally efficacious with respect to the proper parts of the system over time by molding the dynamic patterns of the efficient material

causal processes of the system. Dynamic emergence is essentially diachronic.

There are six further points to be made about dynamic emergence in relation to epistemic emergence and supervenient emergence.

First, both supervenient emergence and dynamic emergence are richer notions than epistemic emergence. Epistemic emergence, which merely captures the limits of our scientific ability to predict certain kinds of physical properties, is a necessary condition for supervenient emergence and dynamic emergence alike. But epistemic emergence is not a *sufficient* condition for either supervenient emergence or dynamic emergence. For example, our scientific knowledge of hydrogen atoms and oxygen atoms alone will not enable us to predict their chemical bonding as H_2O. But the property of being di-hydrogen oxide is neither a superveniently emergent nor a dynamically emergent property of hydrogen atoms and oxygen atoms. On the contrary, the property of being di-hydrogen oxide is literally *identical with* the relational interaction of hydrogen and oxygen.

Second, not only is it the case that nature contains some dynamic systems whose global properties are epistemically emergent but not superveniently emergent—as in the example given in the last paragraph, in which the property of being H_2O is epistemically emergent from relational interactions between hydrogen atoms and oxygen atoms, but not superveniently emergent from them. It is also the case that nature also contains some dynamic systems whose global properties are superveniently emergent but not dynamically emergent. More precisely, every strictly mechanical dynamic system—that is, every dynamical system whose global properties are strictly determined by computable functions (e.g., truth functions, primitive recursive arithmetic functions, linear functions, etc.)—is such that its global properties are superveniently emergent but not dynamically emergent.[35] For example, consider the global digital properties of the implemented programs of word processors (say, the global computational state of my Dell Latitude D810 laptop right now, as I create and save text),

[35] Since all living systems have dynamically emergent properties, it will follow that strictly mechanical systems are not alive. Nevertheless, every living system also has some sub-systems that are mechanical in form—they are isomorphic to the operations of a universal Turing Machine—although these sub-systems are not *strictly* mechanical. Thus, the human mind is *also* a digital computer, in addition to being an intrinsic proper part of self-organizing dynamic system. But the human mind is not *nothing but* a digital computer. So Strong AI, which says that $minds_{lo}$ are nothing but digital computers, is false.

the global arithmetic properties of purely aggregative totalities of particles or other material objects (say, some apples and oranges scattered on a table top) and the global kinematic properties of billiard-ball-like systems (say, the properties of gases under idealized equilibrium conditions). These are all superveniently emergent from the elements of the system together with extrinsic rules, but *not* dynamically emergent. Only self-organizing systems have dynamic emergence.

Third, obviously the difference between an intrinsic *non-relational* physical property of something and an intrinsic *relational* physical property of something is crucial to the concept of emergence. As we said in Section 1.1, by the notion of an *inherent or intrinsic property* of X we mean a *necessary, internal* property of X, namely a property whose instances are necessary proper parts of X. Therefore we do *not* use the notion of an intrinsic property in such a way as to imply that intrinsic properties are exclusively non-relational. On the contrary, we explicitly hold that some intrinsic properties are relational. More precisely, we explicitly postulate the existence of *intrinsic structural properties*, which are necessary, internal, spatiotemporal relational properties. Then any natural thing which has its spatial or temporal properties intrinsically, thereby instantiates intrinsic structural properties. Consider, e.g., all three-dimensional *orientable* material objects like my own living organismic body, which, unlike three-dimensional locally *non-orientable* objects like the Möbius Strip, necessarily include a relation to an egocentric center of directions in space.[36] To the extent that minded animals are essentially embodied, suitably neurobiologically complex living organisms, and also self-organizing thermodynamic systems, they thereby instantiate dynamically emergent intrinsic relational spatial and temporal properties. Animals with conscious, intentional minds_{lo} are necessarily egocentrically oriented in space, and our essentially embodied lives are thermodynamically irreversible with respect to time (or as the Heideggerians would say, every life like ours is a *being-towards-death*). These facts in turn provide for a metaphysically robust interpretation of Nagel's well-known and widely-accepted observation "that every subjective phenomenon is essentially connected with a single point of view."[37]

[36] See Kant, "Concerning the Ultimate Ground of the Differentiation of Directions in Space."
[37] Nagel, "What is it like to be a bat?," 167.

Fourth, whereas dynamic emergence is an essentially *timeflow-sensitive* metaphysical relation that can obtain only in asymmetrically temporally structured possible worlds and over thermodynamically irreversible durations of time, supervenient emergence is *not* an essentially timeflow-sensitive or time-asymmetric metaphysical relation, and can obtain in time-reversible or time-symmetric possible worlds. As we saw in the case of O'Connor's emergentism, supervenient emergence is essentially synchronic even when it is *accidentally* diachronic. Dynamic emergence, by sharp contrast, is *essentially* diachronic. Therefore the dynamically emergent properties of dynamic systems are inherent or intrinsic structural properties of those systems, whereas superveniently emergent properties of dynamic systems are *not* intrinsic structural properties of those systems. It follows that a superveniently emergent property of a dynamic system *cannot* also be a dynamically emergent property of that same system.

Fifth, the basic philosophical problem with supervenient emergence, like all "upwards" physical determination relations, is that it falls within the scope of the Supervenience Causal Exclusion Problem (see Section 6.4). No superveniently emergent property can ever be causally *efficacious*, but at most only causally *relevant* (see Section 6.1), because the causally efficacious properties of the physical base on which it supervenes *exclude* the "downward" causal efficacy of all its strongly supervenient higher-level properties. Otherwise put, superveniently emergent properties are necessarily *epiphenomenal*.

But by a night-and-day contrast, dynamically emergent properties are *causally efficacious* by inherently dominating and thereby molding the efficient material causal processes of the system. Otherwise the dynamic system would begin to fall apart, and either quickly or slowly move towards the dispersal of its energy and matter, and towards its own heat-death. In the dynamic world, stability and stasis are the ways the world ends, to borrow T.S. Eliot's ironically detached and apollonian trope,[38] *not with a bang but a whimper*. By sharp contrast, instability and chaos are inherently naturally creative and life-affirming—big bangs and little bangs galore, and no whimpering permitted!—and the means by which complex dynamic systems seek new temporary forms of "disorderly order" for the defiance

[38] See Eliot, "The Hollow Men."

of entropy and permanent equilibrium. Or as Dylan Thomas puts it with a truly dionysian passion:

> Do not go gentle into that good night.
> Rage, rage against the dying of the light.[39]

Sixth, finally, and most importantly, we are proposing that, along with the dynamically emergent properties of quantum entanglement, the growth of thunderstorms and traffic snarl-ups, the growth of organisms, the growth of planets and planetary ecosystems, the growth of stars and star systems, and so on, the fundamental *mental* properties of animals with conscious, intentional minds$_{lo}$ are *also* dynamically emergent properties of suitably complex dynamic systems—in our case, our suitably neurobiologically complex living organismic bodies. A conscious, intentional mind$_{lo}$ dynamically emerges from and truly globally intrinsically structures its own suitably neurobiologically complex living organismic body. It thereby efficaciously molds the dynamic patterns of that living body, and defies entropy by channeling the inherently unstable non-equilibrium, non-linear thermodynamically irreversible neurobiological processes of that body into the mental causation of intentional body movements. The explicit metaphysical description of this fact may make it seem excessively complicated. But from the first-person standpoint of a minded animal, it is brilliantly simple: all you have to do is *try*.

8.3 Arm-Raising vs. Arm-Rising: Trying as Structuring Causation

As we said in the Introduction, the Essential Embodiment Theory of the mind–body relation, mental causation, and intentional action has six central theses:

(1) **The Essential Embodiment Thesis:** Creatures with conscious, intentional minds$_{lo}$ are necessarily and completely neurobiologically embodied.
(2) **The Essentially Embodied Agency Thesis:** Basic acts (e.g., raising one's arm) are intentional body movements caused by an essentially

[39] Thomas, "Do Not Go Gentle into that Good Night."

embodied mind's synchronous *trying* to make those very movements and its active *guidance* of them.

(3) **The Emotive Causation Thesis:** Trying and its active guidance, as the cause of basic intentional actions, is primarily a pre-reflective, desire-based emotive mental activity and only derivatively a self-conscious or self-reflective, deliberative intellectual mental activity.

(4) **The Mind–Body Animalism Thesis:** The fundamental mental properties of conscious, intentional minds_{lo} are (a) non-logically or strongly metaphysically a priori necessarily reciprocally intrinsically connected to corresponding fundamental physical properties in a living animal's body (mental–physical property fusion), and (b) irreducible truly global or inherently dominating intrinsic global structures of motile, suitably neurobiologically complex, egocentrically-centered and spatially oriented, thermodynamically irreversible living organisms (neo-Aristotelian hylomorphism).

(5) **The Dynamic Emergence Thesis:** The natural world itself is neither fundamentally physical nor fundamentally mental, but is instead essentially a causal-dynamic totality of forces, processes, and patterned movements and changes in real space and real time, all of which exemplify fundamental physical properties (e.g., molecular, atomic, and quantum properties). Some but not all of those physical events *also* exemplify irreducible biological properties (e.g., being a living organism), and some but not all of those biological events *also* exemplify irreducible fundamental mental properties (e.g., $\text{consciousness}_{lo}$ or $\text{intentionality}_{lo}$). And both biological properties and fundamental mental properties are *dynamically emergent* properties of those events.

(6) **The Intentional Causation Thesis:** A mental cause is an event or process involving both $\text{consciousness}_{lo}$ and $\text{intentionality}_{lo}$, such that it is a necessary proper part of a nomologically jointly sufficient essentially mental-and-physical cause of intentional body movements. In so being, it is a dynamically emergent structuring cause of those movements. Then, under the appropriate endogenous and exogenous conditions, by virtue of synchronous trying and its active guidance, conscious, intentional essentially embodied minds_{lo} are mental causes of basic acts from their inception in neurobiological processes to their completion in overt intentional body movements.

So far, we have offered various arguments for the first five theses. These arguments, in turn, are all integral parts of a booklength argument for the sixth thesis. So in this last section, we want to complete our argument for the Intentional Causation Thesis by using the Essential Embodiment Theory to give an adequate explanation of the difference between an intentional arm-*raising* and a mere arm-*rising*—the difference between the things I do and the things that merely happen to me.

