KANT AND THE SIGNIFICANCE OF SELF-CONSCIOUSNESS

by

Matthew Brendan Boyle

B.A., Social Studies and Philosophy, Harvard University, 1994

B.Phil., Philosophy, Oxford University, 1996

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

Matthew Brendan Boyle

It was defended on

August 30, 2005

and approved by

Robert Brandom, Distinguished Service Professor of Philosophy

Stephen Engstrom, Associate Professor of Philosophy

Michael Thompson, Associate Professor of Philosophy

James Conant, Professor of Philosophy, University of Chicago

Dissertation Director: John McDowell, University Professor of Philosophy

KANT AND THE SIGNIFICANCE OF SELF-CONSCIOUSNESS

Matthew Boyle, Ph.D.

University of Pittsburgh, 2005

Human beings who have mastered a natural language are *self-conscious creatures*: they can think, and indeed speak, about themselves in the first person. This dissertation is about the significance of this capacity: what it is and what difference it makes to our minds. My thesis is that the capacity for self-consciousness is essential to rationality, the thing that sets the minds of rational creatures apart from those of mere brutes. This, I argue, is what Kant was getting at in a famous passage of his *Critique of Pure Reason*, when he claimed that a representation which could not be "accompanied with the 'I think'" would be "nothing to me" as a thinking being. I call this claim *the Kantian thesis*.

My dissertation seeks to explain and defend the Kantian thesis, to show how it entails that the advent of self-consciousness brings with it a new kind of mind, and to sketch the implications of this point for a philosophy of mind that seeks to understand the minds of rational creatures. This involves, on the one hand, an investigation of the kinds of capacities that characterize a rational creature, and, on the other hand, an argument connecting reason with self-consciousness. I show that a rational creature, in the interesting sense, is one capable of conceptual representation, and I argue that (1) to represent conceptually is to represent in a way that decouples information from any particular context or purpose, (2) this special form of representation is possible only in a creature that can reflect explicitly on grounds for judging a proposition true, and (3) to have this capacity to reflect explicitly on grounds is necessarily to have the crux of self-consciousness. If this is right, then the representations of a rational creature must differ from those of a nonrational creature not merely in complexity but in kind. The dissertation sketches the implications of this point for various forms of naturalism and reductionism in the philosophy of mind, for debates about how to explain "first person authority," and for our understanding of the sort of failure of self-consciousness involved in self-deception.

TABLE OF CONTENTS

PREFACE

That this dissertation exists at all reflects the kindness of many people. A dissertation is of course a poor sort of response to that kindness and all that it has meant to me. Nevertheless, to whatever extent I've managed, in the pages that follow, really to try to get to the bottom of things – wherever I've cared less about vindicating some view of mine than about understanding what's true – I owe the moral resources for that effort to the people named below. I'm more grateful to them than a preface can express, but I will at least say who they are.

In the first place, I'm grateful to the members of my committee, and above all to my director, John McDowell. I came to Pittsburgh because I had read McDowell's *Mind and World*: it was the work that made me care about philosophy enough to go to graduate school at all. The content of this dissertation is everywhere indebted to that book, most obviously to the footnote on its page 47. And John himself has been even more helpful to me than his book. I could not have had a better advisor.

I've been fortunate, though, to benefit not just from my advisor but from every single member of my committee. Everybody says this, but my impression is that the experience I had is actually quite rare. I won't try to describe what each committee member did for me. All of them read and commented on huge parts of this document in draft, often in several drafts. All of them have had a tremendous influence on what I think, an influence that stands out almost too nakedly in the pages that follow.

What is inevitably not as plain is the influence of innumerable conversations with friends about the topics discussed in this thesis. Nearly all the progress I've ever made in philosophy has been made in conversation, and my main partners in conversation have been: Alp Aker, Ian Blecher, Anton Ford, Matthias Haase, Ben Laurence, Lissa Merritt, Susanna Schellenberg, and Herb Wilson. Kieran Setiya has

also had a large impact on me since his arrival a few years ago, and I'm grateful to him for very helpful comments on several chapters. I should also mention two earlier influences who were very important to me: Jeff Seidman, my friend since high school, who got me interested in philosophy and who has been an ongoing source of intellectual stimulation over a period of decades; and Peter Hacker, the advisor of my B.Phil. thesis at Oxford, who taught me to think carefully and who urged me, at a crucial moment, to work on topics only if I could see real reasons to care about them.

Doug Lavin and Sebastian Rödl have had so large an influence that they merit their own paragraph. Doug is responsible, for better or worse, for my taking what one or the other of us started calling "the metaphysical turn." My whole way of thinking is a product of conversations with him, and his help and support are what made it possible for me to finish this document and to survive the attempt to get a job on the basis of it. As to Sebastian: although he arrived on the scene when this dissertation was already well underway, I can now hardly remember how I thought of the project before he came along. Reading his work – especially his book on the first person, which he generously showed to me in draft – has transformed my conception, not just of the issues treated here, but of philosophy in general.

Finally, there are people who have helped me, not primarily to be a better philosopher, but rather, at least fitfully, to be my better self. Several of the people mentioned above belong to this most precious group, but I'll only list names that haven't yet appeared: Katerina Gakiopoulou, Katherine Ibbett, Alison Lake, my brother Latham Boyle, and of course my parents. I dedicate the dissertation to my brother Alex Boyle, who does not say "I." He was not exactly the impetus for what follows, but he has been on my mind throughout.

*

Portions of this dissertation were written while I held a National Science Foundation Graduate Research Fellowship. Any justification for this unlikely munificence would have to rely heavily on etymology. Be that as it may, I was very grateful for the funding.

INTRODUCTION

> I have to do with nothing save reason itself and its pure thinking; and to obtain knowledge of these, there is no need to go far afield, since I come upon them in my own self.
>
> Immanuel Kant, *Critique of Pure Reason*, Axiv[1]

What sort of a creature is a human being? Heard in one register, this question calls for an empirical investigation. Heard in another, it seems to call for something else, an investigation of ourselves not by observation but by self-reflection. The traditional philosophical answer to the question "What is man?" – namely, "A rational animal" – can be understood as a response to the question taken in the latter way. For if "man" means "the kind of creature *I am*," and if I am able even to entertain the question "What is man?", then the philosophical answer to this question is already available to me without further empirical research. A creature that can ask itself what it is is necessarily a rational creature, and one capable of sensing and acting is necessarily an animal. The fact that there is such a way of hearing the question "What is man?", a way that asks us to look not *at* ourselves but *into* ourselves, is one indication of how fundamental the capacity for self-consciousness is to our nature.

My project in this dissertation is to examine our capacity for self-consciousness: what it is and what it does for us. My aim is to show that it does a great deal, more than many contemporary

---

[1] References to works by Kant in this dissertation will be by volume and page number of the Prussian Academy Edition of his works (1900), except in the case of references to the first *Critique*, where I have followed the usual practice of giving the pagination as in the first ("A") and the second ("B") editions. When not referring to the first *Critique*, I use the following abbreviations: AN = *Anthropology from a Pragmatic Point of View*, CJ = *Critique of Judgment*, CPR = *Critique of Practical Reason*, G = *Grounding for the Metaphysics of Morals*, L = *Logic*, LM = *Lectures on Metaphysics*, P = *Prolegomena to Any Future Metaphysics*. The English translations from which I quote are listed in the Bibliography; in the case of the first *Critique*, I have mostly preferred the Kemp Smith translation, but have occasionally followed the Guyer and Wood. On the rare occasions when I have made a change in translation, I have placed a "T" after the page reference.

philosophers of mind appreciate – that, in fact, self-consciousness is *the* distinctive feature of our kind of mind, the crux of our rationality. The inspiration for this outlook comes from Kant, who claimed that man's capacity to have "the representation *I*" "raises him infinitely above all the other beings living on earth."[2] I will be offering an account of what Kant meant by this and why he thought it was true, and I will be arguing that he was right to think so. The purpose of this introduction is to give some context to this project: on the one hand, to say something about what *kind* of claim I take Kant to be making (§1); and on the other hand, to indicate the relevance of his claim to some topics discussed in contemporary philosophy of mind (§2). At the end, I give a brief overview of the coming chapters (§3).

1. RATIONALITY AS AN ASPECT OF HUMAN NATURE

I shall be arguing that self-consciousness is the essential trait of our kind of mind, the property that qualifies us as rational creatures. But what sorts of claims are these? What is the significance of calling something an "essential" trait of our minds, or of saying that it characterizes our "nature"? Although I cannot offer anything like a full treatment of these topics, I must say something about them in order to avoid certain natural misunderstandings of the claims I will be defending. A useful way to do this will be to reflect on the idea that human beings are *by nature* rational creatures. Once we have a clearer understanding of this idea, we will be in a better position to appreciate the significance of claims about the connection between rationality and self-consciousness.

According to a tradition reaching back at least as far as Aristotle, human beings are set apart from other terrestrial creatures by their rationality. Other animals, according to this tradition, are capable of sensation and appetite, but they are not capable of *thought*, the kind of mental activity associated with the rational part of the soul. Human beings, by contrast, are rational animals, and an understanding of

---

[2] Kant, *Anthropology*, I, §1 (Ak. 7:129).

our minds must begin from a recognition of this distinctiveness. For, the tradition holds, the presence of rationality does not just add one more power to the human mind; it transforms all of our principal mental powers, placing us in a different order of being from the one to which dumb brutes belong.

Although the historical roots of this tradition run deep, I think it is fair to say that many contemporary philosophers regard it with suspicion. No one doubts, of course, that there are all sorts of differences between human beings and other animals, but we do not have much use these days for talk of parts of the soul or orders of being. Our philosophy of mind seeks not primarily to characterize the human mind's distinctiveness but to show how our minds fit into the natural world, and the demand that human mentality be conceived as fundamentally continuous with the mentality of other animals looks to many like just a piece of naturalistic common sense. For whatever we mean by calling our minds "rational," surely this must be compatible with a recognition that the human mind is one kind of animal mind, which has arisen through the same sorts of evolutionary processes that also produced the minds we call "nonrational." And our increasingly detailed sense of the psychological, behavioral, and neurophysiological similarities between ourselves and other animals, and of the ways in which we "rational" creatures frequently think and choose on patently nonrational principles, only adds to the pressure to see the specialness of our minds as a matter of degree rather than one of kind.[3]

The contemporary tendency to deny that the human mind differs categorially from the minds of other animals is not without precedent: it is arguably a legacy of British Empiricism. Thus, although Locke defines Reason as "that Faculty whereby Man is supposed to be distinguished from beasts," he also says of beasts that "if they have any *Ideas* at all, and are not bare Machins (as some would have them) we cannot deny them to have some Reason."[4] In a similar vein, Hume remarks that "no truth appears to

---

[3] For a review of continuities between the cognition of humans and the cognition of other primates, see for instance Michael Tomasello and Josep Call, *Primate Cognition* (1997). Standard works on nonrational tendencies in human thinking include R. E. Nisbett and L. Ross, *Human Inference: Strategies and Shortcomings of Social Judgement* (1980) and D. Kahneman, P. Slovic and A. Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (1982).

[4] *Essay concerning Human Understanding,* IV, XVII, 1 (1975, p. 668) and II, XI, 11 (1975, p. 160).

me more evident, than that beasts are endow'd with thought and reason as well as men."[5] We differ from brutes, according to these philosophers, not in the fundamental nature of our mental states, but merely in the complexity of the contents our minds can entertain and the sophistication of the operations we can perform on such contents.

There is another tradition in modern philosophy, however, which adheres more closely to the Aristotelian position described above. The great figure in this tradition is Kant, who follows Aristotle in holding that, although brutes are capable of sensation and appetite, they are not capable of thought and judgment.[6] In his lectures on metaphysics from the academic year 1784-5, for instance, Kant asked:

> Now how can we conceive animals as beings below human beings? … [W]e can think of things which are *below* us, whose representations are different in species and not merely in degree. [For] we perceive in ourselves a specific feature of the understanding and of reason, namely consciousness; if I take this away there still remains something left yet, namely, sensation <*sensus*>, imagination <*imaginatio*>, the former being intuition with presence, the latter without presence of the object. We can also think a reproduction <*reproduction*>, anticipation <*praevision*>, without the least self-consciousness; but such a being could not prescribe rules to itself… (Ak. 28:449-450)

Thus, although Kant admits that mere brutes can have the sorts of "representations" involved in sensation, imagination, anticipation, and a kind of recollection, he holds that only rational beings can prescribe rules to themselves. Moreover, he elsewhere argues that to be capable of prescribing rules to oneself is to be capable of a new *kind* of representation, the kind he calls a *concept*. And, finally, he takes the fact that rational beings are capable of thinking conceptually to inform all of their representations. For, he holds – and here he emphasizes something that Aristotle does not – a concept-user must be capable of "accompanying all of [its] representations with the 'I think'." A representation that was not thus accompaniable "would be impossible, or at least would be nothing to me" (B131-2).[7] Kant's view is thus that the capacity for rationality not only increases the complexity of our representations or the

---

[5] *Treatise of Human Nature*, I, III, XVI (1978, p. 176).

[6] Kant's inheritance of these Aristotelian ideas was mediated by Leibniz, who self-consciously attempted to defend a modernized version of the Aristotelian position as an alternative to Cartesianism. I will focus on Kant, however, since his defense of the position is fuller and (to my mind) more plausible.

sophistication of operations we can perform on them, but transforms the very nature of our representations, conferring on them a new *kind* of content, the kind appropriate to self-conscious reason.

Contemporary discussions of the connection between mind and reason tend to overlook the possibility of such a view. There are some contemporary philosophers who argue that the very idea of a mind – at least, the sort of mind capable of holding propositional attitudes – is tied to the idea of rationality in such a way that our ascription of beliefs and desires to dumb brutes can only be a kind of anthropomorphism, a fudging of the facts.[8] Others insist, on the contrary, that this view implausibly ties the attribution of mental states to a standard that even very intelligent persons only meet imperfectly, and flies in the face of the fact that we routinely attribute beliefs and desires to nonlinguistic animals and to human infants, although these are presumably not rational creatures in the intended sense.[9] But what does not receive much discussion in the current debate is the idea that brute mentality is genuine, but different in kind from rational mentality.[10] If such a conception could be defended, it would allow us to admit that brutes have beliefs and desires in a nonmetaphorical sense, while insisting that what it is for them to have such attitudes differs fundamentally from what it is for us to have them.

What prevents this sort of position from receiving serious consideration, presumably, is that it is far from clear what it could amount to. To insist on a categorial difference is to insist that the things distinguished differ not just in detail or degree, but in kind. But what is a difference in kind, and what application does this idea have to the case at hand? My aim in here is to sketch the general shape of an Aristotelian answer to these questions. Once we have seen this general shape, we will be able to

---

[7] I take it that this capacity is what Kant means by "consciousness" in the remark quoted earlier (note that he uses the term "self-consciousness" shortly afterwards).

[8] This is famously the view of Donald Davidson. See his "Thought and Talk" (1984) and "Rational Animals" (1982). Similar claims, expressed in more moderate language, can be found in Sydney Shoemaker, "Rationality and Self-Consciousness" (1991).

[9] A few casually-selected instances: Jonathan Bennett, *Rationality* (1964), p. 2; Norman Malcolm, "Thoughtless Brutes" (1977); Alva Noë, "Is Perspectival Self-Consciousness Non-Conceptual?" (2002); Tyler Burge, "Perceptual Entitlement" (2003). It is not necessary to multiply examples: one hears this sort of thing constantly.

[10] There are exceptions: views of this sort are defended by John McDowell in his *Mind and World* (1994), Lecture VI and by Christine Korsgaard in her recent Locke Lectures, *Self-Constitution: Action, Identity and Integrity* (2002), esp. Lecture IV.

recognize Kant's account of the difference between rational and nonrational creatures as an instance of it.

*

What could it mean to say that the difference between rational and nonrational creatures is categorial, a difference in kind rather than one of degree? In an essay attempting to analyze what it means to call a creature "rational," Jonathan Bennett offers the following gloss on the idea that human minds differ in kind from those of other creatures:

> It is commonly believed… that between a genius and a stupid man there is a smooth slide while between a stupid man and an ape there is a sharp drop, not just in the sense that there are no creatures intellectually half-way between apes and stupid men, but in the sense that there could not be such creatures. Any possible creature whose intellectual level was higher than that of normal apes and lower than that of normal men—so the common belief runs—either would or would not have that special something which puts humans importantly above other animals.[11]

This characterization of the claim of categorial difference is evocative, but I do not think it helps us much. To deny that there are creatures whose intellects stand half-way between apes and men is just to insist that the rational-nonrational opposition is exclusive: for any given creature, we will say either that it is rational or that it is not, and we will not admit intermediate cases. But this might be true although the opposition in question was simply stipulative, an arbitrary line drawn at a certain point on what is in fact a continuum. Thus we might draw an exclusive distinction between persons over six feet tall and persons of six feet or less, calling the former group "tall" and the latter "non-tall"; but although this distinction might be useful for certain purposes, it clearly would not mark a categorial difference in the intended sense.

Indeed, I think the very contrast between a continuous transition and a sharp break is a distraction. For there are plenty of discontinuous differences that are not categorial in the sense in which the Aristotelian tradition holds the difference between rational and nonrational creatures to be

---

[11] Jonathan Bennett, *Rationality* (1964), p. 4.

categorial. The difference between a steam engine and a gas engine, for instance, presumably involves a sharp break rather than a continuous transition: for what would the relevant continuum be? It is thus natural enough to say that these are different *kinds* of engines, engines that fall into different *categories*. But the distinction is not categorial in the philosopher's sense: it does not mark fundamentally different orders of being, beings whose very modes of existence are different.[12]

This talk of orders of being and modes of existence is, of course, no less obscure than the special emphatic use of the word "category" that is our topic; but some examples should start to bring out the kind of distinction that is of interest. Take, for instance, the distinction between material things and numbers, or the distinction between living and nonliving things. To hold, as the Aristotelian tradition does, that material things differ categorially from numbers, and that living things differ categorially from nonliving things, is not merely to hold that things of the one kind have certain properties that things of the other kind lack, as a gas engine has certain properties that a steam engine lacks. The "properties" that characterize material things as such – endurance through time, the capacity to stand in causal relations, the capacity for motion and change of state – are not characteristics that we *might* find in a number. To say that something is a number is already to identify it as a kind of thing that could not conceivably fall under such predicates as "is moving" or "is at rest," "has changed" or "has remained the same," "still exists" or "has been destroyed." Such predicates belong to a distinctive *kind* of predication, one that places its subject in a temporal order, and it characterizes the mode of existence of a number to say that predicates of this kind do not apply to it. Similarly, the "properties" that characterize living things as such – growth, self-maintenance, reproduction, and so on – are arguably properties associated with a distinctive a kind of predication, one that places its subject in a teleological order, and it

---

[12] I do not mean to suggest that Bennett was trying to describe this sort of difference. His view seems to be that the difference between rational and nonrational creatures *is* like the difference between steam engines and gas engines, at least to this extent: it is possible to characterize rationality as a certain complex configuration of abilities all of which might separately be possessed by a nonrational creature. My point is just that this way of looking at the matter constitutes a departure – not clearly marked by Bennett – from the traditional view. I shall suggest that Kant, unlike Bennett, takes the difference between rational and nonrational beings to be a categorial difference in the traditional

characterizes the mode of existence of a nonliving thing to say that predicates of this kind do not apply to it.[13] A stone is neither healthy nor sick; it does not either grow or fail to grow. If we say that numbers don't move, or that stones don't grow, we are making remarks of a different kind from the remark that a certain plant is not growing: we are saying that numbers and stones are not even *potential* subjects of the respective sorts of predicate.