Here we need to make explicit a distinction that we already have been implicitly deploying, between two types of causation. If we consider all the causal powers and causal operations of any dynamic system, there is on the one hand

(1) the physical energy of the system together with its compositional stuff,

and on the other hand,

(2) the specific causal organization of the system.

We will call the first causal aspect (drawing on Aristotle's notions of efficient cause and material cause, but also invoking the classical early modern mechanistic account of causation) a dynamic system's *efficient material causation*, and we call the second aspect (drawing on Aristotle's notions of formal and final cause, but also on Dynamic System Theory's dynamic emergentist account of causation) its *structuring causation*.[40] Although no cause operates without both of these complementary efficient material and structuring causal aspects in play, they do each make importantly distinct contributions to causation.

In order to demonstrate this, let us now consider the following six cases:

Robby I is a robot with a mechanical arm connected to a digital clock, that has been programmed to lift its arm every day at 4:00 pm.

Robby II is a robot with a mechanical arm connected to a mercury thermometer, that has been programmed to lift its arm whenever the temperature is at 80 degrees Fahrenheit and to lower its arm whenever the temperature is other than 80 degrees Fahrenheit.

[40] See also Dretske, "Mental Events as Structuring Causes of Behavior"; and Hershfield, "Structural Causation and Psychological Explanation."

Robby III is an adult monkey that has been trained to lift its arm voluntarily whenever its trainer winks at it.

Robby IV is an adult human being who is a voluntary participant in a Wilder Penfield neurological experiment,[41] and who involuntarily lifts his arm whenever his cerebral cortex is electrically stimulated in a certain way.

Robby V is an adult human being who is a participant in a Wilder Penfield neurological experiment, and who involuntarily lifts his arm whenever his cerebral cortex is electrically stimulated in a certain way, but does not know that this is what is happening to him and falsely believes that he has intentionally lifted his arm.

Robby VI is an adult human being under ordinary conditions who self-consciously or self-reflectively lifts his arm for a reason.

In each of these cases, an arm goes up. This is the effect. Correlatively, whatever the relevant Robby does in each case is the nomologically sufficient cause of that effect. In the last example of Robby VI, an arm not only goes up but also is self-consciously or self-reflectively and deliberatively intentionally raised—hence Robby VI is a nomologically sufficient *intentional cause* of his own intentional body movement.

What more can we say about the causal powers and operations of the six Robbies? Each Robby is a dynamic system. The physical energy of each Robby-system and its compositional stuff can operate as the efficient material cause of the arm going up only if each Robby-system has the specific causal organization that it has. Otherwise the physical energy and stuff of that Robby-system would be deployed and distributed in a different way, and so would either constitute a *different* efficient material cause of a *different* effect or else be *dispersed* without any causal efficacy. Correspondingly, the specific organization of each Robby-system can operate as the structuring cause of the arm going up only if the right efficient material cause is already in place. So these two types of causation are individually necessary and jointly sufficient conditions, but not individually sufficient conditions, of causing Robby's arm to go up.

It is important to emphasize that each causal aspect authentically and efficaciously *causes* Robby's arm to go up, but only in conjunction with its

[41] See Penfield, *The Mystery of the Mind*.

complementary causal aspect. Thus the arm's going up has *dual causal aspects*, each of which can be informatively cited in a causal explanation as *the* cause of the arm-rising, but only in the context of a single overarching complete and independent causal explanation that cites both of them in relation to one another. Consider the following bit of imaginary philosophical dialogue.

Q: "What caused your arm to go up, Robby VI?"
*A*1: "Dude, I so totally, like, *raised* it."
*A*2. "The neurobiological processes of my living organismic body."

While either *A*1 or *A*2 is a perfectly correct answer to *Q*, what is crucial from an explanatory point of view is that those two answers always go together, like the coupled vibrating strings of a two-string guitar. Given the essential embodiment of conscious, intentional minds_{lo}, these are not *competing* causal explanations, precisely because they both belong to the *same* overarching complete and independent explanation. Similarly, one could cite the geometric properties of a certain material curved surface considered as concave, or cite a different set of geometric properties of the same material surface considered as convex, but only in the context of a single overarching complete and independent geometric explanation that cites both of them together.

At the same time, however, the distinctive contributions of the structuring causes and efficient material causes within each Robby-system can be conceptually isolated.

First, the structuring cause of each Robby-system exists *before* the effect happens, and in this respect is like a background condition for causation.[42] Nevertheless, unlike a background condition, the structuring cause also remains causally in force and in place *throughout* and even *after* the arm has been caused to go up, by efficaciously molding the dynamic patterns of the arm movement into an arm held motionlessly poised above Robby's head. By contrast, in causing the arm to go up, a definite amount of physical energy of the efficient material cause is irrecoverably and irreversibly *discharged* at a certain time, along with the corresponding changes (if any) in the amount, distribution, or kind (as, e.g., in chemical reactions) of compositional stuff that accompany this energy discharge. In this way, the

[42] See Mackie, *The Cement of the Universe*.

efficient material cause *expires* at the very moment of bringing its effect into existence. So the structuring cause is a temporally simultaneous and continuously extended, or *synchronous*, cause of the effect, whereas the efficient material cause is a temporally perishing and relatively sequential or *antecedent* cause of the same effect.

Second, small changes or modulations in the physical energy or compositional stuff of the efficient material cause will not normally make any significant difference to the specific character of the effect. The arm will normally still go up in just the same way. For example, various small temperature changes will not normally affect the causal powers or operations of any Robby except Robby II, which of course has been specially designed for temperature sensitivity. And various small changes in the amount, distribution, or kind of compositional stuff (for example, the presence or absence of a certain paint job or make-up job) will not normally affect any of the Robbies. By a sharp contrast, small changes or modulations in the structuring cause *will* normally make very significant differences in the specific character of the effect. For example, minute differences in re-programming or re-wiring Robby I and Robby II will lead to the arm's going up at different times, or its going up in very different ways (e.g., quickly or slowly, jerkily or smoothly, by describing an arc or a spiral, etc.), or even its not going up at all. The same thing is true for minute differences in Robby III's training or commands, and also for minute differences in how Wilder Penfield sets up the neurological experiment in the cases of Robby IV and Robby V.

Most importantly of all, the same thing (i.e. the fact that small changes or modulations in the structuring cause *will* normally make any significant differences to the specific character of the effect) remains manifestly true for minute differences in Robby VI's *conscious intentionality*$_{lo}$, that is, his pre-reflectively conscious desire-based emotive effortless *trying* to raise his arm and its active guidance of that very body movement (see Chapters 3–5). More generally, the structuring cause of each Robby *controls* the operations of its complementary efficient material cause in either a *finegrained* or *hyper-finegrained* way, which is to say that for a given unstructured quantity of physical energy and (since physical energy is a function of mass) a correspondingly unstructured given hunk of compositional stuff at a time, there is not only more than one possible way for it to move its body

(finegrainedness), but also a further plurality of agent-relative differences for every finegrained possible body movement (hyper-finegrainedness). And there is clearly also a *direct proportionality* between the complexity of a dynamic system and the level of finegrainedness of control in its structuring cause, such that the more complex and sophisticated the dynamic system is, the more highly finegrained the control of the structuring cause is.

Indeed, Robby VI reaches a maximum degree of highly finegrained structuring control in virtue of his being able, under favorable internal and external natural conditions, to have self-conscious deliberative intentions that precisely match the *hyper-finegrained* character of the *affective frames* of his practical goal focusing (see Section 5.3). For example, an arm-raising that conveys a warm greeting to a loved one is sharply and hyper-finegrainedly different from an arm-raising that conveys a Nazi salute, and the causal-dynamic profiles of the corresponding arm-risings will differ accordingly.

Therefore, we are committed to the thesis that pre-reflectively conscious desire-based emotive effortless trying and its active guidance is the structuring cause of intentional body movements, and that this mental structuring cause operates only in reciprocally necessary conjunction with the physical energy and compositional stuff of the neurobiological process that results in those very movements, and which constitute their efficient material cause.

It should also be obvious by now that there is no question of any theoretical confusion between the causal contributions of the fundamental mental properties of the intentional agent on the one hand, and the causal contributions of the fundamental physical properties of his living organismic body on the other hand. Robby VI's conscious, intentional mind_{lo} does not *push* anything around; nor is his mind_{lo} a ghostly *material substance*; nor does his mind_{lo} *inject* more physical energy into the Robby-system; nor does his mind_{lo} *extract* any physical energy from the Robby-system. What Robby VI's conscious, intentional mind_{lo} does is to synchronously induce a new specific organization in the efficient material causal processes that materially constitute Robby VI's living body. Robby VI's conscious, intentional mind_{lo} thereby brings about physical changes in his own covert (i.e., neurobiological) and overt (i.e., behavioral) body movements by inducing minute and hyper-finegrained changes in the specific organization of his thermodynamic constitution. In this perfectly definite sense, Robby VI

causes his own covert and overt intentional body movements, and in particular his arm-raising, by synchronously consciously and intentionally *structuring* the neurobiological and behavioral processes of his own living body by means of effortless trying and its active guidance.

Our use of the verb 'induce' in the last paragraph should be taken at face value. To induce the occurrence of an event *Y* is to provide the conditions that are nomologically sufficient for *Y*'s occurrence, by doing something *X* that turns out, in context, to be token-identical to the cause of *Y*'s occurrence. For example, one can induce a sneeze by presenting pepper to one's nostrils, and *this* event *also* causes the sneeze. So it cannot be stressed too much that the dynamic structuring cause of Robby VI's covert and overt intentional body movements is *token identical* with Robby VI's synchronous pre-reflectively conscious desire-based emotive effortless trying to raise his arm and its active guidance of the intentional body movements that make up that performance. Again, Robby VI's trying to raise his arm *just is* the same event-token as the event-token that is the organizational specification of the physical energy and compositional stuff of the efficient material cause of his arm's going up. Or yet again, Robby VI's trying to raise his arm just is the *role-player* of the structuring causal role in the dual aspect causation of his arm's going up.