To hold that two kinds of things differ categorially is thus to hold that one of the kinds is potentially the subject of a kind of predication that has no intelligible application to the other kind. This is not to deny that there can be predicates that apply unequivocally to material things and to numbers, or to living and nonliving things. Both material things and numbers are countable, and both living things and nonliving things can be said to weigh five pounds. But there are also kinds of predicates that can apply to the one kind of thing but not to the other. This is what gives point to talk of different orders of being. To put the point in a way that is anachronistic but useful: a *Begriffschrift* would distinguish terms referring to living things from terms referring to nonliving ones, and predicates of living things from predicates of nonliving ones, so as to make clear that certain kinds of combinations of subject and predicate expressions are simply excluded – for instance, the would-be combinations manifested in "The number 3 is moving" and "This stone is dead."

To investigate the hierarchy of such orders of being, what sorts of predicates are associated with each order, and what the presuppositions of such predications are, would be to perform a task that has some claim to be called "metaphysics." This conception of metaphysics is obviously controversial, and my aim at this point is not to defend it, but simply to indicate it as a possibility. For present purposes, the important thing is just to see what sort of claim a claim of categorial difference is. To claim that there is a categorial difference between material things and numbers, or between living and nonliving things, is to claim that we are dealing here not just with distinct *ranges* of objects but with different *types*

---

sense.

[13] For a defense of such a view, see Michael Thompson, "The Representation of Life" (1995), to which the outlook

of objects, objects characterized as such by their ability to possess certain sorts of properties. The claim I shall want to defend is that the distinction between rational and nonrational beings is a distinction of this sort, and that the capacity for self-consciousness is at the root of this distinction.

*

My reasons for holding that self-consciousness is essential to rational mentality will emerge in the chapters that follow. Before proceeding, however, I want to mention two widespread assumptions which prevent this sort of claim from getting a hearing. Once they are articulated, I think it will be obvious that both assumptions are questionable.

The first assumption is on display in the following passage from the beginning of Stephen Stich's essay, "Could Man Be an Irrational Animal?":

> Aristotle thought man was a rational animal. From his time to ours, however, there has been a steady stream of writers who have dissented from this sanguine assessment… During the last decade or so, [the] impressionistic chroniclers of man's cognitive foibles have been joined by a growing group of experimental psychologists who are subjecting human reasoning to careful empirical scrutiny. Much of what they have found would appall Aristotle. Human subjects, it would appear, regularly and systematically invoke inferential and judgmental strategies ranging from the merely invalid to the genuinely bizarre. (1985, p. 115)

The plain assumption here is that the Aristotelian doctrine that man is a rational animal must be taken as a claim about how most men think most of the time. Otherwise how could it be a threat to the Aristotelian thesis that human subjects regularly and even systematically make invalid inferences or judge questions on unsound bases? But the claim that man is a rational animal was plainly never intended as some sort of statistical generalization. It is clearly meant as a claim about human *nature*, and to say that it is a part of the nature of a certain kind of thing to be thus-and-so is not to say that all or even most members of that kind must be thus-and-so. This point has been emphasized by a number of authors: claims such as "Beavers build dams," "Human beings have eyesight," and "Mayflies breed shortly before dying" might be true although worsening environmental conditions rendered most beavers incapable of

---

developed here is heavily indebted.

dam-building, a nuclear attack made most human beings blind, and it turned out that – as a normal matter, environmental degradation and nuclear attacks aside – most actual mayflies die long before breeding.[14] To make these negative observations is not to give a positive account of the truth-conditions of such "nature statements," but it does suggest that such statements are ubiquitous, and that we must come to terms with them if we are to make sense of talk about living things at all.

We can call the assumption that nature statements must be read as involving some sort of implicit (universal or merely statistical) quantification over individuals, the *Assumption of Individual Generality*. As soon as it is articulated, this assumption looks dubious. Nevertheless, it often underlies objections to views that posit a categorial difference between rational and nonrational creatures. For instance, whenever a philosopher with Kantian sympathies claims that there is an essential connection, in a rational creature, between belief and the ability to produce reasons for belief, or between belief and the ability to self-ascribe belief, or whatever, it is routine to hear other philosophers objecting by counterexample. For surely – the objectors say – rational creatures sometimes find themselves unable to give reasons for what they believe, or find themselves unable to change a belief which they know to be irrational, or even discover by observing their own behavior that they hold beliefs of which they have no direct awareness. And the possibility of such cases, it is argued, shows that it cannot be essential to belief, as it occurs in a rational creature, that it meet the relevant conditions. Hence, it is concluded, the beliefs of rational creatures cannot differ categorially from those of nonrational creatures in virtue of being subject to these conditions.[15]

---

[14] The first two examples are from Julius Moravcsik, "Essences, Powers, and Generic Propositions" (1994). The last is from Michael Thompson, "The Representation of Life" (1995). For discussion of similar points in connection with the claim that human beings are rational, self-conscious creatures, see Brian O'Shaughnessy, *Consciousness and the World* (2002), Chapter 3.

[15] I cannot cite a full presentation of this line of thought, for I do not know of any explicit criticism of the claim that rational creatures differ categorially from nonrational ones: the view is simply not discussed as such. However, cases in which a subject is unable to give reasons for her beliefs or to speak straightaway for what she believes are frequently brought forward in criticism of particular theses about what is involved in having a belief or a reason – theses which are best understood, I want to suggest, as claims of essential connection. For this sort of criticism, see for instance Gilbert Harman, *Thought* (1973), Chapter 2, §2. Other instances of the criticism can be found in the literature cited in footnote

Without denying the possibility of such cases, however, we can question their status as counterexamples. For their possibility shows that the relevant claims of essential connection are false only if such claims must be read as holding of all tokens of the relevant types. But this is just a version of the Assumption of Individual Generality. And again, as soon as it is articulated, the assumption looks suspicious: the claim that it is an essential property of horses to have four legs (or put another way: the claim that there is an essential connection between being a horse and being such as to have four legs) is hardly falsified by the occasional three-legged horse. It would not be falsified by any mere *number* of such horses. Similarly, the claim that there is an essential connection, in rational creatures, between being a belief and being open to self-conscious rational reflection would not be falsified by the existence in a rational mind of any mere number of aberrant beliefs. If the proposition that human beings are rational animals, and the more specific claims of essential connection that articulate the content of this proposition, are claims about human nature, then the way to evaluate these claims is not to ask whether human beings *for the most part* can speak authoritatively about what they believe, or produce reasons for what they believe. Nor is it to ask whether human subjects for the most part believe in accordance with the laws of logic, or choose in accordance with the principles of decision theory. The way to evaluate such claims is rather to ask what kinds of *powers* are being exercised in human thinking, what should count as the *normal* operation of such powers, and what should count as *malfunctions* calling for special explanation.

All of these concepts are obviously in need of explanation, and I will not offer, nor can I claim to possess, a general theory of natures, essences, powers, and functions. One aim of my discussion, however, will be to show in practice what it might mean to take the claim that human beings are rational animals as a nature statement, and how the claim might be defended when thus understood. I shall argue that we rational creatures are distinguished from mere brutes by our possession of the power of judgment, which involves the capacity to reflect on the objective significance of our representations, and

---

9.

which presupposes a correlative power to reflect on ourselves as representers. The claim that these powers belong to human nature is perfectly consistent with the observation that we often fail in their exercise. Indeed, it is consistent with the observation that there are human beings who never attain to these powers at all. For even human beings who lack these powers altogether belong to a species whose form of life involves the exercise of these powers: their minds are thus defective in an important way. A mere brute, by contrast, does not count as defective for want of the capacity to judge.

A second assumption that makes the claim that rational creatures differ categorially from mere brutes look untenable is what we can call the *Uniformity Assumption*. This is the assumption that a psychological or epistemic concept which applies both to rational creatures and to mere brutes must be susceptible of a single, undifferentiating account that covers both sorts of application – that, in other words, the relevant concept must be treated as *uniform* in its application to rational creatures and to mere brutes. This assumption manifests itself in the frequently heard insistence that an account of belief, warrant, knowledge, etc. must not make any demands that a mere brute could not meet, since brutes hold beliefs, possess warrant for their beliefs, have knowledge, and so on.[16] It should be clear that this inference is only valid given the Uniformity Assumption, for only if an account of these concepts must not differentiate between rational creatures and mere brutes does the fact that we speak of brutes as holding beliefs, being warranted, etc., show that our account of the application of such concepts to humans cannot make demands that a brute could not meet.

Must we make the Uniformity Assumption? In his recent "Perceptual Entitlement," Tyler Burge writes:

> Children and higher nonhuman animals do not have *reasons* for their perceptual beliefs. They lack concepts like *reliable*, *normal condition*, *perceptual state*, *individuation*, *defeating condition*, that are necessary for having such reasons. Yet they have perceptual beliefs. There is no sound basis for denying that epistemology can evaluate these beliefs with respect to norms governing their formation, given the perspectival limitations and

[16] Again this way of thinking is a legacy of British Empiricism. Thus Hume holds that "[w]hen any hypothesis… is advanc'd to explain a mental operation, which is common to men and beasts, we must apply the same hypothesis to both" (*Treatise* I, III, XVI; 1978, p. 177).

> environmental conditions of the believer. There is no sound basis for denying that epistemology can evaluate their perceptual beliefs for epistemic warrant. (2003, p. 528)

I think there is a reading of what Burge says here on which it is undeniable: animals patently respond to the world on the basis of representations of what is the case, representations that we have every reason to call "beliefs," representations concerning which we can certainly raise questions of warrant. But Burge makes these points in the context of attacking the Sellarsian thesis that a rational creature's warrant for perceptual belief must "lie within 'the space of reasons'" (Burge 2003, p. 526). Indeed, he writes as though these points themselves constituted a refutation of the Sellarsian thesis: if nonhuman animals can have beliefs, and be warranted in having them, then, Burge reasons, the notion of warrant cannot be restricted to a space in which nonhuman animals have no part.

But surely this inference reflects a blinkered view of the options. Whatever exactly it means to claim that a rational creature's warrant must lie within the space of reasons, a sensible Sellarsian should *not* hold this to entail that children and nonhuman animals cannot have perceptual beliefs, or cannot be warranted in having them. He should hold, instead, that the concepts of belief and warrant have a different application here from the one they have in connection with rational creatures – that with the dawn of reason comes a new form of belief and new standards of warrant associated with it. To hold that concepts such as *belief* and *warrant* apply in one way to rational creatures and in a different way to mere brutes is not to suggest that we are merely being ambiguous when we speak of "belief" and "warrant" in connection with creatures of both kinds. The suggestion is that brute belief and rational belief, brute warrant and rational warrant, are different *species* of the same *genus*. Once again, this sort of point is familiar from the Aristotelian tradition. Aristotle is constantly remarking on the fact that a single abstract concept is capable of taking on various different, more determinate significances in connection with different kinds of things. He says this, for instance, about the concepts *cause*, *movement*, and *life*.[17] In each case, he holds that the several more determinate significances have some abstract structure in

common. Nevertheless, he maintains that in applying some determinate form of one of these determinable concepts, we are saying something more specific than what is specified by the abstract structure.

Consider, for instance, the concept of an agent, in the sense in which to call something an agent is opposed to saying that it passively undergoes a change. Aristotle holds that anything whose principle of change is internal to it can be said to be an agent. Hence any substance can be said to be an agent insofar as its undergoing a certain change is a result of its own nature.[18] When we speak of a living thing as an agent, however, and say that certain changes *it* undergoes result from its own nature, we are clearly speaking of agency in a richer sense, one that is a particular specification of the more generic concept. A nonliving material substance does not act in *this* sense, nor does it even make sense to suppose that it might. And an even richer form of the concept *agency* applies only to animals, and a yet richer one only to rational animals. This hierarchy of forms of the concept *agency* is just one reflection of the fact that, on Aristotle's view, living things differ categorially from nonliving ones, animals differ categorially from inanimate living things, and rational animals differ categorially from mere brutes.

For the moment, my aim is not to defend these claims, but merely to note their possibility. This is at least an intelligible sort of position to take, whether or not it is defensible in the case at hand. But what I am calling the Uniformity Assumption in effect preempts the possibility of such a position with regard to concepts such as *belief*, *knowledge*, *inference*, *warrant*, and so on. Yet surely the claim that these concepts have a different and more stringent meaning in application to rational beings from the one they have in application to mere brutes at least deserves a hearing. If such a position were vindicated, then, despite what Burge says, we *would* have sound basis for denying that epistemology can evaluate the perceptual beliefs of children and nonhuman animals for epistemic warrant *in the sense of warrant that is*

---

[17] For the claim about the concept of cause, see *Physics*, II, §§3-7. For movement, see *On the Soul,* 406a13-14. For life, see *On the Soul,* 413a21-25.

[18] For this general account of agency, see *Physics*, II-III.

proper to rational beings.

In the chapters that follow, I shall be arguing that the difference between rational and nonrational beings *is* categorial: that talk of belief, of mental content, and of knowledge, has an irreducibly different application in connection with rational creatures from the one it has in connection with brutes. And I shall claim that self-consciousness, the capacity to have the representation *I*, is both the ground and the purest expression of this difference.

2. THE CONTEMPORARY RELEVANCE OF KANT'S POSITION

At this point, the idea of an irreducible difference between the application of talk of belief, mental content, and knowledge to rational beings and its application to nonrational creatures can only stand as a cryptic promissory note. Already, however, the interest of this idea should begin to be apparent. For such a view would rule out a certain kind of reductionist project: the project of saying what it is to believe something, or know something, or have a representation with a certain content, in the sense proper to rational beings, in terms that are supposed to be capable of applying unequivocally to nonrational creatures. It would rule out this sort of project because, if such talk has a special and irreducible application to rational beings – an application whose very intelligibility presupposes the capacity for self-consciousness – then a rational creature's beliefs cannot be *identified* with states to which self-consciousness is not essential. And nor can the fact that those states have the content they do, or the facts in which their being warranted consists, be identified with facts that have no necessary relation to self-consciousness.

To suppose that such identifications *are* possible is, in effect, to accept a thesis along the following lines:

(T) The fundamental elements out of which rational mentality is constructed are elements that can be present without the capacity for self-consciousness.

A philosophy of mind that accepts (T) will be forced to explain what it is to be self-consciously aware of one's own mental states in terms of some configuration of kinds of states which can exist in the absence of the capacity for self-consciousness. I call such views *tack-on theories of self-consciousness*, for they represent the capacity for self-consciousness as an optional addition to a creature's mind, in the absence of which that mind would remain otherwise essentially the same. It seems to me that a number of influential programs in contemporary philosophy of mind are committed to giving tack-on theories of self-consciousness. To see this commitment, and how Kant's view constitutes a challenge to it, it will be useful to consider a case in point.

David Lewis's influential version of functionalism is a particularly plain case of a view that is committed to giving a tack-on theory of self-consciousness. The general lines of Lewis's view are familiar: mental states are identified with whatever states turn out to "realize" (or most nearly realize) the "causal roles" characterized by the set of platitudes that make up "common-sense psychology" – platitudes about how mental states relate to one another, to sensory stimuli, and to motor responses, platitudes that are supposed to be part of the common knowledge possessed by anyone who understands the meaning of terms like "belief," "desire," "pain," and so on. Building on work done by Frank Ramsey, Lewis shows how functional definitions of individual mental states can be derived from such platitudes.[19] For my purposes here, however, it is not necessary to consider the details of this derivation. To see Lewis's commitment to (T), we need only reflect on how his view seeks to account for our privileged awareness of our own mental states.

It might seem that a view which characterizes mental states in terms of their functional roles in mediating between sensory inputs and behavioral outputs is committed to denying any special significance to the fact that a subject believes herself to be in a given mental state. For what is important, on this sort of view, is that the subject actually *be* in a state that plays a certain role, not that

---

[19] For a succinct statement of the approach, see Lewis's "Psychophysical and Theoretical Identifications" (1972).

she *believe* that she is in the kind of state that fulfills that role. Thus it might appear that functionalist theories must deny any special privilege or authority to a subject's beliefs about her own mental states. According to Lewis, however, this appearance is false:

> Suppose that among the platitudes [that constitute common-sense psychology] are some to the effect that introspection is reliable: 'belief that one is in pain never occurs unless pain occurs' or the like. Suppose further that these platitudes enter the term introducing postulates as conjuncts, not as cluster members; and suppose that they are so important that an *n*-tuple that fails to satisfy them perfectly is not even a near-realization of common-sense psychology. Then the necessary infallibility of introspection is assured. Two states cannot be pain and belief that one is in pain, respectively (in the case of a given individual or species), if the second *ever* occurs without the first (1972, p. 258).[20]

In short, the suggestion is that a functionalist theory can recognize a privileged role for introspection simply by treating *the tendency to cause higher-order beliefs to the effect that one is in a certain state* as an essential component of the role that defines the relevant state. And, at least on first inspection, including this sort of component in the functional definition of a state *does* seem to make presence to self-consciousness essential to the identity of that state. For if the functional definition of *S* requires that being in *S* give rise to the belief "I am in *S*," then no state can be identified as *S* which does not give rise to the relevant belief. What closer connection could a state have with self-consciousness?