The literal or token identity here is not (as in the Layered World picture) a ''downwards'' token identity of reduction, whereby mental properties are type-identical to certain physical properties, because the property of Robby VI's trying to raise his arm is a dynamically emergent property of Robby VI's living organismic body. Hence the literal or token identity here between the event of Robby VI's trying to raise his arm and the event of a self-organizing structuring of a certain complex neurobiological dynamic system is instead (as in the Dynamic World picture) a ''looping'' or ''spiralling'' token identity of non-reduction, whereby the fundamental mental property of Robby VI's desire-based emotive trying and its active guidance is a dynamically emergent property that is a necessary proper part of a jointly sufficient essentially mental-and-physical cause of Robby VI's arm's going up. That further implies

(a) that this fundamental mental property stands in a relation of mental-physical property fusion to some corresponding fundamental physical property of Robby VI's living body,

(b) that this fundamental mental property is the truly global or inherently dominating intrinsic structure of Robby VI's suitably neurobiologically complex living organismic animal body,

and also

(c) that the relation that this fundamental mental property bears to a corresponding fundamental physical property of Robby VI's living organismic animal body is identical to the hylomorphic joint constitution of Robby VI as a particular dynamic system.

If all this is correct, then it establishes the Intentional Causation Thesis.

Now back to the six Robbies. Let us consider them specifically in relation to one another for the purposes of philosophical inspection, comparison, and contrast.

It should be immediately clear that Robby VI's intentional causation by virtue of his trying to raise his arm is nothing but a *species* of the structuring causation that is at work in every dynamic system whatsoever. To be sure, Robby VI is interestingly different from the other Robbies. Unlike Robby I and Robby II, Robby VI is alive and has a conscious, intentional mind_{lo}. Unlike Robby III, Robby VI is capable of self-conscious or self-reflective deliberative action. Unlike Robby IV and Robby V, Robby VI is free to move his own arm when he wants to. And unlike Robby V, Robby VI is under no illusions about his own causal powers and operations. So unlike all of the other Robbies, Robby VI self-consciously or self-reflectively and deliberatively intentionally raises his arm. But otherwise considered, all six Robbies are alike *structuring causes* of their arm-risings, and therefore all the same basic causal principles are applied in each dynamic process.

Some other comparisons and contrasts between the Robbies are also illuminating.

Robby I and Robby II are both *mechanistic* dynamic systems, while Robbies III through VI are all *organismic* dynamic systems. Among other things, this means that while the global properties of Robby I and Robby II are both only *superveniently emergent* properties, the global properties of Robbies III through VI are all *dynamically emergent* properties. Mechanistic systems can be either determined (like Robby I, since its arm's going up depends on the deterministic digital computation of time) or indeterministic (like Robby II, since its arm's going up depends on temperature facts, and the

occurrence of a given temperature at a given time is arguably probabilistic or statistical, not determined). By contrast, organismic complex dynamic systems are *neither* completely determined *nor* completely indeterministic, but instead are *naturally self-created consistently with all the roughgrained general deterministic or probabilistic laws there actually are*. So unlike Robbies I and II, Robbies III through VI are all *natural causal singularities* (see Section 6.1).

Robbies III through VI are also all conscious, intentional dynamic systems. Non-mechanism does not however automatically entail freedom of the will, construed minimally as a person's ability to choose or do something without preventative constraint and without inner or outer compulsion (negative freedom), together with her ability to choose or do what the agent wants (positive freedom), which entails causal or moral responsibility.[43] This is shown by the fact that a heliotropistic sunflower is an organismic dynamic system that moves, but obviously it is not a person and therefore is incapable of free will. Only Robbies III, IV, V, and VI are even capable of free will. Robby IV has freedom of the will, but not freedom of *action*, since he cannot prevent his arm rising even if he wants to. He may even feel helplessly violated by the experimenter. Indeed, the actual subjects in the original Penfield experiments reported feeling as if their movements had been "pulled out of them."[44] Furthermore, Robby V lacks even freedom of *will*, since he wrongly thinks that his arm rising is the result of his own choice. By virtue of the manipulative machinations of the experimenter, Robby V only *believes* that he chooses to raise his own arm.

With these points in place, we can now quite easily liberally conceive (in the APA logic) of a variant on Robby IV, call it *Robby IV**, that captures the Dr Strangelove scenario:

> **Robby IV*** is an adult human being who occasionally experiences seizures that cause him to lift his arm involuntarily.

Dr Strangelove's (and thus Robby IV*'s) phenomenology, presumably, is that of someone who feels helplessly violated, and feels that his arm movement is being pulled out of him. This might take the reflexive form of feeling victimized by another part of himself, or it might take the form of feeling manipulatively victimized by some other agency—perhaps

[43] See, e.g., Kane, *A Contemporary Introduction to Free Will*.
[44] See Penfield, *The Mystery of the Mind*.

a demon or evil super-scientist—as in the pathological schizophrenic phenomenon of intrusive commands and voices. This latter possibility in turn suggests that (in the APA logic) we can also quite easily liberally conceive of a variant on Robby V, call it *Robby V**, that captures the basic scenario of John Frankenheimer's chilling and also mordantly funny 1962 paranoid thriller, the *Manchurian Candidate*, in which a normal military sharpshooter is brainwashed by an evil cognitive super-scientist into becoming a robotic assassin whenever this state is triggered by a chillingly ordinary protocol: "Raymond, why don't you play a little solitaire?," followed by the eventual presentation of a Red Queen card:

> **Robby V*** is an adult human being who has been kidnapped by an evil cognitive super-scientist, brainwashed, and then made to involuntarily lift his arm and shoot people whenever commanded to by the evil super-scientist, but does not know that this is what is happening to him, and falsely believes that he has intentionally lifted his arm and is guilty of these terrible crimes.[45]

Here the unfortunate Robby V* lacks free will. It is also intuitive that Robby V* lacks any moral responsibility for his crimes even though, tragically, he *falsely believes* he is guilty of them.

Robby III and Robby VI are also particularly interesting cases, because their arm risings are both voluntary and *responsive to reasons* (see Sections 3.2 to 3.4), albeit in quite different ways. Robby III responds "impulsively," and merely by dint of habit and training (see Section 2.4), to the practical reasons of *its trainer*, while Robby VI responds self-consciously or self-reflectively and deliberatively to his *own* recognition and adoption of a reason for action. Once we see that Robby III and Robby VI are both reasons-responsive, however, these cases point up the further possibility of two variants, call them *Robby III** and *Robby VI**, that are both engaged

[45] The actual *Manchurian Candidate* scenario is slightly more complicated than this. Part of the brainwashing protocol is that Raymond the sharpshooter is commanded to instantly forget the terrible things he is made to do when he is in "assassin mode." But in fact, when Raymond kills his newlywed wife and father-in-law, he partially breaks free of his brainwashing, and realizes what he has done. Obviously he is not morally responsible, and he knows this self-reflectively. But, even more tragically, he still cannot help pre-reflectively consciously *feeling* responsible, and ultimately commits suicide.

in *pre-reflective or spontaneous impulsive self-expressive* action, and not in reasons-responsive action:

> **Robby III*** is an adult monkey under normal conditions who feels good and suddenly throws his arm in the air, without any instrumental or other reason for doing so.
> **Robby IV*** is an adult human being under ordinary conditions who feels good and suddenly throws his arm in the air while freestyle hip-hop dancing, without any instrumental or other reason for doing so.

These actions are not at all senseless. In fact, they *make complete sense* in the context of the conscious, intentional animal's unique ongoing pre-reflectively conscious desire-based emotive life. But unless one is prepared to say that all and only intentional actions that *make complete sense* in the context of an intentional agent's unique ongoing desire-based emotive life are "reasons-responsive"[46]—which presumably would water down the very ideas of "having a reason" and "adopting a reason" to the point of there being no difference between intentional actions with reasons and without reasons—then these are intentional actions *without* reasons, in the sense that they lack self-conscious or self-reflective instrumental reasons. It is true that these intentional actions do have *internal* reasons, based on pre-reflectively conscious emotive desires. But in these cases the agent is not conscious *of* those reasons. Hence the actions of Robby III* and Robby VI*, although *reason-less* in the sense they lack self-conscious or self-reflective instrumental reasons, are nevertheless authentic intentional *arm-raisings*, and must be metaphysically explained in the same way as above. More generally, the metaphysical explanation of arm-raising offered above, when generalized to any intentional body movement, is an adequate explanation of *all* intentional action.

We should also mention one other factor that is implicit in the causal powers and operations of all of the Robbies, including Robby III* and Robby VI*, and also in the causal powers and operations of any dynamic system whatsoever. And this is the very general *ceteris paribus* assumption that *other things remain equal*. In other words, for a dynamic system to

[46] See, e.g., Arpaly, *Merit, Meaning, and Human Bondage*, ch. 2.

have a certain set of causal powers and operations, and for its structuring cause to fuse with its corresponding efficient material cause, *the world must contingently cooperate*. Inner and outer background conditions must all be appropriate for causation. If worldly conditions are uncooperative, all bets are off, and the causal powers of that dynamic system will either fail to operate, break down in the middle of its operations, or encounter complete disaster—say, by suddenly exploding, like a character in *Monty Python's Flying Circus*. (Mrs Premise might say to Mrs Conclusion, when Jean-Paul Sartre suddenly explodes: ''That's strange.'' And Mrs Conclusion might say: ''No it's not. People are exploding *every* day''). At the level of intentional action, this of course implies that it is *always* possible for a given trying to *fail*, through no fault of the intentional agent and due to unsupportive worldly conditions, at *any* point prior to the completed performance of the intentional act. This is a sad fact.

But looked at more optimistically and realistically, conditions of failure also necessarily imply corresponding conditions of success, and it seems entirely true (if generally unnoticed, because it is not very dramatic or exciting) that most of the time and for the most part, other things *are* equal. Since the purposive goal and causal effect of every basic intentional action is an intentional body movement, and since in fact intentional agents are successfully doing this all the time as a normal part of their lives, it follows that successful tryings are the *norm*. (It would *not* be absurd and hilarious, but rather only weirdly trivially true, if Mrs Premise says, when Jean-Paul suddenly raises his arm to wave to Simone: ''That's strange,'' and Mrs Conclusion says: ''No it's not. People are raising their arms *every* day.'') So this is a happy fact that is substantially bigger than the more dramatic sad fact.

And happy facts always bear repeating. Therefore we will end the book by providing, as a brief summary review of our Essential Embodiment Theory of the mind–body relation, mental causation, and intentional action, a step-by-step metaphysical analysis of a successful arm-raising, which is something that an intentional agent *does* and is truly *up to her*. We will leave a correspondingly detailed analysis of the sharply contrastive case of Dr Strangelove's spasmodic mere arm-rising, which is something that

merely happens to him—even if his overt body movement just happens to be observationally indiscriminable from the overt intentional body movement of the successful arm-raising—as a task left to the reader.

Consider now the not-so-very-strange case of *Elizabeth*, a creature with a conscious, intentional $mind_{lo}$, who is also a rational 23 year-old female human animal, under *ceteris paribus* conditions.

Elizabeth's conscious, intentional $mind_{lo}$ is essentially embodied, and thereby it is necessarily and completely neurobiologically embodied, right out to the skin. She is fully alive and fully awake. Her $consciousness_{lo}$ is primarily manifest as desire-based emotion. Her fundamental mental properties are fused with her fundamental physical properties and she is a non-equilibrium, non-linear, self-organizing thermodynamic system. Her mind is the set of dynamically emergent truly global or inherently dominating intrinsic structural properties of her living organismic animal body, which, together with her fundamental physical properties, hylomorphically jointly constitutes her. Since all dynamic systems engage in circular causality, and thereby have both structuring causation and efficient causation, it follows that her mind is the dynamically emergent structuring cause of whatever is efficiently materially caused by her living animal body.

At time t_1 Elizabeth sees a half-filled glass (of milk, of course) in front of her and forms the self-conscious deliberative intention to pick it up and drink from it. This is the goal focus of her current affective frame. The act of picking up the glass and drinking from it would then be a complex non-basic intentional act carried out by means of a certain arm movement, which would in turn constitute a single basic intentional act.

Elizabeth is also wearing her iPod Nano and listening to classic hip-hop (more specifically, ''Throw Your Hands in the Air,'' downloaded from the most excellent classic 1996 CD *ATLiens*, by OutKast). At time t_2, slightly later than time t_1, Elizabeth suddenly has a pre-reflectively conscious effective first-order desire to *throw* her right arm above her head, and thereby begins e*ffortlessly trying* to throw her right arm in the air. She forms no self-conscious or self-reflective deliberative intention to raise her arm, and indeed has no instrumental reason for raising her arm. But she feels good. In response to the music she is listening to, she simply suddenly effectively desires to raise her arm, and the goal focus of her current affective frame correspondingly shifts from picking up her glass to throwing her arm in the air. If the act successfully comes off, it will be a pre-reflective or spontaneous, impulsive self-expressive intentional act.

At the same time, her brain and other vital systems are in a certain complex dynamic neurobiological state. Because Elizabeth's fundamental physical properties are in a relation of mental-physical property fusion with her fundamental mental properties, and because she is a self-organizing thermodynamic system, her neurobiological state mirrors her conscious intentional state, and therefore her neurobiological processes are in an unstable dynamic transition between instantiating a causal power to produce a reaching and grasping movement, and instantiating a causal power to produce an impulsive self-expressive arm-raising movement. In the jargon of Dynamic Systems Theory, this unstable dynamic transition is called a "bifurcation."

At time t_3, which is later than time t_2 by at the very least 550 milliseconds (see Section 4.3), Elizabeth's effortless trying begins to actively guide her overt body movements. This effortless trying has already been synchronous and efficaciously causally engaged with various covert neurobiological processes in her vital organs and vital systems, including of course her brain, for at least 550 milliseconds. But by now the dynamic instability has been stabilized: the dynamic bifurcation has happened, and Elizabeth's neurobiological and behavioral processes alike have now entered a new complex dynamic regime with a new global dynamic pattern. She actively guides these processes just by continuing to instantiate the fundamental mental property of effortlessly trying to raise her arm and by keeping this goal within the hyper-finegrained focus of her affective frame. The affective frame of her pre-reflectively conscious desire-based emotive intention to raise her arm guarantees that her conscious state is as hyper-finegrained as her goal focus, and as a consequence, the property of effortlessly trying to throw her arm in the air is just as richly structured as it needs to be in order to provide a structuring cause of her arm movement.

At time t_4, which is later than t_3 by at the very least 350–400 milliseconds (again, see Section 4.3), Elizabeth's arm is in motion and going up as a necessary result of her effortless trying and active guiding, together with the neurobiological processes whose dynamic patterns she has efficaciously molded to the requisite level of hyper-finegrainedness conforming to her goal focus on throwing her arm in the air, by instantiating that very mental property. As we have seen, she has been synchronously controlling the efficient material causation of her motile, situated, forward flowing suitably neurobiologically complex living organismic body, so that her effortless trying has been the structuring cause of both her covert and overt intentional body movements, since t_3. But although she is now, at t_4, still trying to throw her arm in the air and still affectively framing her practical goal, and still controlling her covert and overt intentional body movements by inherently dominating and thereby molding the dynamic patterns of her neurobiological

processes into a certain hyper-finegrained truly global or inherently dominating dissipative structure, her synchronous guidance at this moment is monitoring rather than active. Thus she is pre-reflectively relatively mentally relaxed as her arm moves quickly, smoothly, and gracefully towards its intended position above her head. It is also possible now for Elizabeth to become *self-consciously or self-reflectively aware* that she has formed the intention to throw her arm in the air rather than the intention to reach for and pick up the glass.

Finally at time t_5, which is slightly later than t_4, Elizabeth's arm glides into its intended position above her head, which then becomes an overtly motionless body orientation with her arm poised for its next movement, and she stops trying to throw it in the air. Her pre-reflectively conscious desire-based emotive intentions and her neurobiological processes then begin to enter another dynamic bifurcation—perhaps towards the causal power to produce a waving movement as if she just doesn't care. In any case, and in precisely the way we have described, Elizabeth's arm has gone up in a characteristically Elizabeth-like way, and the pre-reflective or spontaneous, impulsive self-expressive intentional act of raising her arm is successfully complete.

And so is our collaborative, rational self-conscious and self-reflective deliberative intentional act of writing this book. We now rest our case, and our arms.

This page intentionally left blank

Bibliography

Allen, C., "Animal Pain," *Noûs* 38 (2004): 617–43.

Allen, C., and Bekoff, M., *Species of Mind*. Cambridge: MIT Press, 1997.

Anderson, M.L., "Embodied Cognition: A Field Guide." *Artificial Intelligence* 149 (2003): 91–130.

Anderson, P.B., et al. (eds.), *Downward Causation*. Denmark: Aarhus University Press, 2000.

Anscombe, E., *Intention*. Ithaca, NY: Cornell University Press, 1957.

Arbib, M.A., "From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics," *Behavioral and Brain Sciences* 28 (2005): 105–24.

Arbib, M.A., and Rizzolatti, G., "Neural Expectations: A Possible Evolutionary Path from Manual Skills to Language." *Communication and Cognition* 29 (1997): 393–423.

Aristotle., *De Anima*. Trans. J. A. Smith. In Aristotle, *The Collected Works of Aristotle*, 535–603.

—— *The Parts of Animals*. Trans. W. Ogle. In Aristotle, *The Collected Works of Aristotle*, 643–61.

—— *The Collected Works of Aristotle*. New York: Random House, 1941.

Arpaly, N., *Merit, Meaning, and Human Bondage*. Princeton, NJ: Princeton University Press, 2006.

Audi, R., *Action, Intention, and Reason*. Ithaca: Cornell University Press, 1993.

Austen, J., *Pride and Prejudice*, in J. Austen, *The Complete Novels of Jane Austen*. Harmondsworth: Penguin, 1983, 223–445.

Austin, J. L., "A Plea for Excuses," in J. L. Austin, *Philosophical Papers*. 3rd edn. Oxford: Oxford University Press, 1979, 175–204.

Baier, A., "What Emotions are About," *Philosophical Perspectives* 4 (1990): 1–29.

Baker, L., "Why Constitution is Not Identity," *Journal of Philosophy* 94 (1997): 599–621.

Bealer, G., "Mental Properties," *Journal of Philosophy* 91 (1994): 185–208.

Beckermann, A., et al., (eds.), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin: De Gruyter, 1992.

Bergson, H., *An Introduction to Metaphysics*. Trans. T. E. Hulme. Indianapolis, IN: Bobbs-Merill, 1955.

Bermúdez, J. "Nonconceptual Self-Consciousness and Cognitive Science," *Synthèse* 129 (2001): 129–49.

——*The Paradox of Self-Consciousness*. Cambridge: MIT Press, 1998.

Bickhard, M., and Campbell, D., "Emergence," in Anderson, et al. (eds.), *Downward Causation*, 322–48.

Blackburn, S., *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Clarendon Press, 1998.

Blakesee, S., and Ramachandran, V.S., *Phantoms in the Brain*. New York: William Morrow, 1998.

Block, N., "Concepts of Consciousness," in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 206–18.

——"The Harder Problem of Consciousness," *Journal of Philosophy* 99 (2002): 391–425.

——"On a Confusion about a Function of Consciousness," in Block, Flanagan, and Güzeldere (eds.), *The Nature of Consciousness*, 375–415.

——"Paradox and Cross Purposes in Recent Work on Consciousness," *Cognition* 79 (2001): 197–219.

——"Troubles with Functionalism," in Block (ed.), *Readings in the Philosophy of Psychology*. vol. i, 268–305.

——(ed.), *Readings in the Philosophy of Psychology*. 2 vols. Cambridge: Harvard University Press, 1980.

——"What is Functionalism?," in Block (ed.), *Readings in the Philosophy of Psychology*. vol. i, 171–84.

——Flanagan, O., and Güzeldere, G., (eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press, 1998.

Boghossian, P., and Peacocke, C., (eds.), *New Essays on the A Priori*. Oxford: Clarendon Press, 2000.

Botterill, G., and Carruthers, P., *The Philosophy of Psychology*. Cambridge: Cambridge University Press, 1999.

Braddon-Mitchell, D., and Jackson, F., *Philosophy of Mind and Cognition: An Introduction*. 2nd edn. Oxford: Blackwell, 2007.

Brading, K., and Castellani, E., "Symmetry and Symmetry Breaking," *The Stanford Encyclopedia of Philosophy (Winter 2004 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2004/entries/symmetry-breaking/>.

Brand, M., and Walton, D. (eds.), *Action Theory*. Dordrecht: D. Reidel, 1973.