But the connection for which Lewis's account provides is not really so close. We can begin to see this by observing that, although such an account does make a connection to self-consciousness essential to the identity of *S* in one sense, there is another sense in which having the belief "I am in *S*" remains distinct from being in *S*. If this is not obvious, consider the role that supposedly defines *S*: let us call it *R*. If we delete from the specification of *R* whatever clauses tie the presence of *S* to the belief "I am in *S*," we are left with a specification of a simpler role, which we can call *R**. On the functionalist view, there can be no reason why a creature could not in principle be in the kind of state characterized

---

[20] An "*n*-tuple" is a set of *n* causally-interacting physiological states, concerning which we can ask how near they come to realizing the set of causal roles defined by the platitudes of common-sense psychology. If a platitude *P* enters the term introducing postulates as a conjunct, then a state that does not have the causal role specified by *P* does not satisfy the relevant definition. If *P* enters as a cluster member, by contrast, then a state might count as satisfying the definition even though it did not play the role specified by *P*; for a definition involving clusters is one that consists of a disjunction of

by *R** (call it *S**) while not being in the kind of state that plays the more complex role *R* (namely, *S*). And even in the case of a subject whose mental apparatus *is* set up in the way that allows it to be in *S*, still it should be clear that all this really amounts to is the capacity to be in *S** plus the capacity to form beliefs about her own states that reliably track the presence of *S**.[21] There is, in short, a clear sense in which what self-consciousness amounts to, on Lewis's view, is just the ability to form higher-order beliefs that reliably track the presence of a distinct sort of state, a state that might in principle exist in the absence of self-consciousness. Lewis is thus committed to accepting a form of (T), and to giving a tack-on theory of self-consciousness. And he is hardly alone in this commitment: many authors explicitly advocate such "higher-order thought" theories of self-consciousness, and many more advocate, on general grounds, theories of mind that presumably must account for self-consciousness in more or less this way.[22] I focus on Lewis's account only because it is a particular lucid and simple instance of this approach.

My reasons for rejecting this sort of view will emerge in the chapters that follow. For the moment, I only want to indicate how the conception of self-consciousness that I shall be defending contrasts with the conception just described. To see the contrast, it is helpful to consider how Lewis's account would look when applied, not to the state of pain, but to a state like believing that *p*.[23] On

---

conjunctions of most of the platitudes, and the specific platitude *P* might figure in some but not all of the conjunctions.

[21] Compare the discussion of "weak special access theories" in Elizabeth Fricker, "Self-Knowledge: Special Access versus Artefact of Grammar —A Dichotomy Rejected" (1998; see esp. p. 182). For simplicity's sake, I am ignoring questions about what gives the relevant higher-order beliefs their contents: on the one hand, what makes their subject first-personal; and on the other hand, what ensures that the mental state they concern is the state with which they are reliably correlated. I think careful consideration of these questions would only strengthen my case.

[22] Prominent advocates of higher-order thought theories include David Armstrong (see Armstrong 1968), David Rosenthal (see Rosenthal 1986) and William Lycan (see Lycan 1998). Anyone familiar with contemporary philosophy of mind should recognize that a large number of positions are committed to giving an account of self-consciousness of the same *general* shape as Lewis's: namely, one that takes self-conscious awareness of one's own mental states to involve the existence of a higher-order state distinct from the state that is the object of awareness. Of course the story about the relationship between the first-level state and the higher-level state of awareness can be more complex: see Shoemaker 1990, §IV for an indication of what a more complex functionalist account of self-consciousness might look like.

[23] The features of the Kantian conception of self-consciousness that I will be emphasizing apply as much to self-conscious awareness of pain as to self-conscious awareness of belief, but I think their plausibility comes out most strikingly in the case of belief, precisely because it is clear that, in self-consciously reflecting on what I believe, I can *make up my mind* about what to believe. Self-conscious awareness of pain does not in this way constitute pain, although I still

Lewis's view, my capacity for self-conscious awareness of my belief that *p* must presumably consist in the fact that, when I believe that *p*, I am also reliably disposed to form the belief that I believe that *p*. To anyone predisposed to sympathy for the view I shall be defending, however, this will seem to put the cart before the horse. For Lewis's account makes my ability to know whether I believe that *p* a matter of my being able to form higher-order beliefs that reliably *track* my own first-order beliefs; whereas, on the Kantian view that I shall be defending, my self-consciously judging "I believe that *p*" can, in the right circumstances, be *my making up my mind* to believe that *p*. And when this is so, it will not be merely a matter of my judgment "I believe that *p*" tending to *cause* a distinct state of believing that *p* to come into existence (as a functionalist who had some sympathy for the Kantian view might hold); my judgment will itself *constitute* my coming to believe that *p*. For the crux of the Kantian account is just this: that in a self-conscious creature, being in a certain mental state and being self-consciously aware of that state are not distinct, causally-interacting states. Rather, self-conscious belief is just an attentive instance of belief.[24]

Without yet attempting to argue for this conception of self-conscious belief, let me at least indicate where a criticism of Lewis's account might begin. The first order of business would be to think more carefully about the supposed platitudes of common-sense psychology. Lewis suggests that the relevant platitudes might take the form:

> When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so motor responses (1972, p. 256).

But it takes only a little reflection on the kinds of things we know about the relation between sensory inputs, mental states, and behavioral outputs to see that this form calls for rather more information than we typically possess. For one thing, the language of "sensory stimuli" and "motor responses" tends to suggest that our platitudinous knowledge about human psychology includes information about the

---

think, *contra* Lewis, that self-conscious awareness of pain does not involve a state of awareness distinct from pain itself.

[24] We can thus think of the "self-conscious" in "self-conscious belief" as functioning adverbially, indicating what is fundamentally a *way* of believing something, not a *relation* in which a person stands to his own belief. For the idea of an adverbial treatment of such modifiers, see Moran 2001, pp. 30-31.

relation between mental states, on the one hand, and inputs and outputs *described in ways that do not employ the language of common-sense psychology*, on the other. Yet this seems implausible: what we know is that, e.g., a person who wants a drink and sees a glass of water in front of her will tend, *ceteris paribus*, to reach for the glass. Both the sensory input and the behavioral output here are plainly described in intentional language. And the *ceteris paribus* clause marks another difficulty: it seems that our platitudinous knowledge about human psychology does not involve any detailed idea of probabilities, and is always relative to a background of conditions that must be "held equal," which conditions we are hardly in a position to spell out. These are familiar difficulties for functionalism, and my purpose here is not to press them. I only want to stress that, if "common-sense psychology" includes platitudes about how mental states are connected with sensory inputs and behavioral outputs, they will be such things as:

(P) If a subject wants a drink and sees a glass of water in front of her, this will tend, *ceteris paribus*, to cause her to reach for the glass and drink from it.

But now what sort of knowledge do we have of platitudes like (P)? Do we know them as inductive generalizations from particular cases? Surely not: the suggestion that my knowledge of the whole body of such platitudes reflects my personal *experience* with what people tend to do given certain beliefs and desires is obviously absurd. And on reflection, I think the frequently-heard suggestion that such platitudes reflect "humanity's accumulated wisdom" about how people tend to act given certain beliefs and desires should seem equally absurd. For if, as is widely admitted, the very concepts of belief and desire have their significance only in relation to a "constitutive ideal of rationality," an ideal that characterizes any given intentional state in terms of its role in the rational mediation between sensory intake and behavioral response, then our knowledge of principles like (P) cannot just reflect our accumulated observations of how things tend to go with people: it must reflect an understanding of how things *ought* to go – a knowledge of what is rational.

How then *do* I know facts like (P)? I think the obvious answer is this: the connection posited in (P) *makes sense* to me when I consider it from the first-person standpoint. I know such facts, that is, by being able to think about *what I would choose to do* under the circumstances (e.g., if I wanted a drink and

saw a glass of water within reach). This is plainly not inductive knowledge about how I am *likely* to behave when I am in certain states; it is knowledge I have by seeing that, absent other pressing considerations, reaching for the glass would be the reasonable thing to do, and consequently would be what I would *choose*, in self-conscious reflection, to do.[25] But if this is right, then my knowledge of the "platitudes" of common-sense psychology – platitudes whose identifiability is required for Lewis's functionalist approach to get off the ground – depends on my conviction that *my self-conscious reflection about what to do can constitute my choosing to do something.* And similarly, where it is a question of knowing a platitude about belief such as

(P*) If a subject believes that if *p* then *q* and notices that *p*, this will tend, *ceteris paribus*, to cause her to form the belief that *q*

my claim to know this will again depend on the assumption that *my self-conscious reflection on what to believe can constitute my coming to believe something.* For only if I am entitled to assume that a rational creature's self-conscious reflection would, in general, settle the question of what it believes, or what it chooses, can I be entitled to regard my conclusions about what would be reasonable as knowledge of what would take place.

This is very quick, and I do not expect it to convince anyone who is committed to an opposing view. But my purpose here is only to prepare the ground for the coming chapters by raising some suspicions about something anti-Kantians tend to take for granted, namely that whatever connections exist between being in a given mental state and being self-consciously aware of that state can be treated as just a few more platitudes that form part of our "folk theory" of human psychology. If I am right,

---

[25] This point is helpfully discussed in Richard Moran's "Interpretation Theory and the First Person" (1994). According to Moran, our ability to explain and predict how people will act on the basis of their beliefs and desires can be neither a matter of pure theory nor a question of mere "simulation":

> [N]ormally a person has very little of an inductive, norm-free basis for predicting his own future or hypothetical behavior. In the ordinary case, involving something more than sheer conjecture, he can answer such first-person questions with confidence only because he can make a here-and-now decision about what course of action to endorse. And without the normative basis he loses the rationale for applying this same prediction to another person, for he cannot assume that whatever inductive basis he had for his first-person prediction would apply to anyone else (1994, p. 163).

none of the genuine platitudes of common-sense psychology can be regarded as theoretical, precisely because they are all facts that we know through self-consciousness. And this means that claims about the connection between, e.g., belief and self-conscious awareness of belief cannot be just a few more such platitudes.

*

I have taken Lewis's causal-role functionalism as my primary example of a view that is committed to giving a tack-on theory of self-consciousness, but recent philosophy of mind abounds with examples of views that take on this commitment. I could have focused instead on Wilfrid Sellars's "Myth of Jones," in which the vocabulary of common-sense psychology is first introduced as a theory of internal, behavior-causing states, and only subsequently given a nonobservational self-ascriptive use.[26] I could have focused on attempts to give a naturalistic account, not of the nature of mental *states*, but of the content of mental *representations*; for such accounts characteristically presuppose that we can say what it is for a representation to have a certain content in terms that do not make reference to a capacity for self-consciousness, and then can build up from such materials to representational content of the kind that mental states of a self-conscious creature possess.[27] Or I could have focused on attempts to say what it is for a self-conscious creature to be *warranted* in believing something in terms of factors which have no necessary relation to self-consciousness.[28] What all such programs have in common, despite their different topics and their diverse forms, is a commitment to (T): they all assume that the basic elements out of which our mental life is constructed are elements that can be present without the capacity for self-consciousness.

---

[26] See Sellars 1963a, Parts XII-XV. Lewis himself likens his view to Sellars's at Lewis 1972, p. 257.

[27] This is presupposed, for instance, by Fred Dretske's influential account of representational content in his *Knowledge and the Flow of Information* (1981). It also seems to be presupposed by Robert Brandom's avowedly non-naturalistic account of the constitution of assertional content in his *Making It Explicit* (1994). According to Brandom,

> there need be nothing incoherent in descriptions of communities of judging and perceiving agents, attributing and undertaking propositionally contentful commitments, giving and asking for reasons, who do not yet have available the expressive resources 'I' provides (1994, p. 559).

[28] Dretske 1981 is again an example here, as is any "externalist" account of knowledge that seeks to sever all connection

If Kant is right, then all such attempts to give a reductive account of the states and statuses characteristic of self-conscious mentality are futile. For either the terms in which the reductive account is couched can apply unequivocally to a creature that lacks the capacity for self-consciousness, or they cannot. If they can, then Kant will hold that the states and statuses described in the account cannot be identified with the states and statuses of a self-conscious creature, for the capacity for self-consciousness is not *essential* to these states and statuses. But if they cannot – if the significance of their application in the present instance can only be explained in a way that involves reference to the capacity for self-consciousness – then the reduction is pointless, for the terms of the reducing theory can only be explained by appeal to the terms supposedly being reduced.

3. OVERVIEW OF THE COMING CHAPTERS

In the six chapters that follow, I first present an interpretation and defense of Kant's claim about the importance of self-consciousness, and then draw lessons from this discussion for two contemporary debates.

The interpretation and defense takes up the first four chapters. I begin, in Chapter I, by presenting an interpretation of and an intuitive motivation for the claim I call *the Kantian thesis*: Kant's claim that "It must be possible for the 'I think' to accompany all of my representations" (B131). The next three chapters offer a more detailed defense of this claim, drawing on Kant when it seems helpful, but aiming primarily to give arguments that are plausible in their own right. In Chapter II, I argue that a *self-conscious* believer must in general be capable of calling to mind her own grounds for belief, since, except in special cases, a subject who is aware that she lacks satisfactory grounds for a given belief cannot retain that belief. Then, in the next two chapters, I defend the claim that a *rational* believer must

---

between warrant and accessibility to reflection. I say more about externalist accounts of knowledge below in Chapter II.

be self-conscious. This involves, on the one hand, an investigation of the kinds of capacities that characterize a rational believer, one capable of conceptual thought (Chapter III), and, on the other hand, an argument that connects these capacities with self-consciousness (Chapter IV). I argue that (1) to represent *conceptually* is to represent in a way that decouples information from any particular context or purpose, (2) this special form of representation is possible only in a creature that can reflect explicitly on grounds for judging a proposition true, and (3) to have this capacity to reflect explicitly on grounds is necessarily to have the crux of self-consciousness. The capacity to represent conceptually thus requires the capacity for self-consciousness. It follows that the beliefs of a rational creature – and also, by the argument of Chapter II, whatever representations can figure as a rational creature's grounds for belief – are states to which the capacity for self-consciousness is essential, and hence cannot be identified with states that might occur in a creature which lacked self-consciousness.

The dissertation concludes with two chapters that apply ideas from the preceding discussion to issues in the wider philosophy of mind. In Chapter V, I argue that much contemporary discussion of "first-person authority" is hampered by a failure to distinguish between the kind of self-knowledge supplied by what Kant calls "inner sense" and the kind supplied by what he calls "pure apperception." I show that the latter kind of self-knowledge, the kind expressed in our capacity to accompany our representations with "I think," is in an important sense the fundamental kind, and that until we recognize its priority, we can have no satisfactory understanding of our authority in speaking about our own minds, or of the nature of the states with respect to which we have such authority. In Chapter VI, I discuss the phenomenon of self-deception, arguing that someone who accepts my claims about the centrality of self-consciousness need not deny the existence of this sort of phenomenon, and that the considerations which speak in favor of the Kantian thesis actually help to explain what makes this phenomenon seem puzzling, and why it is necessarily the exception to the rule.

I. THE KANTIAN THESIS

> That man can have the representation *I* raises him infinitely above all the other beings living on earth. By this he is a person; and by virtue of his unity of consciousness through all the changes he may undergo, he is one and the same person – that is, a being altogether different in rank and dignity from things, such as nonrational animals, which we can dispose of as we please. This holds even if he cannot yet say 'I'; for he still has it in mind. So any language must think 'I' when it speaks in the first person, even if it has no special word to express it. For this power (the ability to think) is understanding.
>
> Kant, *Anthropology*, I, §1 (7:129T)

1. INTRODUCTION

A striking feature of Kant's philosophical outlook is the importance he attaches to our capacity to think first-personally, to have "the representation *I*."[1] In the passage quoted above, which forms the first paragraph of his *Anthropology*, he identifies this as *the* distinctive capacity of human beings. Our ability to think first-personally, he claims, is what makes us persons, what sets us apart from irrational animals. Indeed, he seems to suggest, it is only because we possess this capacity that we are able to think at all.

Kant is certainly not the first philosopher to reflect on our capacity to conceive of ourselves as "I"; his interest in first-person thought has precedent, notably in Descartes. But Kant's account of what it is for a creature to think first-personally, and what role the representation *I* plays in cognition, is more developed than Descartes's, and he seeks to explain the significance of first-person thought in a way that avoids some of the most objectionable features of Descartes's view. Thus, whereas Descartes suggests

---

[1] What Kant actually says is "*Daß der Mensch in seiner Vorstellung das Ich haben kann*" – literally, "That man can have the I in his representing." Constructions of the form "*In meiner Vorstellung ist es…*" are often rendered "As I imagine it…" or "As I conceive it…", so an idiomatic rendering of Kant's claim might be "That man can conceive of the I." I have

that "the I that thinks" must be an immaterial substance, essentially independent of any body, Kant denies that any such conclusion can be drawn about the referent of "I" in the judgment "I think." And whereas Descartes seems not to distinguish clearly between *having* a certain thought and *thinking that* one has that thought[2] – seems not to distinguish, as we might put it, between unreflective mental episodes and episodes accompanied by self-consciousness – Kant explicitly recognizes that not every mental event need actually be accompanied with self-consciousness in a famous passage of the Transcendental Deduction:

> It must be possible for the "I think" to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me (B131-2).

In emphasizing that it must be *possible* for me to accompany any of my representations with "I think," Kant evidently means to leave room for representations which I do not in fact bring to reflective self-consciousness. Nevertheless, he maintains, it is essential that a thinker should have the capacity to reflect on any of her representations self-consciously.

The different conclusions that Descartes and Kant draw about the application of the idea *I* are connected with a difference in the kind of scrutiny they give to this idea. This latter difference might be expressed as follows: although Descartes is interested in the sort of thing "an I" is, Kant is the first philosopher to give serious attention to the fact that it is an *I*. For although Descartes conducts his investigation into the mind and what it is capable of knowing *in* the first person, he does not ask *what it is*

---

opted for a rendering that brings out the use of the term *Vorstellung* (representation), since this term will be important in what follows; but I have not opted for the literal rendering, which seems to me only to obscure Kant's meaning.