Brasil-Neto, J., et al., "Focal Transcranial Magnetic Stimulation and Response Bias in a Forced-Choice Task," *Journal of Neurology, Neurosurgery, and Psychiatry* 53 (1992): 964–6.

Bratman, M., "Two Faces of Intention," in Mele (ed.), *The Philosophy of Action*, 178–203.

Brentano, F., *Psychology from an Empirical Standpoint*. Trans. A.C. Rancurello, D.B. Terrell, and L. McAlister. London: Routledge, 1995.

British Parliamentary Office of Science and Technology Notes 94 (1997). URL = <http://www.parliament.uk/post/pn094.pdf>.

Brown, D., *Descartes and the Passionate Mind*. Cambridge: Cambridge University Press, 2006.

Bruntrup, G., "Is Psycho-Physical Emergentism Committed to Dualism? The Causal Efficacy of Emergent Mental Properties," *Erkenntnis* 48 (1998): 133–51.

Bub, J., "Quantum Entanglement and Information," *The Stanford Encyclopedia of Philosophy (Spring 2006 Edition)*, E. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2006/entries/qt-entangle/>.

Campbell, J., *Past, Space, and Self*. Cambridge: MIT Press, 1994.

——*Reference and Consciousness*. Oxford: Oxford University Press, 2002.

Chalmers, D., "The Components of Content," in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 608–33.

——*The Conscious Mind*. New York: Oxford University Press, 1996.

——"Consciousness and its Place in Nature," in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 247–72.

——"Does Conceivability Entail Possibility?," in Gendler and Hawthorne (eds.), *Conceivability and Possibility*, 145–200.

——"The Foundations of Two-Dimensional Semantics," in M. Garcia-Carpintero and J. Macia (eds.), *Two-Dimensionalism* (Stanford, CA: CSLI, 2002).

——(ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press, 2002.

Chalmers, D., and Jackson, F., "Conceptual Analysis and Reductive Explanation," *Philosophical Review* 110 (2001): 315–60.

Chisholm, R., "Human Freedom and the Self," in Watson (ed.), *Free Will*, 26–37.

Churchland, Patrick, *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge: MIT Press, 1986.

Churchland, Paul, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78 (1981): 67–90.

——*Matter and Consciousness*. Cambridge: MIT Press, 1984.

——*Being There: Putting Brain, Body, and World Together Again*. Cambridge: MIT Press, 1997.

——"Embodiment and the Philosophy of Mind," in A. O'Hear (ed.), *Current Issues in the Philosophy of Mind*. Cambridge: Cambridge University Press, 1998, 35–51.

——*Mindware*. Oxford: Oxford University Press, 2001.
——"Visual Experience and Motor Action: Are the Bonds Too Tight?," *Philosophical Review* 110 (2001): 495–519.
Clark, A., and Chalmers, D., "The Extended Mind," in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 643–51.
Clarke, R., "Agent Causation and Event Causation in the Production of Free Action," *Philosophical Topics* 24 (1998): 19–48.
Cleveland, T., *Trying Without Willing*. Brookfield, VT: Ashgate, 1997.
Crane, T., and Mellor, H., "There Is No Question of Physicalism," *Mind* 99 (1990): 185–206.
Cussins, A., "Content, Conceptual Content, and Nonconceptual Content," in Gunther (ed.), *Essays on Nonconceptual Content*, 133–63.
Damasio, A., *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books, 1994.
——*The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. San Diego, CA: Harcourt, 1999.
——*Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. San Diego, CA: Harcourt, 2003.
Danto, A., *Analytic Philosophy of Action*. Cambridge: Cambridge University Press, 1973.
Davidson, D., "Actions, Reasons, and Causes," in Mele (ed.), *The Philosophy of Action*, 27–41. Also in Davidson, *Essays on Actions and Events*, 3–19.
——"Agency," in Davidson, *Essays on Actions and Events*, 43–62.
——*Essays on Actions and Events*. Oxford: Clarendon Press, 1980.
——"How is Weakness of the Will Possible?," in Davidson, *Essays on Actions and Events*, 21–42.
——"Intending," in Davidson, *Essays on Actions and Events*, 83–102.
——"Mental Events," in Block (ed.), *Readings in the Philosophy of Psychology*. vol. i, 107–19. Also in Davidson, *Essays on Actions and Events*, 207–25.
——"Thinking Causes," in Heil and Mele (eds.), *Mental Causation*, 3–17.
——"Thought and Talk," in D. Davidson, *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, 1984, 155–70.
DeGrazia, D., *Taking Animals Seriously: Moral Life and Moral Status*. New York: Cambridge, 1996.
Dennett, D., "Animal Consciousness: What Matters and Why," in D. Dennett, *Brainchildren: Essays on Designing Minds*. Cambridge: MIT Press, 1998, 337–52.
——*Consciousness Explained*. Boston: Little, Brown, & Co., 1991.
——*The Intentional Stance*. Cambridge: MIT Press, 1987.
——*Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic Books, 1996.

——"Quining Qualia," in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 226–46.

Descartes, R., *Meditations on First Philosophy*. Trans. J. Cottingham, R. Stoothoff, and D. Murdoch. In Descartes, *The Philosophical Writings of Descartes*. vol. ii, 3–62.

——*Passions of the Soul*. Trans. J. Cottingham, R. Stoothoff, and D. Murdoch. In Descartes, *The Philosophical Writings of Descartes*. vol.i, 326–404.

——*Principles of Philosophy*. Trans. J. Cottingham, R. Stoothoff, and D. Murdoch. In Descartes, *The Philosophical Writings of Descartes*. vol. i, 177–291.

——*The Philosophical Writings of Descartes*. 3 vols. Cambridge: Cambridge University Press, 1985.

De Sousa, R., *The Rationality of Emotion*. Cambridge: MIT Press, 1995.

Doring, S., "Explaining Action By Emotion," *Philosophical Quarterly* 53 (2003): 214–30.

Dostoevsky, F., *The Brothers Karamazov*. 2 vols. Trans. D. Magarshack. Harmondsworth, Middlesex: Penguin, 1975.

——*Crime and Punishment*. Trans. D. Magarshack. Harmondsworth, Middlesex: Penguin, 1976.

——*The Devils*. Trans. D. Magarshack. Harmondsworth, Middlesex: Penguin, 1973.

——*The House of the Dead*. Trans. D. McDuff. Harmondsworth, Middlesex: Penguin, 1986.

——*The Idiot*. Trans. D. Magarshack. Harmondsworth, Middlesex: Penguin, 1977.

Dretske, F., "Change Blindness," *Philosophical Studies* 120 (2004): 1–18.

——"Conscious Experience," in Block, Flanagan, and Güzeldere. (eds.), *The Nature of Consciousness*, 773–88.

——*Explaining Behavior: Reasons in a World of Causes*. Cambridge: MIT Press, 1998.

——"Mental Events as Structuring Causes of Behavior," in Heil and Mele (eds.), *Mental Causation*, 121–36.

——*Naturalizing the Mind*. Cambridge: MIT Press, 1995.

——*Seeing and Knowing*. Chicago: University of Chicago Press, 1969.

Earman, J., *A Primer on Determinism*. Dordrecht: D. Reidel, 1986.

Eilan, N., McCarthy, R., and Brewer, B., (eds.), *Spatial Representation*. Oxford: Blackwell, 1993.

Eliot, T. S., "The Hollow Men," in T. S. Eliot, *Collected Poems: 1909–1962*. London: Faber & Faber, 1974, 89–92.

Elster, J., *Strong Feelings: Emotion, Addiction, and Human Behavior*. Cambridge: MIT Press, 2000.

Evans, G., "Demonstrative Identification," in Gunther (ed.), *Essays on Nonconceptual Content*, 43–74.

Fine, K., "Essence and Modality," *Philosophers' Annual* (1994): 151–66.

Fischer, J. M., and Ravizza, M., *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998.

Fodor, J. "Making Mind Matter More," in J. Fodor, *A Theory of Content and Other Essays*. Cambridge: MIT Press, 1990, 137–59.

Frankfurt, H., "Alternate Possibilities and Moral Responsibility," in Frankfurt, *The Importance of What We Care About*, 1–10.

——"Freedom of the Will and the Concept of a Person," in Frankfurt, *The Importance of What We Care About*, 11–25.

——"Identification and Externality," in Frankfurt, *The Importance of What We Care About*, 58–68.

——"Identification and Wholeheartedness," in Frankfurt, *The Importance of What We Care About*, 159–76.

——"The Importance of What We Care About," in Frankfurt, *The Importance of What We Care About*, 80–94.

——*The Importance of What We Care About*. Cambridge: Cambridge University Press, 1988.

——"The Problem of Action," in Frankfurt, *The Importance of What We Care About*, 69–79. Also in Mele (ed.), *The Philosophy of Action*, 42–52.

Freeman, W., "Emotion is Essential to All Intentional Behaviors," in M. Lewis and I. Granic (eds.), *Emotion, Development, and Self-Organization: Dynamic Systems Approaches to Emotional Development*. Cambridge: Cambridge University Press, 2000, 209–35.

Frege, G., *Collected Papers on Mathematics, Logic, and Philosophy*. Trans. M. Black, et al. Oxford: Blackwell, 1984.

——"Concept and Object," in G. Frege, *Translations from the Writings of Gottlob Frege*. Trans. P. Geach and M. Black. Oxford: Blackwell, 1960, 42–55.

——"Function and Concept," in Frege, *Collected Papers on Mathematics, Logic, and Philosophy*, 137–56.

——"Logic [1897]," in G. Frege, *Posthumous Writings*. Trans. P. Long, et al. Chicago, IL: University of Chicago Press, 1979, 127–51.

——"Thoughts," in Frege, *Collected Papers on Mathematics, Logic, and Philosophy*, 351–72.

Galison, P., *Image and Logic: A Material Culture of Microphysics*. Chicago, IL: University of Chicago Press, 1997.

Gallagher, S., *How the Body Shapes the Mind*. Oxford: Clarendon Press, 2005.

Gallese, V., "The 'Shared Manifold' Hypothesis: From Mirror Neurons to Empathy," in Thompson (ed.), *Between Ourselves: Second-Person Issues in the Study of Consciousness*, 33–50.

Gallese, V., Keysers, C., and Rizzolatti, G., "A Unifying View of the Basis of Social Cognition," *Trends in Cognitive Sciences* 8 (2004): 396–403.

Gendler, T., and Hawthorne, J., (eds.), *Conceivability and Possibility*. Oxford: Clarendon Press, 2002.

——*Perceptual Experience*. Oxford: Clarendon Press, 2006.

Gershon, M., *The Second Brain*. New York: Harper Collins, 1998.

Ginet, C., *On Action*. Cambridge: Cambridge University Press, 1990.

Godfrey-Smith, P., *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press, 1996.

Goldie, P., *The Emotions*. Oxford: Clarendon Press, 2000.

——"Emotions, Feelings, and Intentionality," *Phenomenology and the Cognitive Sciences* (2002): 235–54.

Grahek, N., *Feeling Pain and Being in Pain*. 2nd edn. Cambridge: MIT Press, 2007.

Green, O. H., "Toe Wiggling and Starting Cars: A Re-examination of Trying," *Philosophia* 23 (1994): 171–91.

Greenspan, P., "A Case of Mixed Feelings: Ambivalence and the Logic of Emotion," in Rorty (ed.), *Explaining Emotions*, 223–50.

Gregory, R. (ed.), *Oxford Companion to the Mind*. Oxford: Oxford University Press, 1987.

Griffin, D., *Animal Minds*. Chicago: University of Chicago Press, 2001.

——*Animal Thinking*. Cambridge: Harvard University Press, 1984.

——*The Question of Animal Awareness*. New York: Rockefeller University Press, 1976.

Gunther, Y., "Emotion and Force," in Gunther (ed.), *Essays on Nonconceptual Content*, 279–88.

——(ed.), *Essays on Nonconceptual Content*. Cambridge: MIT Press, 2003.

Güzeldere, G., "The Many Faces of Consciousness," in Block, et al. (eds.), *The Nature of Consciousness: Philosophical Debates*, 1–67.

Haack, S., *Deviant Logic*. Cambridge: Cambridge University Press, 1974.

Haggard, P., "Conscious Awareness of Intention and Action," in N. Eilan and J. Roessler (eds.), *Agency and Self-Awareness*. Oxford: Clarendon Press, 2003, 111–27.

Haken, H., *Principles of Brain Functioning: A Synergetic Approach to Brain Activity, Behavior, and Cognition*. Berlin: Springer, 1996.

Hammett, D., *The Maltese Falcon*. New York: Vintage, 1992.

Hanna, R., "Kant and Nonconceptual Content," *European Journal of Philosophy* 13 (2005): 247–90.

——"Kantian Non-Conceptualism," *Philosophical Studies* 137 (2008): 41–64.

——*Kant and the Foundations of Analytic Philosophy*. Oxford: Clarendon Press, 2001.

——*Kant, Science, and Human Nature*. Oxford: Oxford University Press, 2006.

——"Mathematical Truth and Knowledge Regained: A Positive Solution to Benacerraf's Dilemma," Unpublished MS.

——*Rationality and Logic*. Cambridge: MIT Press, 2006.

Hanna, R., and Ivy, D., "Review of Rockwell's *Neither Brain Nor Ghost*," *Philosophical Psychology* 20 (2007): 277–82.

Hanna, R., and Thompson, E., "The Mind-Body-Body Problem," *Theoria et Historia Scientiarum* 7 (2003): 24–44.

——"Neurophenomenology and the Spontaneity of Consciousness," in E. Thompson (ed.), *The Problem of Consciousness*. Calgary, AL: University of Alberta Press, 2005, 133–62.

Hawking, S., *A Brief History of Time*. New York: Bantam, 1988.

Hawkins, J. M., and Allen, R., (eds.), *Oxford Encyclopedic English Dictionary*. Oxford: Clarendon Press, 1991.

Heidegger, M., *Being and Time*. Trans. J. Macquarrie and E. Robinson. New York: Harper & Row, 1962.

Heil, J., and Mele, A., "Mental Causes," *American Philosophical Quarterly* 28 (1991): 61–71.

——(eds.), *Mental Causation*. Oxford: Oxford University Press, 1993.

Hellman, L., "Introduction," to Hammett, D., *The Big Knockover*. London: Orion, 2005, v–xxii.

Helm, B., "Emotions and Practical Reason: Rethinking Evaluation and Motivation," *Noûs* 35:2 (2001): 190–213.

Hershfield, J., "Structural Causation and Psychological Explanation," *Journal of Mind and Behavior* 22 (2001): 249–62.

Horgan, T., "From Supervenience to Superdupervenience: Meeting the Demands of a Material World," *Mind* 102 (1993): 555–86.

Horgan, T., and Tienson, J., "The Intentionality of Phenomenology and the Phenomenology of Intentionality," in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 520–33.

Huemer, M., and Kovitz, B., "Causation as Simultaneous and Continuous." *Philosophical Quarterly* 53 (2003): 556–65.

Humberstone, L., "Intrinsic/Extrinsic," *Synthese* 108 (1996): 205–67.

Hume, D., *Treatise of Human Nature*. Oxford: Oxford University Press, 1975.

Humphrey, N., *Seeing Red: A Study in Consciousness*. Cambridge: Harvard University Press, 2006.

Humphreys, P., "Aspects of Emergence," *Philosophical Topics* 24 (1996): 53–70.

——"Emergence, Not Supervenience," *Philosophy of Science* 64 (1997): S337–S345.

——"How Properties Emerge," *Philosophy of Science* 64 (1997): 1–17.

Hurley, S., *Consciousness in Action*. Cambridge: Harvard University Press, 1998.

Hurley, S., and Noë, A., "Neural Plasticity and Consciousness," *Biology and Philosophy* 18 (2003): 131–68.

Hursthouse, R., "Arational Actions," *Journal of Philosophy* 88 (1991): 57–68.

Husserl, E., *The Phenomenology of Internal Time Consciousness*. Trans. J. S. Churchill. Bloomington, IN: Indiana University Press, 1964.

Ismael, J. *The Situated Self*. Cambridge: MIT Press, 2007.

Jackendoff, R., *Consciousness and the Computational Mind*. Cambridge: MIT Press, 1987.

Jackson, F., "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982): 127–36.

——"Mental Causation," *Mind* 105 (1996): 377–413.

Jantsch, E. *The Self-Organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution*. New York: Pergamon, 1980.

Jeannerod, M., "Consciousness of Action as an Embodied Consciousness," in Pockett, Banks, and Gallagher (eds.), *Does Consciousness Cause Behavior?*, 25–38.

Johnson-Laird, P., *Mental Models*. Cambridge: Harvard University Press, 1983.

——*How We Reason*. Oxford: Oxford University Press, 2006.

Johnston, M., "Better than Mere Knowledge? The Function of Sensory Awareness," in Gendler and Hawthorne (eds.), *Perceptual Experience*, 260–90.

Juarrero, A., *Dynamics in Action*. Cambridge: MIT Press, 1999.

Judson, H. F., *The Eighth Day of Creation: Makers of the Revolution in Biology*. New York: Simon & Schuster, 1979.

Kane, R., *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press, 2005.

——*The Oxford Handbook of Free Will*. Oxford: Oxford University Press, 2002.

Kant, I., "Concerning the Ultimate Ground of the Differentiation of Directions in Space," in I. Kant, *Theoretical Philosophy: 1755–1770*. Trans. D. Walford and R. Meerbote. Cambridge: Cambridge University Press, 1992, 365–72.

——*Critique of the Power of Judgment*. Trans. P. Guyer and E. Matthews. Cambridge: Cambridge University Press, 2000.

——*Critique of Pure Reason*. Trans. P. Guyer and A. Wood. Cambridge: Cambridge University Press, 1997.

——*Groundwork of the Metaphysics of Morals*. Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy*. Cambridge: Cambridge University Press, 1996, 37–108.

——*Prolegomena to Any Future Metaphysics*. Trans. J. Ellington. Indianapolis, IN: Hackett, 1977.

Kauffman, S. A., *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. New York: Oxford University Press., 1995.

——*The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford University Press., 1993.

Kelso, J. A. S., *Dynamic Patterns*. Cambridge: MIT Press, 1995.

Kenny, A., *Action, Emotion, and Will*. London: Routledge & Kegan Paul, 1963.

Kihlstrom, J., "The Cognitive Unconscious," *Science* 237 (1987): 1445–52.

Kim, J., "Can Supervenience and 'Non-Strict Laws' Save Anomalous Monism?," in Heil and Mele (eds.), *Mental Causation*, 19–26.

——"Epiphenomenal and Supervenient Causation," in Kim, *Supervenience and Mind*, 92–108.

——"Making Sense of Downward Causation," in P. B. Anderson, et al. (eds.), *Downward Causation*. Denmark: Aarhus University Press, 2000, 305–21.

——"Mechanism, Purpose, and Explanatory Exclusion," in Kim, *Supervenience and Mind*, 237–64.

——*Mind in a Physical World*. Cambridge: MIT Press, 1998.

——"Multiple Realization and the Metaphysics of Reduction," in Kim, *Supervenience and Mind*, 309–35.

——"The Myth of Nonreductive Materialism," in Kim, *Supervenience and Mind*, 265–84.

——"The Non-Reductivist's Troubles with Mental Causation," in Kim, *Supervenience and Mind*, 336–57.

——*Philosophy of Mind*. 2nd edn. Boulder, CO: Westview Press, 2005.

——*Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press, 2005.

——*Supervenience and Mind*. Cambridge: Cambridge University Press, 1993.

Kirk, R., *Zombies and Consciousness*. Oxford: Oxford University Press, 2005.

Korsgaard, C., *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996.

Koslicki, K., *The Structure of Objects*. Oxford: Oxford University Press, 2008.

Kripke, S., "Identity and Necessity," in A.W. Moore (ed.), *Meaning and Reference*. Oxford: Oxford University Press, 1993, 162–91.

——*Naming and Necessity*. 2nd edn. Cambridge, MA: Harvard University Press, 1982.

Kuhse, H., and Singer, P., "Individuals, Humans, and Persons: The Issue of Moral Status," in P. Singer, *Unsanctifying Human Life*. Oxford: Blackwell, 2002, 188–98.

Langton, R., and Lewis, D., "Defining 'Intrinsic'," *Philosophy and Phenomenological Research* 58 (1998): 333–45.