[2] In claiming that Descartes does not distinguish clearly between having a certain thought and thinking that one has that thought, I do not mean to suggest that he simply identifies the two. I mean that his account of mentality does not provide resources to make good sense of the distinction. This failure shows up, for instance, in his notorious claim that "we cannot have any thought of which we are not aware at the very moment when it is in us" (1984, Vol. II, p. 171). As far as I can see, Descartes never explains the notion of awareness at issue in this claim. Does being aware of one's thoughts involve *thinking that* one has them? If not, what does it consist in, and how should we locate the distinction between having a thought and thinking that one has that thought in relation to it? Kant's advance over Descartes, in my view, is that he gives us the tools to begin to get our bearings in this area.

for a creature to think of itself in the first person, what is distinctive of this *way* of thinking of oneself. Kant, by contrast, explicitly takes up this question, and in so doing he takes a significant step beyond Descartes. Once this step is taken, it becomes possible to ask what relationship there is between the capacity for first-person thought and the capacity for thought generally. Kant's answer, evidently, is that these two capacities, while not identical, are importantly interdependent. This characterization of his view, although vague, should begin to bring out the interest of his position – the challenge it poses to anyone who thinks we can explain what it is to be a thinking being in terms of mental goings-on whose occurrence does not require that the subject undergoing them have a capacity for self-consciousness.

My aim in the first part of this dissertation (Chapters I-IV) is to understand what sort of interdependence Kant thought there was between the capacity for first-person thought and the capacity for thought generally, and to consider what reasons he gives us for believing that such an interdependence obtains. I begin with Kant not to honor his historical importance, but because I think that we can still learn from his discussion of these matters. I think he had insights in this area with which contemporary writers have not yet come to terms, insights which contemporary conceptions of self-consciousness in fact tend to obscure. My aim is to bring out these insights, and – to the extent that this is possible – to express them in an idiom that allows them to be evaluated independently of other major commitments of Kant's thinking. Thus, although my discussion will involve exegesis, my aim is not finally exegetical: my ultimate concern is not to discover why *Kant* believed that the capacity for first-person thought and the capacity for thought generally are interdependent, but what reasons *we* can find in his texts for believing such a thing.

I take the recently quoted passage from B131-2 to express Kant's core view about the significance of first-person thought. In claiming that this passage presents Kant's core view, I of course do not mean to deny that he holds other views, both critical and constructive, about the significance of "the representation *I*". Kant's account of the self and self-consciousness involves many distinguishable

theses. We can see several of these theses at work in the passage from Kant's *Anthropology* quoted at the head of this chapter. In that passage, Kant seems to claim:

(1) that our capacity to have the representation *I* is what makes us persons;

(2) that the fact that we can think of ourselves, without equivocation, as one and the same *I* – the fact that the various moments of our consciousness stand together in this kind of unity – is what makes us one and the same person throughout the various changes we undergo;

(3) that beings capable of having the representation *I* possess a special "rank and dignity" that sets them apart from "things... which we can dispose of as we please";

(4) that our capacity to have the representation *I* is somehow implicitly at work in our ability to understand language;

(5) that the capacity to have the representation *I* is essentially connected with the ability to think, which is understanding.

It is not immediately clear how these various claims are meant to be related. One thing that does seem clear, however, is that Kant regards thesis (5) as crucial to the explanation of the other theses. In the case of thesis (4), Kant explicitly claims that it depends on (5): the reason why the capacity to have the representation *I* is implicitly at work in our ability to understand language, apparently, is that the capacity to understand language requires possession of the faculty Kant calls "understanding," and, by (5), possession of this faculty requires a capacity to have the representation *I*. The relationship between theses (1)-(3) and theses (4) and (5) is less clear, but the flow of the paragraph suggests that thesis (5), in particular, should help to clarify why possession of the representation *I* gives us a special rank and dignity, why the distinction between persons and other things marks a significant difference. Thesis (5) thus seems to express one of Kant's core commitments, a doctrine which lies at the root of many of his other views about the importance of the representation *I*. Of course, this does not imply that someone who accepts (5) must accept all the other views Kant regards as consequent upon it. It may be that these follow from (5) – if they follow at all – only given further premises.

I am interested in the claim Kant seems to regard as fundamental, the claim that the capacity to have the representation *I* is essentially connected with the ability to think, which is the defining

characteristic of a creature with understanding. This, I would argue, is precisely the claim being explained and defended in B131-2. For B131-2 asserts that it must be possible for me to accompany all my representations with "I think", and the context of the passage makes clear that this requirement applies to just those creatures which possess the faculty of understanding.[3] Thus it seems that we could restate Kant's claim as follows:

(KT) Any creature that possesses the faculty of understanding, the faculty exercised in thinking, must be able to accompany all of its representations with "I think".

This is the claim whose interpretation and justification I want to explore. I will call it *the Kantian thesis*.

2. INTERPRETING KANT'S REQUIREMENT

The Kantian thesis poses obvious problems of interpretation: What is a representation? What is it to "accompany" a representation with "I think"? In what sense must it be possible for a subject to do this? Despite a profusion of recent work on Kant's theory of mind, these sorts of questions have received surprisingly little attention.[4] The aim of this section is to bring out some of the interpretative questions

---

[3] The context is as follows: B131-2 occurs near the beginning of the second-edition version of the Transcendental Deduction. In this section of the *Critique*, Kant argues from principles that are supposed to articulate necessary conditions of any exercise of the faculty of understanding to the conclusion that certain basic concepts of the understanding must have application to any intuition we might have. The claim that it must be possible for the "I think" to accompany all my representations is one of the principles on which the argument depends – indeed, Kant says that it is the most basic principle, "the supreme principle of all use of the understanding" (B136).

[4] Recent works in English on Kant's theory of mind include: Karl Ameriks, *Kant's Theory of Mind* (Second Edition, 2000); Andrew Brook, *Kant and the Mind* (1994); Pierre Keller, *Kant and the Demands of Self-Consciousness* (1998); Patricia Kitcher, *Kant's Transcendental Psychology* (1990); and C. Thomas Powell, *Kant's Theory of Self-Consciousness* (1990). All of these books except Ameriks's give a central place to Kant's views on self-consciousness, and Ameriks has discussed the topic in several recent papers (see especially Ameriks 1994). A list of recent articles that have discussed Kant's views on self-consciousness would be unmanageably large. It would certainly have to include the sizeable German literature that has developed around the question whether Kant subscribes to a "reflection theory of self-consciousness" and whether such a theory is objectionable. The foundational papers in this debate (both now available in English translation) are Dieter Henrich's "Self-Consciousness: A Critical Introduction to a Theory" (1971) and "Fichte's Original Insight" (1982).

I have consulted various of these works, and have learned things from them, but for the most part their concerns seem orthogonal to mine. A central issue in this literature, for instance, concerns whether Kant provides a coherent and attractive alternative to Descartes's conception of the self as an immaterial substance, and to Hume's picture of the self as a mere bundle of perceptions. Authors who focus on this issue tend to concentrate mainly on the

raised by the Kantian thesis and to canvass possible answers to them. My goal at this point is not to decide what thesis Kant should be read as putting forward, but to distinguish several different theses he might be read as advocating, so that we can consider which seems most defensible.

2.1 TO WHOM DOES THE REQUIREMENT APPLY?

Kant states his requirement, and argues for it, in the first person:

> It must be possible for the 'I think' to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me.

He evidently takes this to be an argument that each of his readers could repeat in his or her own case. But this leaves the scope of the point unclear, for it is not clear what capacities a creature would need to have in order to put this argument to itself; and in any case it is not clear whether the point applies *only* to creatures capable of putting the argument to themselves, or whether its application is meant to be more general. One interpretative question we need to consider, then, is: what must a creature be like if Kant's requirement is to apply to it?

*

The requirement is sometimes read as a claim about what capacities a creature must have if it is to possess mental states deserving to be called "representations."[5] This might be called the *universal reading* of the Kantian thesis, since it would hold that

---

Paralogisms, and to read the Deduction in light of these later concerns. But although I concede the importance of the question how Kant thinks about the self, I think we can only reach an adequate understanding of his views on this topic if we begin by considering his ideas about what function first-person thought plays in the life of a thinking being. The latter question is my topic in this chapter.

[5] Thus Jonathan Bennett reads Kant as claiming that "every representation must occur not just in some mind but specifically in the mind of a self-conscious or self-aware being" (1966, p. 104). Bennett immediately rejects this claim as absurd and seeks to defend a more limited thesis on Kant's behalf. More recently, however, Susan Hurley has argued for a thesis which she takes to have a Kantian lineage, and which seems to be essentially the same as the thesis Bennett rejects: the thesis, as she puts it, that "consciousness requires self-consciousness" (1998, p. 136). For more on Hurley, see footnote 6.

(UR) *Any* creature, if it is to count as having representations at all, must be capable of accompanying all of its representations with "I think".

Put thus universally, however, the requirement looks too demanding. The natural objection to (UR) is that it imposes an implausibly high standard: one that would rule out, as beings capable of having representations, creatures that obviously should not be ruled out of this class. For isn't it plain that animals and infants, although they seem incapable of the sort of higher-order reflection involved in accompanying representations with "I think", can nevertheless have representations? No doubt the answer to this question will depend on what we mean by "representations" – a topic I shall come to presently – but the danger for the unlimited reading of the Kantian thesis is that, if the sense of the term "representation" is adjusted sufficiently to exclude the mental states of animals and infants from consideration, the resulting account will be committed to denying that animals and infants are capable of forms of world-directed mentation of which they obviously are capable.[6]

In any case, there is strong textual evidence that Kant did not mean to deny that animals and infants have representations. Although he does not often discuss the mentality of animals and infants, he does mention this topic occasionally, and on these occasions he consistently maintains that such creatures can and do have representations. For instance, in a much-discussed letter to Marcus Herz,

---

[6] One way to escape this objection while maintaining the universal reading of the Kantian thesis would be to weaken the requirement of accompaniability with "I think" until it looked like a standard which animals and infants *could* meet. This, in effect, is the strategy pursued by Susan Hurley in Chapter 4 of her book *Consciousness in Action* (1998). Hurley takes Kant's thought to be that only self-conscious creatures can have conscious states, and she seeks to reconcile this thought with the observation that animals and infants lack the conceptual resources to ascribe conscious states to themselves by suggesting that there are forms of self-consciousness which do not require a capacity explicitly to self-ascribe mental states. The resulting understanding of the Kantian thesis, however, strikes me as both unfaithful to Kant and unpromising in its own right. It is unfaithful to Kant inasmuch as Kant is plainly concerned with our capacity to think thoughts whose content is explicitly self-ascriptive – thoughts involving the representation *I*. It is unpromising in that it leaves the Kantian thesis without evident motivation. One striking fact about Hurley's defense of her version of Kantianism is that she only defuses objections to the view; she says nothing about what recommends it. Now, I do not wish to deny that there are intelligible senses in which animals and prelinguistic infants may be said to exhibit self-consciousness. But I think there is a reason for Hurley's silence on the question why animals and prelinguistic infants *must* exhibit self-consciousness in some such lenient sense: namely, that the strongest considerations in favor of Kantianism only come into view when we consider how the capacity to have states with representational content is related to the capacity to self-ascribe such states. Or so I shall suggest when we come to consider grounds for the Kantian thesis in §3.

written shortly after the second edition of the *Critique* was published, Kant remarks that, were I not able to apply the categories to the data supplied by my senses,

> I would not even be able to know that I have sense data; consequently for me, as a knowing being [*als erkennendes Wesen*], they would be absolutely nothing. They might still (if I imagine myself to be an animal) exist in me (a being unconscious of my own existence) as representations connected according to empirical laws of association, carrying on their play in an orderly fashion, and thus even having an influence on my feeling and desire (assuming I were even conscious of each individual representation, but not of their relations to the unity of representation of their object by means of the synthetic unity of their apperception). This might be so without my knowing the slightest thing thereby, not even what my own condition was.[7]

This passage involves a number of notions (notably "empirical laws of association" and the "synthetic unity of apperception") that we are not yet in a position to gloss. Even lacking an interpretation of these notions, however, we can observe that the argument of the passage depends on a distinction between a creature's merely *having* representations and its *being conscious* of its representations. Kant is evidently willing to concede (in the first parenthetical remark) that an animal might *have* sensory representations, although he insists that, inasmuch as an animal is not a knowing being, such representations would be "absolutely nothing" to it. His formulation here echoes the formulation in B131-2, where he says that a representation which I could not accompany with "I think" would be "nothing to me." His view thus seems to be that, because animals lack the ability to accompany their representations with "I think", they are not capable of *knowing* anything on the basis of their representations – they do not know "the slightest thing thereby." But he plainly does not take this to show that animals lack representations. On the contrary, he seems quite ready to allow that animals can have representations which succeed one another "in an orderly fashion" and which have an influence on their feelings and desires. And in other discussions of animals and infants, from the pre-critical period to the end of his life, he consistently

---

[7] Letter to Marcus Herz, May 26, 1789 (11:52T).

holds to this view.[8]

*

In light of all this, it seems more plausible to adopt a limited reading of the Kantian thesis, according to which the requirement of accompaniability with "I think" applies only to the representations of creatures of a certain kind. When I formulated the Kantian thesis in the previous section, I incorporated just such a limitation: I suggested that the requirement applies to just those creatures that possess the faculty of understanding. Admittedly, this way of expressing the limitation is not very illuminating, for it does not clarify what a creature must be like in order to count as possessing this faculty. At this point, however, I will not try to make the scope of the Kantian thesis more definite. Kant offers several suggestive characterizations of the understanding: it is the faculty which brings forth representations for itself, the "spontaneity of cognitions," the faculty of thinking, of concepts, of judgments, and of rules (see A51/B75, A68-9/B92-4, A126, B137). But his understanding of these various glosses is bound up with his reasons for thinking that representations available to the understanding must be capable of being accompanied with "I think", to such an extent that it would be hopeless to try to understand what he means by "cognition," or "thinking," or "judgment" without considering those reasons. For the present, then, it must suffice to note that the understanding is meant to be a faculty of discursive thought, whose characteristic exercise occurs in judging, and whose presence in us sets us apart from mere brutes.[9] What I will call the *limited reading* of the Kantian thesis maintains

---

[8] An especially plain statement occurs at CJ 5:464n. For evidence of how little Kant's views on these topics changed over the course of his life, see the index entries for "animals" and "children" in his *Lectures on Metaphysics* (1997b).

[9] An adequate discussion of the way in which understanding sets us apart from mere brutes would, I think, need to invoke both the distinction between first and second potentiality and the notion of human nature. A normal human infant who has not yet learned to speak has the faculty of understanding in first potentiality: she has what can develop into a standing capacity for thought and judgment. A normal adult has understanding in second potentiality: her capacity for thought and judgment is actual, and only awaits occasions for its exercise. But even a human being who altogether lacks the capacity for thought and judgment, either from birth or as a result of injury, belongs to a species whose proper function involves the exercise of this capacity: lack of understanding in such a person is a form of natural *defect*, whereas a mere brute is not defective for want of the power to judge. Thus the capacity for understanding sets *all* human beings apart from mere brutes, inasmuch as we are *by nature* rational animals.

that

(LR) Any creature that possesses the faculty of understanding must be capable of accompanying all of its representations with "I think".

*

The limited reading of the Kantian thesis thus takes Kant's requirement to apply only to creatures of a certain kind. For any such creature, however, it takes the requirement to apply to every one of that creature's representations. This certainly seems to be the implication of Kant's claim that it must be possible for the 'I think' to accompany "*all* my representations," and of his suggestion that a representation not thus accompaniable "would be *impossible*." But Kant goes on to qualify the latter claim in a way that suggests another possible reading of his claim. A representation that I could not accompany with "I think", he claims, "would be impossible, or at least would be nothing to me." As I noted earlier, this seems of a piece with his remark in the letter to Herz that representations which could not be brought to the synthetic unity of apperception would be "for me, as a knowing being, …absolutely nothing." The suggestion of these formulations seems to be, not that there can be no such thing as a representation of mine which I cannot accompany with "I think", but rather that a representation of mine which did not meet this condition could not have a certain kind of significance for me. Kant says that such a representation would be nothing to me "as a knowing being," by which he seems to mean that my having it could not supply me with knowledge of anything. As he puts it in the letter to Herz, I could have such a representation "without knowing the slightest thing thereby." All this suggests a *qualified reading* of the Kantian thesis:

(QR) Any representation which can contribute to a creature's knowing anything must be accompaniable by that creature with "I think".

I call this reading "qualified" to emphasize the possibility of expressing it as a claim about any representation that is significant to a creature *qua* knowing being.

The crucial difference between the limited reading of the Kantian thesis and the qualified reading is that the latter allows that a creature might possess the faculty of understanding and yet have

representations which were not accompaniable with "I think." Such representations would be nothing to it "as a knowing being," but they might be significant for it in some other way. On this reading of his claim, in other words, Kant could allow that unapperceptible representations might exist "in me" and influence my feelings and desires, just as the representations of animals, although not apperceptible, may influence their feelings and desires. Indeed, he could allow that unapperceptible representations might exert a *kind* of influence on my thoughts and judgments: they might cause me to be disposed toward a certain view of some matter, albeit not by presenting me with grounds for that view. To say that such representations would be nothing to me as a knowing being is not to deny that they might *affect* my beliefs; it is only to insist that they could not *justify* them in the way that would bear on the question whether the relevant beliefs should count as knowledge.

Kant seems not to regard the choice between (QR) and (LR) as a matter of much consequence: he treats "nothing to me as a knowing being" and "nothing at all" as if they were interchangeable formulations, at least for the purposes of his argument.[10] Since we do not yet know the grounds for this indifference, however, it will be useful for us to keep the distinction between (QR) and (LR) in mind. One benefit of drawing this distinction is that we will not be led to overestimate what Kant's thesis must commit him to: if the best reading of his thesis turns out to be the qualified reading, then he need not deny either that there are unapperceptible representations or that apperceived representations are identical with states of the subject that are not apperceived as such.