Lee, G., "The Experience of Right and Left," in Gendler and Hawthorne (eds.), *Perceptual Experience*, 291–315.

Leibniz, G. W. F., *Monadology*, in R. Ariew and D. Garber (eds.), *Leibniz: Philosophical Essays*. Indianapolis: Hackett, 1989, 213–34.

Levine, J., "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983): 354–61.

——"On Leaving Out What It's Like," in M. Davies and G. Humphreys (eds.), *Consciousness: Psychological and Philosopical Essays*. (Oxford: Blackwell, 1993).

Lewis, D., "Mad Pain and Martian Pain," in Block, (ed.), *Readings in the Philosophy of Psychology*. vol. i, 216–22.

——"What Experience Teaches," in Lycan (ed.), *Mind and Cognition*, 499–518.

Libet, B., "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action," *Behavioral and Brain Sciences* 8 (1985): 529–66.

Libet, B., Gleason, C., Wright, E., and Pearl, D., "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Unconscious Initation of a Freely Voluntary Act," *Brain* 106 (1983): 623–42.

Libet, B., and Haggard, P., "Conscious Intention and Brain Activity," *Journal of Consciousness Studies* 8 (2000): 47–63.

Lowe, E. J., *An Introduction to the Philosophy of Mind*. Cambridge: Cambridge University Press, 2000.

——"The Causal Autonomy of the Mental," *Mind* 102 (1993): 629–44.

Lycan, W., (ed.), *Mind and Cognition*. Oxford: Blackwell, 1990.

Lyons, W., *The Disappearance of Introspection*. Cambridge: MIT Press, 1986.

——*Emotion*. Cambridge: Cambridge University Press, 1980.

McCulloch, G., *The Mind and its World*. London: Routledge, 1995.

MacDonald, G., and Macdonald, C., "The Metaphysics of Mental Causation," *Journal of Philosophy* (2006): 539–76.

McGinn, C., "Can We Solve the Mind-Body Problem?," in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 394–405.

Mackie, J. L., "Causes and Conditions," in Sosa and Tooley (eds.), *Causation*, 33–55.

McLaughlin, B., "On Davidson's Response to the Charge of Epiphenomenalism," in Heil and Mele (eds.), *Mental Causation*, 27–40.

——"The Rise and Fall of British Emergentism," in Beckermann et al. (eds.), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin: De Gruyter, 1992.

——*The Cement of the Universe*. Oxford: Oxford University Press, 1974.

Margolis, E., and Laurence, S., *Concepts: Core Readings*. Cambridge: MIT Press, 1999.

Matthews, G., "Consciousness and Life," *Philosophy* 52 (1977): 13–26.

——*Mental Content*. Oxford: Blackwell, 1989.

Meinong, A., "The Theory of Objects," in R. Chisholm (ed.), *The Background of Phenomenology*. New York: Free Press, 1960, 76–117.

Mele, A., "Introduction," in Mele (ed.), *The Philosophy of Action*, 1–26.

——*Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford: Oxford University Press, 1987.

——*Motivation and Agency*. Oxford: Oxford University Press, 2003.

——(ed.), *The Philosophy of Action*. Oxford: Oxford University Press, 1997.

Melzack, R., "Pain," in Gregory (ed.), *Oxford Companion to the Mind*, 574–5.

Merleau-Ponty, M., *Phenomenology of Perception*. Trans. C. Smith. London: Routledge & Kegan Paul, 1962.

Montague, R., "Logical Necessity, Physical Necessity, Ethics, and Quantifiers," in R. Montague, *Formal Philosophy*. New Haven, CT: Yale University Press, 1974, 71–83.

Montero, B., "The Body Problem," *Noûs* 33 (1999): 183–200.

——"Post-Physicalism," *Journal of Consciousness Studies* 8 (2001): 61–80.

Moore, G. E., "The Refutation of Idealism," in G. E. Moore, *Selected Writings*. London: Routledge, 1993, 23–44.

Moran, D., *Introduction to Phenomenology*. London: Routledge, 2000.

Nagel, T., "Brain Bisection and the Unity of Consciousness," in Nagel, *Mortal Questions*, 147–64.

——"Conceiving the Impossible and the Mind-Body Problem," *Philosophy* 73 (1998): 337–52.

——"Death," in Nagel, *Mortal Questions*, 1–10.

——*Mortal Questions*. Cambridge: Cambridge University Press, 1979.

——"Panpsychism," in Nagel, *Mortal Questions*, 181–95.

——"The Psychophysical Nexus," in P. Boghossian and C. Peacocke (eds.), *New Essays on the A Priori*. Oxford: Clarendon Press, 2000, 433–71.

——"What is it like to be a bat?," in Block (ed.), *Readings in the Philosophy of Psychology*. vol. i, 159–68. Also in Nagel, *Mortal Questions*, 165–80.

Nemirow, L., "Physicalism and the Cognitive Role of Acquaintance," in Lycan (ed.), *Mind and Cognition*, 490–8.

Nicolis, G., and Prigogine, I., *Self-Organization in Nonequilibrium Systems*. New York: Wiley, 1977.

Nietzsche, F., *Beyond Good and Evil*. Trans. W. Kaufmann. New York: Vintage, 1966.

Noë, A., *Action in Perception*. Cambridge: MIT Press, 2004.

O'Connor, T., "Emergent Properties," *American Philosophical Quarterly* 31 (1994): 91–104.

——*Persons and Causes*. New York: Oxford University Press, 2000.

O'Connor, T., and Wong, H. Y., "Emergent Properties," *Stanford Encyclopedia of Philosophy (Winter 2006 Edition)*, E. N. Zalta (ed.), URL = <http://plato.stanford.edu/archives.win2006/entries/properties-emergent/>.

——"The Metaphysics of Emergence," *Noûs* 39 (2005): 658–78.

Olson, E., *The Human Animal*. Oxford: Oxford University Press, 1997.

Oppenheim, P., and Putnam, H., "Unity of Science as a Working Hypothesis," in H. Feigl, et al. (eds.), *Minnesota Studies in the Philosophy of Science*. Minneapolis, MN: University of Minnesota Press, 1958. Vols. 2, 3–36.

O'Regan, K., Rensink, R., and Clark, J., "Change Blindness as a Result of 'Mudsplashes'," *Nature* 398 (1999): 34.

O'Shaughnessy, B., "Trying (as the Mental 'Pineal Gland')," in Mele (ed.), *The Philosophy of Action*, 53–74.

——*The Will*. 2 vols. Cambridge: Cambridge University Press, 1980.

Pascal, B., *Pensées*. Trans. A. J. Krailsheimer. Harmondsworth, Middlesex: Penguin, 1966.

Penfield, W., *The Mystery of the Mind*. Princeton, NJ: Princeton University Press, 1975.

Pereboom, D., "Robust Nonreductive Materialism," *Journal of Philosophy* 99 (2002): 499–531.

Perry, J., *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press, 2001.

——"The Problem of the Essential Indexical," *Noûs* 13 (1979): 3–21.

Pert, C., *Molecules of Emotion*. New York: Scribner, 1997.

Place, U.T., "Is Consciousness a Brain Process?," *British Journal of Psychology* 47 (1956): 44–50. Also in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 55–60.

Pleydell-Pearce, I., "Biofeedback," in Gregory (ed.), *Oxford Companion to the Mind*, 88–92.

Pockett, S., Banks, W.B., and Gallagher, S. (eds.), *Does Consciousness Cause Behavior?* Cambridge: MIT Press, 2006.

Poellner, P., "Non-Conceptual Content, Experience, and the Self," *Journal of Consciousness Studies* 10 (2003): 32–57.

Port, R. F., and Van Gelder, T., (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge: MIT Press, 1995.

Pred, R., *Onflow: Dynamics of Consciousness and Experience*. Cambridge: MIT Press, 2005.

Priest, G., *An Introduction to Non-Classical Logic*. Cambridge: Cambridge University Press, 2001.

Prigogine, I., *Being and Becoming: Time and Complexity in the Physical Sciences*. New York: W. H. Freeman, 1980.

Prinz, J., *Gut Reactions: A Perceptual Theory of Emotion*. New York: Oxford University Press, 2004.

Putnam, H., "Brains and Behavior," in Putnam, *Mind, Language, and Reality: Philosophical Papers, vol. 2*, 325–41.

——*Mind, Language, and Reality: Philosophical Papers, vol. 2*. Cambridge: Cambridge University Press, 1975.

——"The Nature of Mental States," in Putnam, *Mind, Language, and Reality: Philosophical Papers, vol. 2*, 429–40.

——"There is at Least One A Priori Truth," in H. Putnam, *Realism and Reason: Philosophical Papers, vol. 3*. Cambridge: Cambridge University Press, 1983, 98–114.

——*Reason, Truth, and History*. Cambridge: Cambridge University Press, 1981.

Quine, W.V. O., *Word and Object*. Cambridge: MIT Press, 1960.

Rachels, J., *The Elements of Moral Philosophy*. 4th edn. New York: McGraw-Hill, 2003.

Rensink, R., O'Regan, K., and Clark, J., "On the Failure to Detect Changes in Scenes Across Brief Interruptions," *Visual Cognition* 7 (2000): 17–42.

Roberts, R., "Solomon on Control of Emotions," *Philosophy and Phenomenological Research* 44 (1984): 395–403.

Robinson, J., "Emotion, Judgment, and Desire," *Journal of Philosophy* 80 (1983): 731–41.

Rockwell, W.T., *Neither Brain Nor Ghost: A Nondualist Alternative to the Mind-Brain Identity Theory*. Cambridge: MIT Press, 2005.

Rorty, A. O., "Explaining Emotions," in Rorty (ed.), *Explaining Emotions*, 103–26.

——(ed.), *Explaining Emotions*. Berkeley, CA: University of California Press, 1980.

Rosenberg, G., *A Place for Consciousness*. Oxford: Oxford University Press, 2005.

Rosenthal, D., "A Theory of Consciousness," in Block, Flanagan, and Guzeldere (eds.), *The Nature of Consciousness*, 729–53.

——"Two Concepts of Consciousness," *Philosophical Studies* 94 (1986): 329–59.

Rowlands, M., *Body Language*. Cambridge: MIT Press, 2006.