2.2 WHAT IS A REPRESENTATION?

The significance of the Kantian thesis will obviously depend on what the term "representation" is taken

[10] In a footnote to the A-Deduction, for instance, he writes that "All representations have a necessary relation to a *possible* empirical consciousness: for if they did not have this, and if it were entirely impossible to become conscious of them, that would be as much as to say that they did not exist at all" (A117n).

to mean.[11] In my discussion to this point, I have left the term unexplained. This is more or less the situation in which Kant himself leaves it: nowhere in the *Critique* does he give a general explanation of what a representation is. He does note that "all representations... belong, in themselves, as determinations of the mind, to our inner state" (A34/B50), and that "[a]ll representations, as representations, have their object" (A108). Also, in an important passage early in the Dialectic, he gives a taxonomy of different kinds of representations:

> The genus is *representation* in general (*repraesentatio*). Under it stands the representation with consciousness (*perceptio*). A *perception* that refers to the subject as a modification of its state is a *sensation* (*sensatio*); an objective perception is a *cognition* (*cognitio*). The latter is either an *intuition* or a *concept* (*intuitus vel conceptus*). The former is immediately related to the object and is singular; the latter is mediate, by means of a mark, which can be common to several things (A320/B376-7).

But this taxonomy does not answer the question what it is in virtue of which all these kinds of things count as representations. And in any event, Kant begins to use the term "representation" long before he gives even this much explanation of it. Evidently, he expects his readers to bring to the text some intuitive sense of what a representation is. The natural thought, surely, is that a representation must be a kind of mental state or event (a "determination of mind") which has a certain intentional content ("has an object").[12] But how exactly are we to conceive of such states or events, and how do they fit into the

---

[11] There is some controversy about whether to translate the German word *Vorstellung* as "representation" at all: in his recent translation of the *Critique*, for instance, Werner Pluhar opts for "presentation" on the grounds that "representation" tends to suggest that Kant is committed to a "representational theory of perception," on which intentional states inform us about the world by somehow resembling or standing for the outward things they inform us about (see Kant 1996b, p. 22n73). Kant himself, however, gives "repraesentatio" as a Latin gloss on *Vorstellung* (see A320/B376, quoted in the main text), and I do not think the corresponding English term carries any necessary suggestion of the unhappy epistemological theory on which mental states represent by somehow resembling their objects, or indeed of any very particular view about epistemology or intentionality. The term "representation" can be heard simply as a generic term for a particular kind of mental occurrence – one which consists, as it is natural to say, in the subject's representing something as this way or that – without importing any particular theory about what such representing consists in.

[12] Admittedly, this gloss fits imperfectly with the suggestion that sensations are a species of representation: sensations as Kant understands them seem not to have objects. Kant's reason for classifying sensations as representations, I think, is this: they constitute an *aspect* of representational states – an aspect we can consider by abstracting from (what it is natural to call) the representational import of the relevant states. Suppose, for instance, that it looks to me as if there is something red in front of me. Abstracting from the representational significance of this appearance, it seems possible for me to contemplate the sensory quality of my experience as such, to consider what it is like for me to be appeared-to-

more familiar taxonomy of mental goings-on recognized by commonsense psychology?

Most commentators on the first *Critique* follow Kant in not trying to explain what a representation is except by letting it figure in various contexts as a general term for content-bearing mental states or events.[13] And there may be good reasons for this policy. It seems plausible that the reason why Kant does not explicitly define the term "representation" is that this is just his generic term for a potentially knowledge-supplying "determination of mind," and his answer to the question what characterizes this sort of determination as such is really the whole of the Aesthetic and the Analytic. If this is right, it would be rash to suppose we could understand what kind of thing a representation is before grasping the overall shape of Kant's epistemological outlook. Having noted this need for caution, however, we can at least draw a few distinctions and mark some different routes along which an understanding of what representations are might develop.

*

One distinction worth remarking is between the sense of "representation" in which the term is a count noun ranging over possible representational *contents*, and the sense in which it ranges over particular mental *acts* (representings) by particular subjects. In the former sense, a representation of a red sphere and a representation of a green cube are different representations, but if you and I are both representing a red sphere, we share the same representation, and if I have this representation at one time,

---

redly. This nonrepresentational remainder, the sensory quality of my experience considered merely as such, is my sensation: it is, as Kant puts it, "the effect of an object on our capacity for representation" (A20/B34) considered merely as such. But if this is the right way to conceive of sensations, then it is misleading for Kant to classify them as a *species* of representation – for they are a kind of determination of mind we conceive of precisely by abstracting from one of the defining features of representations, their object-directedness. In any case, the apparent inconsistency between Kant's claim that all representations have objects and his classification of sensations as representations is a difficulty that any interpretation of his views on representations will have to confront. I say more about Kant's conception of sensation in Chapter III.

[13] A notable exception is Wilfrid Sellars' *Science and Metaphysics* (1967). Sellars's work has had a great influence on my conception of what a representation is, although I would be wary of claiming fully to understand his view. I am also indebted to John McDowell's writings on this topic, as will be obvious to anyone who has read his *Mind and World* (1994) or his Woodbridge Lectures (1998b).

and then again at a later time, I have had the same representation twice.[14] In the latter sense, by contrast, representations are individuated not just by their content but by when they occur and who does (or undergoes) the representing. A representation in the act sense is a subject's representing things as thus-and-so.[15] It is natural, once we have introduced the count noun "representation," to speak of a subject's representations as things she *has*, but we should be careful not to read too much into this idiom. To have a representation in the act sense is just to be representing things in a certain way.

When Kant claims that "It must be possible for me to accompany all of my representations with the 'I think'," it is natural to understand him as speaking of representations in the act sense: "my representations" are presumably those content-*tokenings* which I am undergoing at a given time. This is probably not the right way to interpret "representation" in all contexts: when Kant says that concepts are a species of the genus "representation in general," for instance, he seems plainly to mean that a concept is a kind of representational *content*. But in the sense in which a representation is a "determination of mind" which its subject can accompany with "I think", it seems that a representation must be a content-tokening, my representing something as thus-and-so. What will count as a representation in this sense is not, e.g., the abstract concept *red*, but rather an *actualization* of my capacity to represent things as red in the context of a representation of some object or state of affairs. Suppose, for instance, that a subject has an intuition (a singular representation that is "immediately related" to a particular object) whose content can be expressed – following Wilfrid Sellars's useful proposal about how to represent the

---

[14] There is also a further distinction, which is not crucial for present purposes but which I certainly would not wish to deny, between representational contents that are *generically* the same and representational contents that present the very same *individual* in the very same way. If you and I are each representing a red sphere, then the content of our representations may be generically the same, at least as far as redness and sphericality are concerned, but our representations may nevertheless be distinguishable in that you are representing one red sphere and I am representing another. Some philosophers would insist that, if our two representations are qualitatively indistinguishable, the difference between them can only be a matter of their having different causes, not of their differing in content. But I would not want to endorse such a view, or to suggest that Kant is committed to endorsing it.

[15] This is an act in a very permissive sense, one that applies whenever a capacity to do something is actualized. To say that representings are acts in this sense need not carry any implication of intention on the subject's part: my representing a red sphere need not be any more intentional than my breathing is normally intentional.

content of an intuition – as

this red sphere[16]

In virtue of having this intuitive representation, a subject will be actualizing her capacity to represent something as red. She will thus have a representation of red in the act sense of "representation": she will be tokening the conceptual content *red* in application to a particular object, a certain sphere.

*

In the sense relevant to the Kantian thesis, then, representations are mental determinations identified by their content, the subject to which they belong, and the time at which they occur or the period over which they persist. But of course there are various kinds of mental states and events that meet this description. For instance, states that involve my holding something to be true – states of or involving belief – certainly seem to merit the title "representations"; but so do states that differ in kind from beliefs. Indeed, given that Kant's primary concern in the part of the *Critique* in which he states the Kantian thesis is to show that representations deriving from *intuition* must have a certain kind of unity if they are to be possible objects of cognition, the kind of representation that he has foremost in mind is presumably not belief-like. It will be useful, therefore, to say something here about how we might conceive of this kind of representation.

Intuitions are sensible representations through which we are informed about particular objects. My seeing a red sphere sitting in front of me is presumably an example. But plainly, I can undergo such

---

[16] For the time being, I will follow Sellars in taking *this* φ to be the general form of an intuitive representation. His idea, I take it, is that at any given moment the "manifold of intuition" will consist of an indefinitely large number of object-related representations, which manifold will be only partly captured by any particular such specification. For my purposes, the crucial features of Sellars's proposal are: (1) that it captures an intuition's "immediate relation to an object" by including a demonstrative in its content-specification, and (2) that it assumes that the demonstrative in question will not be a bare "this," but a "this" determined by some predicative matter. The content of an intuitive representation will thus be essentially related to the content of particular *judgments* about the subject's environment. (Related to particular judgments but not yet committing the subject to those judgments. I take it that this is why Sellars expresses the contents of intuitions in noun-phrases, even while he insists that their contents are essentially related to propositional contents: the absence of a copula in the noun phrase marks the fact that the subject does not in merely intuiting take any stand on whether the content at issue is true.) For Sellars's views on intuitions, see *Science and Metaphysics* (1967), ch. 1, and "Some Remarks on Kant's Theory of Experience" (1974). I say more about the notion of an intuition in Chapter IV, §3.

an episode without *believing* that there is a red sphere in front of me – for I may doubt the evidence of my senses. Nor, however, does my intuiting a red sphere merely involve my being *disposed* to form a certain belief or make a certain judgment. To have a visual representation as of a red sphere's being in front of me is not merely to find myself with an unaccountable yen to think "There is a red sphere in front of me." It is, surely, not merely to acquire a dispositional state, but to undergo an actual episode – an episode that involves its at least seeming to me as if the obtaining of a certain state of affairs (e.g., there being a red sphere in front of me) is the *cause* of things looking to me as they do. In short, the relevant kind of episode will involve its at least seeming as if a certain fact is making itself visually manifest to me.[17] To have such a representation is not just to be disposed to form a certain belief or to make a certain judgment; it is to be actually presented with an apparent *ground* for belief. Intuitions, we could say, embody purported information about our sensibly given environment.

What does it mean, though, to "embody purported information"? Well, I mean my talk of representations "figuring in our lives as knowing beings" to suggest an answer. An intuition is the sort of thing that, if it is veridical and if it occurs in a self-conscious creature, normally puts that creature in a position to *know* the fact of which it is a representation. Intuitions are thus information-bearing states in a particular sense: they are not merely states of a subject whose presence more or less reliably *indicates* that some object has some characteristic (to someone aware of the correlation between the state and the object's having that characteristic); they are states whose obtaining is, at least in some cases, by itself sufficient to *inform the subject* of the relevant fact. Obviously, not every state of the subject whose characteristics are systematically correlated with the characteristics of some worldly thing informs the subject about that thing. The concentration of carboxyhemoglobin in a person's blood, for instance, is systematically correlated with, and thus in one undemanding sense a representation of, the quantity of

---

[17] Here I echo a formulation of John McDowell's. For McDowell's own elaboration of this point, see his 1998b, Lecture I, §3.

carbon monoxide in her environment. But the concentration of carboxyhemoglobin in my blood does not inform me of the quantity of carbon monoxide in my environment: if it did, carbon monoxide poisoning would be much easier to avoid. Moreover, even if I were *aware* of the concentration of carboxyhemoglobin in my blood, this awareness would not necessarily put me in a position to know the concentration of carbon monoxide in my environment; I would have to know that there is a correlation between the former sort of fact and the latter. By contrast, I take it that an intution in Kant's sense is a kind of state of the subject such that, if the subject is aware of it at all, she is necessarily (purportedly) informed of how things stand with the object of that representation.

This is not to say that every intuition does inform us of some fact: some representations are not veridical, and even when they are veridical, our overall epistemic situation may be such that we are not entitled to assume that they are. We can, however, say this: representations in Kant's sense belong essentially to a subject's cognitive economy, to the set of states that bear on the question of what the subject knows. The Kantian thesis applies only to intuitions that are cognitively available in this sense.[18]

2.3 WHAT IS IT TO ACCOMPANY A REPRESENTATION WITH THE "I THINK"?

Given this interpretation of representations as content-tokenings, there remains a question about what it is to accompany a representation with the "I think." What is "the 'I think'", and what kind of connection between representation and self-ascription should we take the word "accompany" to require? Any account of what Kant has in mind when he speaks of "accompanying representations with the 'I think'" will involve some speculation: this is another phrase he begins to use without explanation, leaving

---

[18] A fuller treatment of Kant's views on epistemology would need to say more about the sense in which representations "purport to" embody information even when they are not veridical. As I understand him, Kant's view is that, although the case in which a representation merely purports to present information is possible, it is necessarily an exceptional case whose occurrence requires a special explanation. In other words, the case in which representations actually present information (i.e., in which they are veridical and well-grounded) is necessarily primary in the order of understanding. I

his readers to surmise what sort of act he has in mind. Let me begin by stating some assumptions I am making about what the relevant act must involve.

*

If a representation is, as we have been supposing, a particular actualization of representational powers, then it seems reasonable to assume that accompanying a representation with the "I think" involves ascribing such an actualization to oneself. If we confine our attention to the case where the content of the representation in question is propositional, the form of the relevant self-ascription would presumably be

I *R* that *p*

where *R* is a verb that ascribes an act of representation and *p* is a propositional content.

But what can go in the *R* position here? Well, one permissible substitution for *R*, clearly, would be "think" itself. If one has a representation with the content that *p*, and says (or thinks) "I think that *p*," one has made explicit (to the world at large, or merely to oneself) one's endorsement of the content of one's representation. It is a familiar point about statements of this form that they have a double significance: on the one hand, "I think that *p*" says something about how things stand with the subject; on the other hand, it amounts to an affirmation (accompanied, as a matter of pragmatics, by a note of hesitancy) of the proposition contained in the "that"-clause. In appropriate circumstances, to make such a statement is to ascribe a representational state to oneself and at the same time to endorse the relevant representation as true.[19] Surely this should count as "accompanying a representation with 'I think'."

The statement "I think that *p*," however, is only one member of a wider class of self-ascriptions which can perform this sort of function: to say, for instance, "I see that *p*" or "I remember that *q*" is

---

cannot defend this interpretation here, however. For helpful discussion of related issues, see Engstrom, "Kant on Objective Validity, Truth, and Judgment" (unpublished).

[19] Nor does this double significance characterize only the *statement* "I think that *p*"; the mere *thought* that one thinks that *p* has the same two valences. I intend everything I say about accompanying representations with "I think" to apply

likewise both to ascribe to oneself a certain mental state and to affirm the contained proposition. Moreover, corresponding to each of these judgments is a related judgment that does not commit the subject to the truth of the contained proposition: "I am tempted to think that *p*," "It looks to me as if *q*," "I seem to remember that *r*." *All* of these judgments, both those that commit the subject to the truth of the relevant proposition and those that do not, are alike in that they explicitly register that things seem thus-and-so to the subject. All of them seem capable, in appropriate circumstances, of expressing self-conscious recognition of a representation.

When Kant speaks of "accompanying representations with 'I think'," I take him to be using the expression "I think" as a generic representative of expressions capable of fulfilling the latter function. I assume, therefore, that a self-ascription involving *any* of the kinds of expression indicated in the preceding paragraph could count as a case of the subject's accompanying one of her representations with "I think". Which expression was appropriate would depend on whether the subject wished to specify the epistemic basis of her inclination to take the relevant proposition to be true and whether she wished to endorse the representation. But whichever expression she used, she would be explicitly ascribing a representational state to herself. I take the Kantian thesis to require that knowing beings have the capacity to do just this: they must be able explicitly to ascribe to themselves representational states, and either endorse or withhold their endorsement from the relevant representations, using (1) a first-person pronoun and (2) a verb that ascribes a representational state.

*

Even with these stipulations in place, the Kantian thesis remains unclear at a crucial point. To see the unclarity, consider a creature with conceptual capacities sufficient to judge of itself "I *R* that *p*" for any propositional content *p* determined by one of its representations. Suppose this creature has a visual representation with the content

---

indifferently to overt (spoken or written) self-ascriptions and to the mere thoughts that such overt ascriptions may be

(R) A pale pink flower here

and judges

(J) I think that there is a pale pink flower here.

Will the creature have thereby "accompanied its representation with 'I think'"? The right answer, it seems to me, is: not necessarily. For all that has been said so far, it seems that the relationship between the creature's having the representation (R) and its making the judgment (J) might be entirely coincidental: the creature might have (R), and judge (J), and yet (J) might be just an idle speculation, not a judgment prompted by its awareness of the representation in question. When we speak of a creature "accompanying a representation with 'I think'," we surely mean to require that it ascribe to itself a thought which reflects an *awareness* of the relevant representation. It seems, then, that a capacity to accompany my representations with "I think" must be more than a mere capacity to self-ascribe thoughts whose contents in fact coincide with the contents of my representations; it must be a capacity to self-ascribe such thoughts in a way that is intelligible as *expressing an awareness* of the relevant representations.

To understand the requirement of accompaniability with "I think" in this way is to understand the Kantian thesis as requiring that knowing beings possess a particular kind of self-consciousness: such beings must *know* what representations they have, and be capable of thinking of those representations as their own. This understanding of Kant's requirement contrasts with another understanding, on which it sets a weaker condition. Consider the following gloss on the Kantian thesis, from Jonathan Bennett's book *Kant's Dialectic*:

> Kant's doctrine about self-consciousness, namely that any judgment I make can be accompanied with 'I think …' …merely implies that given any judgment (*P*) which I make there is a *correlated* true judgment with myself as the subject matter (I judge that *P*) (1974, p. 74).

On Bennett's interpretation, all Kant requires is that, for each judgment I make, there be "a correlated

---

used to express.

true judgment with myself as the subject matter." That is to say, for any judgment *p* made by me, it must be the case that I *could* also judge truly of myself, "I judge that *p*." There is no requirement that I should be aware of this entitlement when I have it – no requirement that I should have the sort of consciousness of my representations that enables me to self-ascribe them precisely because I know I have them. Bennett's requirement will thus be met trivially by any creature capable of ascribing representational states to itself: if a creature is representing things as thus-and-so, then of course it will be making a true judgment if it judges that it is representing things as thus-and-so. But it is not obvious that a creature with the capacity to make such judgments will necessarily know when it is entitled to make them; so the reading of the Kantian thesis I am advocating seems to make it a stronger requirement than it is on Bennett's reading. We can call Bennett's interpretation the *mere capacity* reading to mark its silence on the question what sort of awareness a subject must have of the states she has the capacity to ascribe to herself.[20] On this reading, Kant's claim is simply that

(MC) Given any representation I have, there must be a correlated true judgment with myself as the subject matter which I could make (presumably, something along the lines of "I am representing things as thus-and-so").

I do not think Bennett is alone in his reading of the Kantian thesis. Bennett's advocacy of the mere capacity reading is unusually clear-cut, but a number of commentators seem to reveal, by the arguments they offer for the Kantian thesis, that they have this reading (more or less definitely) in mind. Nor is the motivation for the mere capacity reading hard to understand: it tends to smooth the path for Kant's argument, making the Kantian thesis appear to depend mainly on the truistic observation that if a creature is representing things as thus-and-so, then it will be making a true judgment if it judges that it is

---

[20] In claming that the relationship between being able to make self-ascriptive judgments and knowing when one is entitled to make such judgments is not obvious, I do not mean to suggest that there is no relationship. In fact, I think there are good reasons not count a creature as capable of making judgments with genuinely first-personal contents if it is does not know without observation when it is entitled to make certain ranges of such judgments. If this is right, then Bennett's formulation of the Kantian requirement may turn out to be only apparently weaker than mine. Even if this is so, however, it seems preferable to have a formulation that makes clear the connection between a capacity for self-ascription and knowledge of what representations one has.

representing things as thus-and-so. There are certainly passages in the *Critique* in which Kant seems to rest his case on this point – as when he remarks that the principle that I must be able to accompany all my representations with "I think" is analytic, since

> it says nothing more than that all *my* representations… must stand under the condition under which alone I can ascribe them to the identical self as *my* representations (B138).

It is important to recognize, however, that the truistic observation cannot by itself establish the Kantian thesis, even on the mere capacity reading. Certainly a creature that is representing things as thus-and-so, and that judges of itself that it is representing things as thus-and-so, will be making a true judgment – but why *must* knowing beings be capable of making judgments of this kind? Even on the mere capacity reading, the Kantian thesis makes a significant claim: namely, that a capacity to make *judgments that self-ascribe representational states* is one of the intellectual capacities a creature must possess in order to count as a knowing being. This claim is by no means a truism, and recognizing this, we should look again at the passages in which Kant seems to base his argument for it on the truistic point.

The language in which Kant frames his discussion of the capacity to accompany representations with "I think" – his frequent use of the terms "consciousness," "self-consciousness," and "apperception" in characterizing this capacity – surely tends to suggest something stronger than the mere capacity reading. One of Kant's characteristic formulations of the function of the "I think" is that it serves "to introduce all thinking as belonging to consciousness" (A341/B399). This suggests that when a creature accompanies one of its representations with "I think", in the way that Kant has in mind, it will be expressing consciousness of its own state. This suggestion would be hard to understand if the capacity to accompany representations with "I think" did not involve the subject's being aware of its representations. But what does this requirement that one be aware of one's own representations amount to?

A first stab at articulating Kant's thesis in a way that took account of this need for awareness

would be to require that

> (KC) Whenever I have a representation with propositional content *p*, I must *know* that I have such a representation, and be able to express this knowledge in a judgment of the form "I *R* that *p*."[21]

This would be to require that the capacity to make such judgments be a *knowledgeable capacity*: a capacity not just to make judgments that self-ascribe representations, but to make them because one knows when they are true.

This would certainly be a step in the right direction, but I think there is room for doubt about whether (KC) brings out the full force of the Kantian requirement. Some philosophers have maintained that any privileged knowledge we have of our own minds is due merely to our having a specially good vantage point on our own behavior.[22] Others have held that our knowledge of our own mental states is the product of something like inwardly-directed perception or self-scanning, so that to know that one has a certain representation is to have a representation and be caused thereby to believe that one has such a representation.[23] Inasmuch as (KC) does not say anything about the *way* in which a subject must know that she has a given representation, it seems to leave open the possibility of such accounts of self-consciousness. But a creature that knew its own mind only in one of these ways – if such a being is really conceivable – surely should not count as meeting the Kantian requirement. For such conceptions of self-consciousness imply that my having a certain representation and my knowing that I have that

---

[21] Note that the requirement is that I know what representations I have, not that I know whether the relevant representations are veridical. This means that although I must know, e.g, that I have a visual representation as of its being the case that *p*, I may in some cases not know whether this representation constitutes my actually seeing that *p* or my merely seeming to see that *p*. (KC) thus does not imply that I must always be right in my view about what should go in the *R*-position in apperceptive judgments of the form "I *R* that *p*." (This is not to say, of course, that I never am in a position to say, on the basis of apperceptive knowledge, that I am seeing that *p* rather than merely seeming to see that *p*. It is possible, and I think attractive, to hold that when I *am* actually seeing that *p*, this forms part of what is available to my apperceptive awareness. I take this to be the position of McDowell 1995.) These points also apply to the next formulation I consider, (CC).

[22] Such a view is often attributed to Gilbert Ryle, and Ryle does say things that suggest it. For instance, he claims that "[t]he sorts of things I can find out about myself are the same sorts of things that I can find out about other people, and the methods of finding them out are much the same," and that the difference between how I know about my own mind and how I know about the minds of others turns on a "difference in the supplies of the requisite data" (1949, p. 155). If this is not to deny that we know our own minds in a special way, it is at least to minimize the significance of the point.

representation are distinct mental states, the latter of which merely happens, in virtue of the existence of some fortunate circumstance or mechanism, to accompany the former. But this seems to be precisely the sort of picture of self-awareness that Kant means to rule out when he claims that a representation which I could not accompany with "I think" would be nothing at all, or at least nothing to me: he seems to be denying that we can make sense of the idea of a representational state of a knowing being without reference to that being's potentially self-conscious awareness of that state (or, at least, insisting that any sense we *can* make of the idea of a representational state of which a knowing being is unaware will leave such states irrelevant to the question what the subject knows).

To bring out the full force of the Kantian requirement, then, we must demand not only that the subject know when she has a given representation and that she be able to express this knowledge in first-person judgments, but that her knowledge of such matters be a particular *kind* of knowledge: knowledge that is *immediate*, in the sense that to be in the state that is the object of the knowledge just is, at least in part, to have knowledge of that state. If we call this sort of cognitive relationship to a mental state "consciousness," we can express the claim that a knowing subject must have the ability to accompany its representations with "I think" by saying that such a subject must have the capacity consciously to ascribe representations to itself. I will call this the *conscious capacity* reading of Kant's thesis:

(CC) My having a representation with the propositional content that *p* must involve my *consciously representing* that *p* in such a way that I can express this consciousness in a judgment of the form "I *R* that *p*."

I do not want to place much weight on this formulation of the Kantian requirement, however, for I do not think we have a very clear understanding of the logical grammar of claims involving the term "conscious." I intend my formulation of (CC) merely to mark the need to invoke a specially immediate kind of knowledge in explaining what it is to accompany a representation with "I think." One benefit of reflecting on the Kantian thesis and Kant's argument for it is that it will help us to clarify what kind of

---

[23] A classic example is David M. Armstrong's *A Materialist Theory of the Mind* (1968).

knowledge this must be.

3. THE INTUITIVE ARGUMENT

In the preceding sections, we saw several different ways of understanding the Kantian thesis. The following table summarizes the possibilities:

	Universal Reading	*Limited Reading*	*Qualified Reading*
Mere Capacity	KT1	KT2	KT3
Knowledgeable Capacity	KT4	KT5	KT6
Conscious Capacity	KT7	KT8	KT9

I suggested that the readings associated with the first column are far-fetched, and I gave some reasons for thinking that Kant's intention is better captured by a reading from the last row than by one from either of the first two rows. But what sort of argument can be offered in defense of the Kantian thesis, and what reading does the argument support?

Plainly, the place to begin looking for an answer to this question is in §16 of the B-Deduction. There, in the same sentence in which he announces his thesis, Kant offers a brief but suggestive defense of it:

> It must be possible for the "I think" to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me (B131-2).

The justification Kant offers here has a kind of superficial reasonableness that disappears on closer inspection. He seems to reason: (1) a representation that I could not accompany with "I think" would represent something that could not be thought (by me), and (2) that amounts to saying that such a representation would be impossible, or at least would be of no significance to me. But step (1) is only obvious if "could not be thought" means "could not be the object of *an explicit self-ascription* of thought," and if that is what is meant, then it is not clear that step (2) follows. Why would a representation that was not open to this kind of self-conscious consideration be "nothing to me"?

3.1 A first step toward finding a plausible argument here is to recall a point from the letter to Herz quoted earlier. As I noted in §2.1, Kant's formulations in that letter seem to parallel his formulations in B131-2, but in the letter he says specifically that a representation which I could not bring to consciousness would be nothing to me *as a knowing being*. If we read this specification back into the argument of B131-2, the resulting claim is:

(C) A representation that I could not accompany with "I think" would be nothing to me as a knowing being.

If (C) represents Kant's reason for holding that it must be possible for me to accompany all of my representations with "I think," then it seems we must read him as arguing for this thesis on its qualified reading, (QR). We must read him as claiming that any representation which can contribute to a creature's knowing anything must be accompaniable by that creature with "I think."

Now, (C) is hardly indisputable, but I think it has some intuitive appeal. The intuition that supports it can be provoked by reflecting on the question: How could a representation of which I was unaware present me with a reason for thinking anything? We saw in §2.2 that there is an intuitive distinction between states of the subject that can be regarded as *indications* of environing facts and states that *inform the subject* of some such fact. The obtaining of a state of the former kind (e.g., the concentration of carboxyhemoglobin in the subject's blood having reached a certain level) may certainly

be a reason to believe that a certain environing fact obtains, in the following sense: someone who knew that the state obtained, and knew its significance, would have a reason to form a certain belief about his environment. But for a state's obtaining to be a reason to believe something in this sense is not necessarily for the state to *present the subject with a reason* to form the relevant belief.

Intuitively, a state only presents a subject with a reason for belief if the state impinges in the right way on the subject's consciousness: the subject must be made aware of a consideration which she could call on in a rationale for believing what she does. A representational state which did not in this sense present a subject with a reason would in one significant sense be "nothing to her": although it might *be a reason* (in the sense specified above) to believe something, and although its obtaining might have an influence on the subject's mental life (perhaps even engendering the very belief for which it was a reason), it seems that this could be at best a *subliminal* influence, one that did not engage the subject's own rational powers. For if the state in question did not impinge on the subject's consciousness, then the information embodied in that state presumably could not figure in her thinking about what to believe, for the simple reason that one's thinking cannot take account of considerations of which one is not made cognizant.[24]

I suggested earlier that when Kant speaks of "representations," he means states that purport to inform the subject of facts – states that, we can now say, purport to *present the subject with reasons* to think that what those states represent as the case is the case. The basic idea underlying (C), I think, is that a representation which I could not accompany with "I think" could not play this role in my mental life. Kant's vacillation between the claim that representations which did not meet this condition would be impossible and the claim that such states would be nothing to me as a knowing being reflects his shifting between a sense of the term "representation" in which nothing that did not present the subject with a

---

[24] These formulations are obviously too full of obscurities to bear much weight, but all I want from them at this point is that they should evoke an intuition that speaks in favor of the Kantian thesis.

reason would count as a representation, and a sense – which Kant allows for the sake of argument – in which we might call states of the subject "representations" even though they did not meet this condition. The fundamental thought is that only representations which I can accompany with "I think" present me with reasons. If this is right, and if only reasons with which I am presented can bear on the question of what I know, then a representation which I cannot accompany with "I think" will be nothing to me as a knowing being.[25]

3.2 What, then, can be said in support of the thought that I am only presented with reasons by representations I can accompany with "I think"? Presumably no one would deny that it is a sufficient condition for a representational state to present me with a reason that I be able to accompany that state with "I think," at least if the phrase "able to accompany" is taken to imply what I have called a conscious capacity. But why should we suppose that it is a *necessary* condition of a state's presenting me with a reason that I should be able to accompany it with "I think"? Why *must* I have the capacity to express my awareness in this way?

The intuitive thought is this: what distinguishes a reason with which I am presented from a mere source of blind impulses to believe is that a reason is the sort of thing I can *scrutinize* – the sort of thing whose significance for my thinking I can consider and make decisions about. If I am to count as forming beliefs for reasons grasped as such, I must not simply find myself with an impulse to believe that comes from I-know-not-where; I must be able to consider the basis of my impulse, in a way that

---

[25] The proposition in the second "if"-clause here amounts to a form of internalism about knowledge. Most everyone would concede that, for a belief to count as knowledge, it must not only be true but must in some sense be held reasonably. The distinctively internalist thought is that only considerations a subject is *presented with* can bear on the question of her reasonability. And there is surely something plausible in this thought. We can if we like assess a subject's belief as reasonable or unreasonable relative to "representations" of which she is unaware, but if we do, we are severing the connection between these evaluations and anything the subject could be expected to take account of in deliberation. How could the obtaining of a state of which the subject *could not* take account in deliberation show that she had conducted herself responsibly or irresponsibly in forming a given belief? The internalist thought would be that it

allows me to assess whether the relevant belief would be warranted. Furthermore, my assessment must matter: if I judge that a certain belief is not warranted by a given consideration, that must make a difference to what I actually believe. If my representations are to present me with candidate reasons for belief, then, they must not simply work on me subliminally; they must be available for me to consider, to accept or reject.

But – the thought continues – in order for me to be able to subject my representations to such scrutiny, I must be able to accompany them with "I think." I must, for instance, be able to recognize that *I (at least) seem to see that* p. For if I am not able to recognize that I am in such a state, then neither will I be able to reflect on what grounds my being in that state gives me, in this context, for believing that *p*. A subject who can accompany her visual representations with "I think" can, so to speak, *take a step back* from those representations, and this is precisely what allows her to relate to them as reasons: it is what allows her to bring them within the scope of her rational scrutiny.[26] To recognize that I seem to see that *p* just is to conceptualize for myself the fact that I am in a particular kind of state that recommends the belief that *p*, a state that will in some kinds of circumstances tend to be veridical and in other kinds misleading. In recognizing that I seem to see that *p*, I thus recognize a distinction, and the possibility of a discrepancy, between how things seem to me and how they are. And by the same token, I recognize that I need not go along with my inclination to believe that *p*: my view of the truth can now take account of the fact that this representation may be a mere semblance. In short, I bring into view the source of my inclination to believe that *p*, and articulate it in a form that allows me to weigh it as a reason for belief, in the context of other things I take myself to know. A creature that lacks the capacity to do this cannot literally be said to grasp its visual representations as reasons for belief. And analogous points

---

could not, and that such states are therefore nothing to the subject as a knowing being. I will say more in support of such internalism in Chapter II.

[26] Philosophers writing in a Kantian tradition about the importance of self-consciousness frequently appeal to the metaphor of stepping back. See for example Hampshire 1965, pp. 89-90; Allison 1990, p. 82; Korsgaard 1996, pp. 92-

will apply to any kind of representation that is capable of figuring as a subject's ground for belief or judgment. So in general, a subject must be able to accompany its representations with "I think" if those representations are to present it with reasons for belief. This is the intuitive argument for the Kantian thesis.

3.3 There is of course much else that takes place in §16 of the B-Deduction. Having argued that it must be possible for me to accompany any of my representations with "I think", Kant next points out that the "I" that figures in each of these possible accompaniments must always refer to one and the same thinking subject.[27] This point follows analytically from the supposition that a given set of representations are all mine: whenever any of these representations is accompanied with "I think", the "I" at issue will always refer to the same thinking subject, for by hypothesis all the representations belong to a single subject who will on each occasion be doing the accompanying. But from this analytic point, Kant argues, we can draw important conclusions about how representations belonging to a single subject must be related to one another, and what role the thinking subject must have in making this relatedness possible. As Kant puts it, the *analytic* unity of apperception – the fact that the "I" that can accompany *my* representations will in any instance refer to the very same thinking subject – presupposes a *synthetic* unity – a relatedness among my representations which is not deducible from the contents of these several representations. The main task of the remainder of the Deduction, as I read it, is to explicate this latter form of unity, and to show that its existence requires that the representations given to me in intuition should admit of classification under the categories.

100; and Moran 2001, pp. 138-51. I owe my own sense of the aptness of the image to conversation with John McDowell. I say more about the literal content of the image in Chapters III-IV.

[27] This point is actually implicit in Kant's claim that it must be possible for *the* "I think" to accompany all of my representations. In my discussion to this point, I have ignored the definite article Kant includes here in order to concentrate on aspects of the Kantian thesis that seem to me more fundamental. But the reason why the article is included, seemingly, is to stress the point that, whenever a given subject accompanies any of its representations with "I

My purpose here is not to give an account of all this, although I will in due course say something about the claim that the analytic unity of apperception presupposes a synthesis. For the present, however, the point I wish to stress is that there is a prior claim of which we should not lose sight in the glare of these subsequent fireworks. The observation that the "I" which figures in any of my possible accompaniments of a representation with "I think" must always refer to the very same thinking subject – the observation on which the whole subsequent argument depends – only attains its significance when connected with the thought that a knowing being must be able to accompany its representations with "I think". If Kant did not suppose that he had established this more basic point, he would not regard himself as entitled to infer that the relevant representations must be connected in a way that makes such accompaniment possible.

Some commentators on the Deduction take it that the reason why Kant supposes that I must be able to accompany all of my representations with "I think" is that a set of diverse representations simply would not count as all belonging to the very same subject if the subject in question could not accompany those several representations with "I think."[28] If I am right, this reading gets the order of Kant's

---

think", the relevant tokening of "I" will refer to the very same thinking subject. I am indebted to Steve Engstrom for making me see this.

[28] Commentators who read Kant's remarks on apperception as addressed, in the first instance, to Hume's "bundle theory" of the self characteristically take the argument of §16 in this way. As against Hume's claim that the identity of the subject comes to nothing more than the existence of a heap of intrinsically disparate perceptions connected by relations of resemblance and cause-and-effect, these commentators read Kant as claiming that for a set of mental states to belong to a single subject, these states must be related in a way for which Hume's account makes no room: they must all be accompaniable with the very same "I think." On Patricia Kitcher's interpretation of the Deduction, for instance, Kant's claim is roughly that distinct mental states are "co-mental" (i.e., belong to the very same mind) if and only if they stand in certain relations of "contentual dependence," relations whose existence Kant takes to entail accompaniability with the same "I think". (I am skirting over much complexity here: see Kitcher's "Kant on Self-Identity" [1982] and, for a fuller discussion, her book *Kant's Transcendental Psychology* [1990].)