Royce, J., *The Letters of Josiah Royce*. Chicago: University of Chicago Press, 1970.
Rupert, R., "*Ceteris Paribus* Laws, Component Forces, and the Nature of Special Science Properties," *Noûs* 42(2008): 349–80.
Russell, B., *The Analysis of Matter*. London: Kegan Paul, 1927.
Ryle, G., *The Concept of Mind*. London: Hutchinson, 1949.
Sacks, O., *A Leg to Stand On*. New York: Summit Books, 1984.
——*The Man Who Mistook His Wife for a Hat*. New York: Harper Perennial, 1987.
Sartre, J.-P., *Being and Nothingness*. Trans. H. Barnes. New York: Philosophical Library, 1956.
——*The Psychology of Imaginination*. Secaucas, NJ: Citadel Press, 1965.
——*The Transcendence of the Ego*. Trans. F. Williams and R. Kirkpatrick. New York: Farrar, Strauss, & Geroux, 1987.
Savitt, S., "Introduction," in S. Savitt (ed.), *Time's Arrows Today*. Cambridge: Cambridge University Press, 1995, 1–19.
Schaffer, J., Lewis, D., Hall, N., Collins, J., and Paul, L., "Special Issue: Causation," *Journal of Philosophy* 97 (2000): 165–256.
Schrödinger, E., *What is Life?: The Physical Aspect of the Living Cell*. Cambridge: Cambridge University Press, 1992.
Schutz, A., *The Phenomenology of the Social World*. Trans. G. Walsh and F. Lehnert. Evanston, IL: Northwestern University Press, 1967.
Searle, J., *Intentionality*. Cambridge: Cambridge University Press, 1983.
——*Mind: A Brief Introduction*. Oxford: Oxford University Press, 2004.
——*Minds, Brains, and Science*. Cambridge: Harvard University Press, 1984.
——*Rationality in Action*. Cambridge: MIT Press, 2001.
——*The Rediscovery of the Mind*. Cambridge: MIT Press, 1992.
Sehon, S., "An Argument Against the Causal Theory of Action Explanation," *Philosophy and Phenomenological Research* 60 (2000): 67–85.
——"Connectionism and the Causal Theory of Action Explanation," *Philosophical Psychology* 11 (1998): 511–31.
Sellars, W., "Philosophy and the Scientific Image of Man," in W. Sellars, *Science, Perception, and Reality*. New York: Humanities Press, 1963, 1–40.
Shakespeare, W., *Hamlet*. Ed. G. L. Kittredge. Lexington, MA: Xerox, 1967.
——*The Merchant of Venice*. Ed. G. L. Kittredge. Lexington, MA: Xerox, 1968.
Shapiro, L., *The Mind Incarnate*. Cambridge: MIT Press, 2004.
——"Multiple Realizations," *Journal of Philosophy* 97 (2000): 635–54.
Siegel, D., *Invasion of the Body Snatchers* (1956).
Silberstein, M., "Converging On Emergence: Consciousness, Causation and Explanation," *Journal of Consciousness Studies* 8 (2001): 61–98.

Silberstein, M., and McGeever, J., "The Search For Ontological Emergence," *Philosophical Quarterly* 49 (1999): 182–200.

Simons, D., and Levin, D., "Change Blindness," *Trends in Cognitive Sciences* 1 (1997): 261–7.

Smart, J. J. C., "Sensations and Brain Processes," *Philosophical Review* 68 (1959): 141–56. Also in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, 60–8.

——and Williams, B., *Utilitarianism: For and Against*. Cambridge: Cambridge University Press, 1973.

Smiley, T., "Relative Necessity," *Journal of Symbolic Logic* 28 (1963): 113–34.

Smith, M., *The Moral Problem*. Oxford: Blackwell, 1994.

Solomon, R., "Emotions and Choice," in Rorty (ed.), *Explaining Emotions*, 251–81.

——*Not Passion's Slave: Emotion and Choice*. Oxford: Oxford University Press, 2003.

——*The Passions*. Indianapolis: Hackett 1993.

——"On the Passivity of the Passions," in Solomon, *Not Passion's Slave: Emotions and Choice*, 195–232.

——"Thoughts and Feelings: What is a 'Cognitive Theory' of the Emotions, and Does it Neglect Affectivity?," in Solomon, *Not Passion's Slave: Emotion and Choice*, 178–94.

Sosa, E., "Davidson's Thinking Causes," in Heil and Mele (eds.), *Mental Causation*, 41–50.

Sosa, E., and Tooley, M. (eds.), *Causation*. Oxford: Oxford University Press, 1993.

Speaks, J., "Is There a Problem about Nonconceptual Content?," *Philosophical Review* 114 (2005): 359–98.

Spinoza, B., *The Ethics and Selected Letters*. Trans. S. Shirley. Indianapolis: Hackett, 1982.

Stapp, H., *Mind, Matter, and Quantum Mechanics*. Munich: Springer, 1993.

Stein, E., *On the Problem of Empathy*. Trans. W. Stein. The Hague: Martinus Nijhoff, 1964.

Stephan, A., "Emergentism, Irreducibility, and Downward Causation," *Grazer Philsophische Studien* 65 (2002): 77–93.

——"Emergence: A Systematic View of its Historical Facets," in Beckermann, et al. (eds), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*, 25–47.

Stoecker, R., "Climbers, Pigs, and Wiggled Ears: The Problem of Waywardness in Action Theory," in S. Walter and H.-D. Heckmann (eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Exeter, UK: Imprint Academic, 2003, 295–322.

Stoutland, F., "Davidson on Intentional Behavior," in E. LePore and B. McLaughlin (eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell, 1985, 44–59.

Strawson, P.F., *Individuals*. London: Methuen, 1959.

Sudnow, D., *Ways of the Hand*. Cambridge: MIT Press, 2001.

Thelen, E., and Smith, L., *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge: MIT Press, 1994.

Thomas, D., "Do Not Go Gentle into that Good Night," in O. Williams (ed.), *The Pocket Book of Modern Verse*. New York: Washington Square, 1973, 486.

Thompson, E., (ed.), *Between Ourselves: Second-Person Issues in the Study of Consciousness*. Charlottesville, VA: Imprint Academic, 2001.

——"Empathy and Consciousness," in Thompson (ed.), *Between Ourselves: Second-Person Issues in the Study of Consciousness*, 1–32.

——*Mind in Life*. Cambridge: Harvard University Press, 2007.

——"Sensorimotor Subjectivity and the Enactive Approach to Experience," *Phenomenology and the Cognitive Sciences* 4 (2005): 407–27.

Thompson, E., and Varela, F., "Radical Embodiment: Neural Dynamics and Consciousness," *Trends in Cognitive Sciences* 5 (2001): 418–25.

Tye, M., *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge: MIT Press, 1996.

Van Fraassen, B., *Laws and Symmetry*. Oxford: Clarendon Press, 1989.

——*The Scientific Image*. Oxford: Oxford University Press, 1980.

Van Gulick, R., "Who's in Charge Here? And Who's Doing All the Work?," in Heil and Mele (eds.), *Mental Causation*, 233–56.

Varela, F., *Principles of Biological Autonomy*. New York: Elsevier/North-Holland, 1979.

Varela, F., Thompson, E., and Rosch, E., *The Embodied Mind*. Cambridge: MIT Press, 1991.

Velleman, D., "The Way of the Wanton" (Aug. 13, 2007). Available at SSRN: <http://ssrn.com/abstract=1006893>.

Wallace, R. J., "Addiction as Defect of the Will: Some Philosophical Reflections," *Law and Philosophy* 18 (1999): 621–55.

Watkins, E., *Kant and the Metaphysics of Causality*. Cambridge: Cambridge University Press, 2005.

Watson, G. (ed)., *Free Will*. 2nd edn. Oxford: Oxford University Press, 2003.

Weatherson, B., "Intrinsic vs. Extrinsic Properties," *The Stanford Encyclopedia of Philosophy (Spring 2007 Edition)*, E. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2007/entries/intrinsic-extrinsic/>.

Weber, A., and Varela, F., "Life After Kant: Natural Purposes and the Autopoietic Foundations of Biological Individuality," *Phenomenology and the Cognitive Sciences* 1 (2002): 97–125.

Weber, B., "Life," *The Stanford Encyclopedia of Philosophy (Spring 2006 Edition)*, E. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2006/entries/life/>.

Wegner, D., *The Illusion of Conscious Will*. Cambridge: MIT Press, 2002.

——"The Mind's Best Trick: How We Experience Conscious Will," *Trends in Cognitive Sciences* 7 (2003): 65–9.

Wegner, D., and Wheatley, T., "Apparent Mental Causation: Sources of the Experience of Will," *American Psychologist* 54 (1999): 480–92.

Weiskrantz, L., *Blindsight*. Oxford: Clarendon Press, 1986.

Whitehead, A. N., *Process and Reality*. New York: Free Press, 1978.

Wider, K., *The Bodily Basis of Consciousness*. Ithaca, NY: Cornell University Press, 1997.

Wilson, C., *The Invisible World: Early Modern Philosophy and the Invention of the Microscope*. Princeton, NJ: Princeton University Press, 1995.

Wilson, G., *The Intentionality of Human Action*. Amsterdam: North-Holland Pub. Co., 1980.

Wittgenstein, L., *Philosophical Investigations*. 3rd edn. Trans. G. E. M. Anscombe. New York: Macmillan, 1953.

——*Remarks on the Philosophy of Psychology*. 2 vols. Trans. G. E. M. Anscombe. Chicago, IL: University of Chicago Press, 1980.

——*Tractatus Logico-Philosophicus*. Trans. C. K. Ogden. London: Routledge & Kegan Paul, 1981.

Wolf, S., "Moral Saints," *Journal of Philosophy* 79 (1982): 419–39.

Yablo, S., "Concepts and Consciousness," *Philosophy and Phenomenological Research* 59 (1999): 455–63.

——"Is Conceivability a Guide to Possibility?," *Philosophy and Phenomenological Research* 53 (1993): 1–42.

——"Mental Causation," *Philosophical Review* 101 (1992): 245–80.

Yeats, W. B., "Among School Children," in *Collected Poems of W.B. Yeats*. London: Macmillan, 1973, 242–5.

Zemach, E., "What is Emotion?," *American Philosophical Quarterly* 38 (2001): 197–207.

Zhu, J., and Thagard, P., "Emotion and Action," *Philosophical Psychology* 15 (2003): 19–36.

Index