My objection to this sort of interpretation is not that it is wrong in attributing to Kant an objection to Hume's views on the self, but that it mislocates the ground of this objection. On my reading, Kant's point is not that there is some relationship of co-mentality whose importance we can recognize prior to seeing why the representations of a knowing subject must be accompaniable with "I think", and from which we can infer the necessity of such accompaniability. Rather, the reasons why the representations of a knowing subject must be accompaniable with "I think" (which turn on considerations about the kind of relation in which a subject must stand to its own representations if those representations are to put it in a position to know anything) come first, and only subsequently do we see that a *single* knowing subject must be able to accompany all of its representations with the *same* "I think".

argument backwards: the point is not that the accompaniability of a set of representations with the very same "I think" (i.e., with tokenings of "I think" in all of which "I" refers to the very same subject) is what *makes* that set of representations count as belonging to a single subject, but rather that, *given* that a set of representations all belong to a single subject, and given that a knowing being must be able to accompany its representations with "I think", various conclusions can be drawn about the form that the representations in question must have and the relations in which they must stand. What I am calling "the intuitive argument" for the Kantian thesis is an argument for the premise that a knowing being must be able to accompany its representations with "I think." The fact that Kant argues for this point in a sentence does not, I think, indicate that it is an insignificant part of his thinking; it indicates, rather, how fundamental and unquestionable he takes the point to be.

4. DIFFICULTIES FOR THE INTUITIVE ARGUMENT

The argument just outlined involves two main claims: (1) the claim that the only reasons that are relevant to my life as a knowing being are reasons with which I am presented, and (2) the claim that a representation presents me with a reason only if I can self-ascribe that representation. There are, accordingly, two kinds of objection to consider: objections which claim that my status as a knower can be affected by factors which do not present me with reasons, and objections which insist that I can be presented with reasons by representations that I cannot self-ascribe. Reflecting on these lines of objection will help us to move toward a clearer and more conclusive presentation of the argument for the Kantian thesis.

The Kantian thesis can indeed be thought of as identifying a sort of ownership of mental states for which Hume's view does not provide (namely, belonging to a subject *qua* knowing being). Many of Kant's claims about "my representations" in the Deduction seem best understood in this way: as claims about representations that are mine *qua*

4.1 The former kind of objection would presumably concede that a subject can only attain knowledge if her thinking is in *some* sense guided by information that justifies her beliefs, but would resist the suggestion that the relevant sort of guidance need involve the subject's being "presented with reasons." In setting out the intuitive argument for the Kantian thesis, I assumed that the question of what a subject knows turns on what considerations can figure in her thinking about what to believe – that is to say, on what considerations she can produce for herself as part of a rationale for what she believes. As I noted, however, this assumption amounts to a form of internalism about knowledge, and such internalism is controversial.

No one will deny that, if a belief is to count as knowledge, it must be more than a lucky guess. But it seems that there might be ways of accommodating this point without requiring that a subject be "presented with reasons" for her beliefs. According to one influential "externalist" approach to epistemology, for instance, a subject's belief that *p* should count as knowledge just if the belief in question is true and is the product of a belief-producing mechanism that is reliable (i.e., normally produces beliefs just when they are true) for beliefs of this kind. Such a "reliabilist" account of knowledge accommodates the point that a belief must be nonaccidentally true if it is to count as knowledge, but it does not require that a subject should be presented with reasons on the basis of which she can justify her beliefs: it demands only that a subject's beliefs should in fact track the truth, not that she should herself be able to supply grounds for them. But if this reliabilist thought is right, then claim (1) faces a dilemma: either the requirement that I be "presented with" reasons by my representations rules out the reliabilist view, in which case it appears to set too high a standard for knowledge; or else it is compatible with reliabilism, in which case it looks too weak to ground an argument for the Kantian thesis.

---

knowing being. But the importance of this sort of ownership is only evident once the reasons for the Kantian thesis

My way of motivating claim (1) was to ask how the obtaining of a state of which a subject was unaware could show that she had conducted herself responsibly or irresponsibly in forming a given belief. This question, however, may seem tendentious: to focus on how a subject has "conducted herself in forming" a given belief suggests that acquiring knowledge must involve something like *reasoning* from grounds to a conclusion. But surely (it will be objected) people know many things which they do not conclude on the basis of reasoning. Indeed, if we concede the plausible principle that a conclusion can only count as knowledge if the beliefs on which it is based count as knowledge, then it appears that a person who could only attain knowledge by reasoning to conclusions could never come to know anything. So if we want to hold onto the thought that a person can only attain knowledge if her thinking is guided by the representations that justify it, it seems that we must understand the relevant sort of guidance in a way that does not in general require reasoning. And if knowledge does not in general require reasoning, why should it require that the subject be "presented with reasons" for what she believes?

4.2 Even if some plausible reading of claim (1) were found, there might still be resistance to claim (2) – the claim that a representation only presents me with a reason if I can self-ascribe that representation. For why should the requirement that I be presented with reasons for what I believe imply that I must be able to self-ascribe the representations on which my beliefs are based? To draw such an implication might seem to involve sliding from the already disputable point that one must be cognizant of the *information* embodied in one's representations to the even more controversial claim that one must be cognizant of the fact *that one has representations embodying* that information. It seems that a creature might count as meeting the requirement imposed by claim (1) – however exactly that requirement is understood – simply in virtue of *having* a representation with a certain content. Why

---

have been grasped.

should we require in addition that a creature be cognizant of the fact *that it has the relevant representation*, in such a way that it could self-ascribe the representation by making a judgment of the form "I *R* that *p*"? Surely the facts of which a knowing subject needs to be informed are facts about the objects on which its beliefs bear. But why can't it be informed of such facts by *having* representations without being informed *of* the representations themselves?

The ground I gave for supposing that a creature must be able to self-ascribe its representations was that only a creature which can take this sort of "step back" from its representations can reflect on their rational significance. But what exactly is "reflection," and why should it require self-consciousness? Certainly a creature that has no concept of itself and its own representations cannot reflect on the rational significance of *its having* a given representation (i.e., on considerations of the form "I *R* that *p*"), since by hypothesis it cannot think about itself and its representations at all. But why should this entail that such a creature cannot reflect on the rational significance of *its representations* (i.e., on considerations of the form "*p*")? Plainly, if this transition is to be more than a non-sequitur, the metaphor of stepping back will need to be supplemented by a more exact account of the relationship between rationality and self-consciousness.

5. A LOOK AHEAD

In the next three chapters, I try to answer these challenges: the next chapter addresses the former, while the latter occupies Chapters III and IV. The answers I suggest are the ones that seem right to me, and although I try to trace a Kantian lineage for the main points in my argument, much that I say has no direct counterpart in Kant's texts. To claim that the argument I present is Kant's argument would thus be preposterous – obviously so, since the objections to the Kantian thesis that I consider reflect the philosophical preoccupations of our time, not his. Nevertheless, I do mean my argument to be Kantian

in spirit, in the following sense: I mean it to bring out a set of relationships between reason, judgment and self-consciousness, relationships which I think Kant recognized, and in view of which various of his characteristic thoughts acquire a motivation and a pattern. So although I do not exactly put forward what follows as a reading of Kant, I do mean it to make him more readable – to make it easier to see why he says the kinds of things he does about the significance of first-person thought.

II. KNOWLEDGE, BELIEF, AND GROUNDS FOR BELIEF

> [W]e cannot possibly think of a reason that consciously lets itself be directed from outside as regards its judgments; for in that case the subject would ascribe the determination of his faculty of judgment not to his reason, but to an impulse.
>
> Kant, *Grounding for the Metaphysics of Morals*, III (Ak. 4:448)

1. INTRODUCTION

In the preceding chapter, I sketched an intuitive rationale for Kant's claim that it must be possible for me to accompany all of my representations with "I think." The rationale involved two claims: (1) that the only reasons that are relevant to my life as a knowing being are reasons with which I am presented; and (2) that a representation only presents me with a reason if I can self-ascribe that representation. At the end of the chapter, a difficulty was raised for each of these claims. The purpose of this chapter is to address the first difficulty; the second will be addressed in Chapters III and IV.

The first difficulty was that, although a subject's belief must in some sense be guided by relevant information if it is to count as knowledge, it is not obvious that this requires that the subject be *presented with* reasons for her belief. The claim that only reasons with which a subject is presented can bear on the question of what she knows amounts to a form of *internalism* about knowledge – a form of the doctrine that a subject's belief can only count as justified, in the sense relevant to knowledge, if she can produce the relevant justification. But such internalism has been challenged by epistemological *externalists* – philosophers who hold that the facts in virtue of which a subject's belief counts as knowledge need not themselves be known to the subject. According to one influential form of externalism, for instance, what matters for knowledge is that the subject's true belief be the output of a mechanism that reliably

produces true beliefs about the subject-matter in question. If this sort of "reliabilist" account of knowledge is correct, then it looks as though a subject could count as knowing that *p* even if she were not presented with *any* reasons for believing that *p*. For what matters, on this view, is not the subject's ability to cite grounds for taking her belief to be reliable, but the actual reliability of her belief-forming mechanism.

My aim in this chapter is to fill in the argument connecting knowledge with grounds accessible to the subject, and, to this extent, to answer the externalist challenge. My primary concern, however, is not to adjudicate a dispute in epistemology. In fact, I think the real ground of the requirement that a subject be presented with reasons for her beliefs cuts across disputes about the analysis of knowledge. For, as I shall try to explain, the ground of this requirement lies in the very nature of belief, and in the way that a creature capable of asking itself the question "Why do I believe that?" must relate to its own beliefs. But both sides in the internalism-externalism debate concede that, to count as knowing that *p*, a subject must at least *believe* that *p*. Hence, if my argument succeeds, it will set a constraint on knowledge that is prior to, and at least partly independent of, the issues about justification that cause controversy among epistemologists.

The bulk of this chapter will therefore be devoted to an examination of the concept of belief. For the purposes of this examination, I will simply assume that the kind of creature under consideration is one that can reflect self-consciously on its beliefs – that is to say, one such that, for any proposition *p* that it understands, (1) it can understand the question "Do you believe that *p*?", and can normally respond to this question with an authoritative "Yes" or "No"; and (2) it can understand the question "Why do you believe that *p*?", where this is a question about what grounds it has for taking *p* to be true.[1] The question of what is involved in having these abilities, and how having them transforms a creature's

[1] For brevity's sake, I will sometimes speak of a subject's "taking *p* to be true" (and similarly of "a subject's attitude toward *p*", "the truth of *p*", etc.), although this sort of construction is strictly ungrammatical when a declarative sentence

cognitive life more generally, is of course one of the main concerns of this dissertation. I will ultimately want to argue that the ability to reflect self-consciously on one's own beliefs is a crucial cognitive capacity, the lynch-pin of theoretical rationality; but I will not offer any such argument in this chapter. Whatever the connection between our ability to reflect self-consciously on our own beliefs and our other cognitive capacities, it is at least clear that a normal adult human being, who has mastered a language that includes a first-person expression and a vocabulary of cognitive states, will have this ability. My purpose in this chapter will be to bring out some consequences that possession of this ability has for a person's relationship to his own reasons for belief.

2. BEING PRESENTED WITH A REASON

Before turning to these topics, however, I need to clarify a notion that I relied on, but did not really explain, in presenting the intuitive argument for the Kantian thesis. In the last chapter, I suggested that the only reasons that are relevant to my life as a knowing being are reasons *with which I am presented*, and I contrasted states that present me with reasons with states that influence me "subliminally." But what is it to be presented with a reason, and how exactly is this relationship to reasons relevant to cognition?

To be presented with a reason, in the sense I intend, is to have a fact made available to one's thinking, in such a way that one can take account of it in thinking about what to believe and what to do. It is to have a certain fact, as we say, within one's ken.[2] Perceiving that things are thus-and-so is a

---

is substituted for *p*. The alternative – constantly to be saying "the subject's taking it to be true that *p*", etc. – seems to me even more awkward, and I do not think my practice will cause any confusion.

[2] As I shall use the term, reasons are facts, so a proposition that figures in a subject's thinking about what to believe counts as articulating a reason only if it is true. This usage of course does not prevent us from recognizing that a subject can, e.g., be under the perceptual impression that something is the case when it is not. This would be a case of its merely seeming to her that she was being presented with a reason. To *seem to* be presented with a reason to the effect that *p* is to be in some mental state that purports to inform me that *p*. Purported information may or may not be *merely* purported information.

paradigm case: if I perceive that the air has a funny smell, or see that exhaust fumes are swirling around me, I am in a position to draw conclusions from these facts, in a way that I am not normally in a position to draw conclusions from the fact that the concentration of carboxyhemoglobin in my blood is rising. But although I take perceiving that things are thus-and-so to be a paradigm case of being presented with a reason, I mean the notion also to apply to other kinds of cases of a fact's being within a subject's ken. If, for instance, a subject remembers having perceived that *p*, or knows that *p* on some other basis (by inference, or testimony, or whatever), these also count for my purposes as cases of her being presented with a reason. In general, to be presented with the fact that *p* is to be in some factive epistemic state (seeing, knowing by inference, knowing by testimony, …) with the content that *p*. For a self-conscious creature, being presented with a reason will presumably involve being able to ascribe the relevant state to itself, and to invoke it in support of other beliefs. When such a creature is presented with a reason, in short, it will have a fact available for use in reasoning and self-justification.

We can begin to see the importance of this sort of relationship to facts (or apparent facts) by noting its connection with a certain kind of explanation of belief. In general, when a person believes that *p*, it makes sense to ask her why she believes it. There is more than one way to hear this question, but in the sense that is relevant for our purposes, it asks about the subject's grounds for believing that *p* is *true*. In this sense, "Because *q*" is a candidate answer to the question only if (1) the subject is (apparently) presented with the fact that *q* and (2) the subject believes that there is some connection between *q*'s being true and *p*'s being true. If either of these conditions is not met, then even if it is true that *q*, the fact that *q* cannot figure in this kind of explanation of a subject's belief: it cannot be the subject's ground for believing that *p*, for to take *p* to be true on a certain ground just is to take the grounding proposition to be a fact with a bearing on the truth of *p*.

To mark off this "Why?"-question and its corresponding "because" is not to deny that there can be other kinds of explanation of a subject's believing something. For instance, a subject's disposition to

believe that there is something green in front of her undoubtedly stands in some complex relationship to facts about the wavelengths of light reaching her eyes. One possible answer to the question why a person believes that there is something green in front of her, therefore, might be "Because light of such-and-such a wavelength is reaching her eyes." Or to take another kind of case: persons sometimes come to believe that *p* as a result of wishful thinking, because they fervently wish that *p*, or cannot bear that thought that not-*p*. We might explain such a person's belief by saying "She believes that *p* because she desperately wants it to be the case that *p*." It is obvious, however, that neither of these sorts of explanation gives the subject's *grounds* for belief. The fact that light of a certain wavelength is reaching my eyes is not ordinarily my ground for believing that there is something green in front of me, for I am not normally aware that light of a certain wavelength is reaching my eyes, except perhaps indirectly, in virtue of finding that the thing in front of me looks green, and knowing something about the range of wavelengths of light that normally produce this impression. This cannot, then, be *what convinces me* that there is something green in front of me, for it is not something I am in a position to know prior to reaching my conclusion about the color of what is in front of me. But equally, the fact that a subject wishes that *p* cannot be her ground for believing that *p*, for a subject's wishing that *p* is not (except in special cases) any reason to think that *p* is true, and no sane subject will believe that it is. The two explanations just mentioned thus do not cite the kind of cause that I am calling "a ground." They correspond to other ways of hearing the question "Why does *S* believe that *p*?" – ways that do not inquire about the grounds on which *S* believes that *p* is true.

When I speak of a subject's belief that *p* being *guided* by the consideration that *q*, I will mean that the relevant consideration could figure in a true answer to the "Why?"-question just identified. Now, taking "guided" in *this* sense, we can affirm the following principle:

(P) A consideration can only guide my thinking if I am presented with that consideration.

(P) says nothing controversial; it simply follows from our characterization of the relevant "Why?"-

question. But the point noted in (P) will be epistemologically significant to whatever extent this "Why?"-question, and the form of explanation associated with it, prove to be epistemologically significant. The purpose of the remainder of this chapter will be to bring out the relevance of this sort of explanation to the status of beliefs as knowledge.

3. INFERENTIAL KNOWLEDGE, PERCEPTUAL KNOWLEDGE, AND RELIABILISM

One approach to making out this relevance would be to stress that, if a belief is to count as knowledge, the believer must have some justification for what she believes, and must believe what she does precisely because she has this justification. Intuitively, there is a difference between coming to know something and merely coming accidentally to believe something true: for a belief to count as knowledge, it must depend in the right way on the fact it is a belief about. Moreover, there is some plausibility in the thought that, at least for certain kinds of knowledge, "the right way" involves the subject's being presented with the relevant fact and having her thinking guided by it. When a subject comes to know something on the basis of reasoning, for instance, it certainly seems that she must meet this requirement: the facts that justify her reasoning must be facts of which she is aware, and she must come to the conclusion she does precisely because she is aware of these facts. To acquire knowledge on the basis of reasoning, in other words, involves having one's thinking guided in the sense subject to principle (P).

It is a good deal less obvious, however, that this principle has any interesting bearing on perceptual knowledge – which is, after all, the sort of knowledge with which Kant is primarily concerned when he claims that it must be possible for the "I think" to accompany all of my representations. In an ordinary case of perceptual awareness, it seems that the only fact that need figure in the explanation of my belief is the very fact I come to believe obtains. I may come to know, for instance, that the table in front of me is green, not by reasoning to this conclusion from other known facts, but simply by seeing

that the table is green. But then the suggestion that I must be presented with this fact in order to know it looks either trivial or wrong: trivial if it means that I don't know that the table is green unless I know it; wrong if it means that before I can have a basis for believing this, I must *already* know it.

No doubt my belief that the table is green must, if it is to count as knowledge, be in some sense a response to the fact that the table is green, but just what sort of responsiveness is required here is a vexed question. As I mentioned in the introduction to this chapter, many contemporary epistemologists hold that one is justified in one's perceptual belief that *p* just if (roughly) one's belief is caused by a mechanism that reliably produces true beliefs about environing circumstances.[3] If this view is correct, it is hard to see how an argument connecting the notion of justification relevant to perceptual knowledge with principle (P) could get off the ground. Reliabilists will concede that, where *p* is a fact about my environment, my perceptual belief that *p* should only count as knowledge if I believe that *p* because it is so. But they will deny that this requires that my thinking be *guided* by the fact that *p* in the sense that requires my being presented with the fact that *p*. On the contrary, they will insist that I only come to *be* presented, perceptually, with the fact that the table is green when I am brought to believe that the table is green by a reliable mechanism, and in that case I already have knowledge and do not need further reasons.

Anyone sympathetic to the Kantian view of mind will want to object that this reliabilist picture falsifies the phenomenology of perception. Our perceptual beliefs do not seem just to come to us, as if from nowhere; they seem motivated and justified by the things we perceive. According to the reliabilist, however, this appearance is an illusion: our perceptual beliefs really do "just come to us" as the output of

---

[3] For influential formulations, see Goldman 1976 and Dretske 1981. Of course, not all philosophers who are sympathetic to the reliabilist intuition hold that truth plus causation by a reliable mechanism *suffices* for knowledge; some attempt to combine sympathy for reliabilism with a measure of respect for the Kantian thought I am concerned to defend. For a sophisticated attempt to accommodate both thoughts, see Brandom 1994, 1995 and 2000. I am skeptical of Brandom's attempt at reconciliation, but I will not attempt to justify that skepticism here. My aim is more basic: I want to explain why we should feel obliged to accommodate the Kantian thought at all. Given this aim, the *pure*

a mechanism whose workings we do not oversee. On this view, our epistemic situation with regard to our perceptual beliefs is not essentially different from the situation of a blindsighted person: we find ourselves with certain beliefs about our environment, and may perhaps have reason to think that these beliefs reliably track the truth, but we do not obtain in perception some further justification for these beliefs.[4] Whatever the significance of the fact that ordinary perception has a distinctive "phenomenology," whereas blindsighted information-gathering does not, this fact does not, according to the reliabilist, reflect a fundamental difference in the sort of justification we acquire for believing what we do.

The observation that the reliabilist picture of perceptual knowledge is committed to denying that ordinary perception supplies us with a distinctive kind of justification for belief is certainly suggestive, but by itself this observation is no objection to reliabilism. The claim that ordinary perception involves our being informed of facts in a way that a blindsighted person is not, and that this gives ordinary perceivers a kind of justification that blindsighted persons lack, is just what the reliabilist denies: to treat the phenomenological difference between the two cases as though it proved the existence of an epistemic difference is to fail to join the issue with him. Indeed, the whole strategy of arguing against reliabilism by appeal to intuitions about justification seems unpromising. So long as our question is "What must be the connection between a belief and a fact in order for the belief to count as knowledge of that fact?", the reliabilist position will be tempting. After all, for a belief to be generated by a reliable mechanism is certainly for it to be connected nonaccidentally with the fact it concerns. To concede this point but continue to insist that something else is required for knowledge – not just some further refinement of the reliabilist criterion, but something else altogether – will inevitably look to tough-minded philosophers like mere romanticism: an appeal to an obscure "something extra" that sets

---

reliabilist becomes a particularly significant adversary, for she denies that the sort of justification relevant to knowledge has anything to do with grounds accessible to the subject.

genuine knowledge apart from counterfeits. What, it will be asked, could this something extra be, and what could be its importance relative to the task of living flourishingly in the world?

4. BELIEF AND GROUNDS FOR BELIEF

My aim is not to answer this challenge, but only to prevent the apparent possibility of a reliabilist analysis from blocking the intuitive argument for the Kantian thesis. We have been trying to find an argument that would connect the possibility of coming to know facts in virtue of having representations with the following principle:

(P) A consideration can only guide my thinking if I am presented with that consideration.

Our initial approach was to try to connect the notion of guidance that figures in (P) with the requirement that a belief must be justified if it is to count as knowledge. The connection seemed evident enough in cases where the justification for a belief involved reasoning, but it seemed less evident in the cases of primary concern to us, cases in which our claim to know rests simply on our having a certain perceptual representation and forming a belief on the basis of it. The possibility of giving a reliabilist account of justification in these cases made it look doubtful that principle (P) has any interesting bearing on this sort of knowledge, for the kind of nonaccidental connection between fact and belief invoked by reliabilists does not require that beliefs be guided in the sense of principle (P). So long as reliabilism remains a possibility, then, it appears that the task of arguing for a connection between justification for perceptual beliefs and principle (P) cannot succeed.

We might avoid this problem altogether, however, if we took a different approach to arguing for a connection between principle (P) and the capacity to acquire knowledge from representations. Rather

---

[4] I assume that the phenomenon of blindsight is familiar enough to philosophers that it is not necessary to summarize

than focusing on the notion of justification, I suggest that we turn our attention to the very idea of belief. Our difficulties over reliabilism, I think, are a symptom of the fact that we have been taking this idea too much for granted: we have been asking what is required for a belief to count as knowledge, but not what is required for a creature even to count as having a belief about a particular matter.

To have a belief is to have a specific sort of attitude toward a proposition, distinct from the sort one has when one imagines a proposition true, or supposes it true for the sake of argument, or has it as a "fixed idea." A number of philosophers have emphasized the point that if a person's attitude toward a given proposition is even to count as belief, the attitude must be responsive to her own judgments about her grounds for accepting the relevant proposition.[5] We can begin to see the significance of this requirement by trying to imagine what we would make of a person who maintained that *p* even while conceding that there was conclusive evidence that not-*p*. It would be hard to make sense of such a person while holding onto the idea that his attitude toward *p* was one of belief. If he admits that there is conclusive evidence that not-*p*, what can he mean by continuing to insist that *p*? Is the thought that *p* one that keeps occurring to him no matter how conclusive he knows the evidence against it to be? In that case, it seems to be some sort of obsession, not something he really believes. Is he playing a game of make-believe, to which the question whether it is actually true that *p* is irrelevant? That would certainly make him intelligible, but to pretend that something is the case is not to believe it. Is he being flippant, or insincere, or deliberately obtuse? Does he mean that he fervently wants to believe that *p*? Is he trying to make a philosophical point? Each of these hypotheses might shed light on his behavior, but each would jeopardize the idea that his assertion that *p* reflects genuine belief.

We could of course ask the person in question to explain himself, and he might do so. But if he

---

the characteristics of the syndrome. For such a summary, see Weiskrantz 1986.

[5] The point is made forcefully in Stuart Hampshire's *Freedom of the Individual* (1965), Chapter 3, and is one of the leading ideas of Richard Moran's recent book *Authority and Estrangement* (2001). It also figures importantly, although in a different way, in Donald Davidson's writings on belief and interpretation (see the essays grouped under the heading "Radical Interpretation" in Davidson 1984).

continued to insist, in apparent sincerity, that he really *believed* that *p*, and also that he really believed there was conclusive evidence that not-*p,* I think we would have to question either his understanding of the concepts of belief and evidence, his grip on the very idea of an objective world, or (if this is something different) his soundness of mind. To say that we would have to question one of these things is not to say that there is nothing he could go on to say that would put our question to rest. Perhaps he subscribes to a paraconsistent logic, or to a form of mysticism which doubts the capacities of human reason and recommends as a corrective that he believe the very opposite of what his reason tells him: such convictions might give his combination of attitudes a tenuous coherence. The important point, however, is that absent *some* special explanation, the combination of attitudes he purports to hold would be unintelligible, for to *believe* that *p* just is to take it to be true that *p*, and a subject who admits that there is conclusive evidence that not-*p* seems thereby to have admitted that he *does not* take the proposition that *p* to be true. Nor, I should emphasize, is this merely a point about what a person can coherently say about himself: the tension is between the attitudes themselves.

Furthermore, although the example we have been considering is outlandish, the point it brings out is not restricted to outlandish cases. The tension between believing that *p* and believing that there is conclusive evidence that not-*p* depends simply on the fact that one's belief about *p* is one's "take" on the truth of *p*. But this fact implies that the question whether a subject believes that *p* cannot *ever* come entirely unhinged from the question of the subject's assessment of considerations she takes to bear on the truth of *p*. A subject need not suppose that there is conclusive evidence that not-*p* in order to face a threat to her ability coherently to believe that *p*: if she allows that there is *any* evidence that not-*p*, she cannot consciously treat this evidence as irrelevant to her maintaining that *p*, on pain of jeopardizing the claim of her attitude toward *p* to count as belief.[6] Moreover, even when a subject does not have any

[6] The possibility of self-deception may seem to disprove this claim. Does not the self-deceiver characteristically recognize that there is evidence that not-*p* and yet continue to believe that *p* without heed to this contrary evidence? I

evidence that not-*p*, she cannot in general just heedlessly form the belief that *p*. The constraint here is not empirical: an attitude toward a proposition formed or maintained in a manner that was consciously heedless of the evidence for *p* would not, special circumstances apart, count as a belief, for to believe that *p* just is to take *p* to be true, and to consciously pay no heed to evidence is to ignore one's own assessment of what speaks in favor of *p*'s being true.

To see the nature of the limitation here, imagine a person who maintained that *p* while admitting that she had no reason to believe that *p* was true. The question would arise: if she admits that she has no reason to believe that *p* is true, how can she insist that *p*? Once again, we can imagine various sorts of answer that would explain her insistence but would undermine our basis for saying that her attitude toward *p* was one of belief: her apparent conviction might really be an obsession, a mere fancy, a lie, a wish. It may seem, however, that there is at least one possible explanation of her insistence that is perfectly consistent with her believing what she says. Suppose we ask her why she insists that *p* and she says "I don't have any reason; I just feel sure." This response seems at least intelligible, and its intelligibility may seem to demonstrate that it is possible for a subject to hold a belief for no reason whatsoever. But before conceding this conclusion, we should think carefully about the kinds of context in which the remark "I just feel sure" makes sense.

One thing such a remark might express is the conviction that one's feeling is *itself* evidence that *p*. Someone who believed that she had psychic powers, for instance, might intelligibly regard her own inclination to think that *p* in this light: when she conceded that she had no reason, initially, to have the hunch that *p*, she would not be conceding that she has no reason, now, to believe that *p* is true. Her case would not, therefore, disprove my claim that an attitude that paid no heed to evidence would not be a

---

do not deny that such a thing is possible; what I deny is that it is possible *in full consciousness* – possible without deceiving oneself. Even the self-deceiver must take her belief to be true, on pain of its not counting as a belief at all. She must, therefore, somehow put the part of herself that genuinely appreciates the force of the contrary evidence out of contact with the relevant belief. (In speaking here of parts of the self, I mean only note the phenomenon that needs explaining, not to offer a theory that would explain it. I will say more about self-deception in Chapter VI.)

belief.

Can we conceive of a different sort of case in which a person who did not believe herself to possess any occult powers "just felt sure" that something was the case? The most natural circumstance to imagine, I think, is one in which some outcome is hoped for or feared: "I just feel sure that the gold is buried here (that the roulette wheel will come up red, that he'll injure himself going over that jump)." Our grounds for saying that the speaker really believes what she asserts will then be her willingness, e.g., to risk something she values on its truth. In this sort of case, however, the line between a genuine belief and, say, a dogged assumption seems vanishingly fine. True, the person in question is prepared to act on her "conviction," which gives some point to our calling it a belief; but precisely because she is conscious that her attitude is held groundlessly, there is also some reason to look for another classification. What this shows, I think, is that such baseless convictions are at best marginal cases of belief, and necessarily so.[7]

---

[7] Some philosophers would argue that the attitude I have been describing is not ordinary belief, but something closer to a stance they call "acceptance." Ordinary belief, they would claim, is not so closely connected with grounds-for-thinking-true as I have been suggesting. Keith Lehrer, for instance, defines acceptance in terms of a subject's willingness "to assent to [a proposition] when [his] only purpose is to assent to what is true and to refuse to assent to what is false," and argues that it is not necessary that a subject either accept everything he believes or believe everything he accepts. For, Lehrer asserts:

> What a person believes is not entirely up to him. One is endowed with certain beliefs, and one may conclude that some of what one believes one should not accept as a truth-seeker (Lehrer 1979, pp. 65-6).

I have no quarrel with Lehrer's definition of acceptance, or with his claim that acceptance thus defined is one thing and belief is another. Lehrer's way of spelling out the difference between these stances, however, seems to me to presuppose an untenable conception of belief. It is true that one does not form beliefs *for a purpose*, even the purpose of believing what is true. Forming a belief is not an action in the sense that assenting to a proposition is an action: it is not something one can decide to do (although one can intelligibly intend to *bring it about* that one forms a certain belief, for instance by the Pascalian method of putting oneself in situations which conduce to one's acquiring that belief). In this sense, I grant that what we believe is not "up to us." But these observations hardly show that one could find oneself simply "endowed with" a belief although one recognized that "as a truth-seeker" one should not accept it. A "belief" that was in this way heedless of my judgments about truth and evidence would be a pathological phenomenon, which we could only make sense of by positing something like a schism in the self. Absent such a schism, to conclude that "as a truth-seeker" I should not accept the proposition that *p* just is to renounce my belief that *p*.

In support of his conception of belief, Lehrer cites such facts as the following: (1) we sometimes say of a person that she "cannot help believing" something despite being aware of strong considerations to the contrary, and (2) people sometimes form beliefs on grounds (e.g., tarot, ouija) that it is hard for an unprejudiced observer to regard as evidential. But surely these observations do not show that one can find oneself simply "endowed" with beliefs that one consciously does not accept. A person who believes what the ouija board tells her must presumably believe, for

I do not wish to deny that a person whose beliefs are in general truth-heeding can on occasion give credence to a mere inkling. The very fact that an inkling is unaccountable may lead a person to think "There must be something in it," especially if her sense of what is reasonable is influenced by a wish or some other passion. But if a subject does give credence to such an inkling, what she has done is to take her inkling (e.g., that she is being watched, or that the roulette wheel will come up red) *to be true*, and the point I would stress is that taking to be true is only intelligible against a certain background. After all, to take a thought to be true is not just to have certain words or pictures pass before one's mind. Hume held that beliefs differ from mere figments of the imagination only in being accompanied by a certain "feeling" or "sentiment," but even he admitted that the relevant feeling is only definable as the one characteristic of *belief*.[8] Our question should be: Could the relevant sort of feeling conceivably accompany just any thought, regardless of our own assessment of our evidence for it? Can we take just anything to be true?

Try the following experiment: suppose that there is someone standing behind you holding a mallet. It seems easy enough to *suppose* this, and in supposing it you have already entertained a proposition that is true just if there is someone standing behind you holding a mallet. But now try to *take it actually to be true* that there is someone standing behind you holding a mallet – to believe this situation to obtain. Is it merely a contingent fact about us that we cannot readily bring this about, cannot

---

whatever reason, that the board tells the truth. She may admit that other evidence points to a different conclusion, but if she does not believe that the board's answer shows this evidence to be misleading, then she does not *believe* the board's answer. Similarly, someone who "cannot help" believing that X is her true friend despite mounting evidence that X does not wish her well must find some way to discount the mounting evidence. Stepping back from her situation, she may note her tendency to discount evidence of X's unfaithfulness, stressing its perversity by saying that she "cannot help" believing what she does. Nevertheless, if she sincerely concedes the decisiveness of the mounting evidence that X does not wish her well, she by that very concession changes her mind and disproves her claim of incapacity. She may subsequently revert to her former opinion, but if she does, this will involve reevaluating the weight of the evidence (perhaps wishfully), not merely ignoring it. Otherwise, however forceful or fervent her attitude toward the proposition that X is her true friend, what she has is not a belief.

[8] "I confess, that 'tis impossible to explain perfectly this feeling or manner of conception. We may make use of words, that express something near it. But its true and proper name is belief, which is a term that every one sufficiently understands in common life. And in philosophy we can go no farther, than assert, that it is something felt by the mind,

summon up a belief by, say, squinting in a particular way?

If the feeling we had when we "just felt sure" of something were just a particular phenomenological profile, a certain coloring of consciousness, then this profile ought in principle to be able to accompany just any thought.[9] In that case, it would be hard to see how it could be anything but a contingent fact about us that we cannot produce this feeling in ourselves in connection with any thought whatsoever. But surely it is not merely a contingent fact about us that we cannot get ourselves to believe without regard to truth. Even if squinting could summon up an impulse to picture a person behind me holding a mallet, or to blurt out the words "There is someone behind me holding a mallet," or even – allowing this for the sake of argument – to believe that there was someone standing behind me with a mallet, consciousness that I had no grounds for taking this to be true would necessarily undermine the impulse.[10] Any sort of force or vivacity that *could* persist in the face of such consciousness, any impulse that persisted in spite of my awareness that it was groundless, would not be a normal belief, but a miniature psychosis: it would not reflect my take on the truth.

Perhaps someone will want to object that this is simply a stipulation about what we should call "belief." What rules out a different understanding of the term? Well, nothing rules it out. My point is

---

which distinguishes the ideas of the judgment from the fictions of the imagination" (*Treatise of Human Nature* [1978], Appendix, p. 629).

[9] Although I have framed this paragraph as an objection to the Humean suggestion that what makes a thought count as a belief is just its being accompanied by a certain feeling (if this means: a particular introspectible character), an analogous point should apply to any view on which belief can in principle come apart from a subject's take on the truth. For example, many philosophers claim that the state of belief can be characterized by its specific causal-functional role. In the course of defending such a conception, John Heil remarks that "If one takes belief exemplifications to be states of a certain sort (perhaps states with certain kinds of causal properties) there seems to be no *a priori* reason why a belief could not be created by 'directly' willing it" (1983, p. 358). If the considerations rehearsed here are convincing, they should work as a *modus tollens* against any form of functionalism that implies such a conditional.

[10] Projects of self-manipulation are possible, of course, but such a project can only succeed by providing me with considerations which I take to be epistemic reasons for believing what I do. Awareness that I have simply induced a given thought in myself would seem to preclude this. (The point I am making here is obviously related to points made by Bernard Williams in his seminal paper "Deciding to Believe" [1973]. But the point is stressed by Kant himself in his *Logic*: "The will has no immediate influence on our holding-to-be-true; if it had, that would be very absurd.... [W]e would constantly make chimeras for ourselves of a happy state and then hold them to be true. The will, however, cannot contend *against* convincing proofs of truth that run counter to its wishes and inclinations" [L 9:73-4]. See also his discussion of holding-to-be-true in Doctrine of Method [A820-831/B848-59].)

not that it is impossible for us to conceive of a use for the term "belief" on which belief does not involve responsiveness to grounds-for-taking-true: we can of course use this *word* however we like. What I deny is that it would be an adequate stipulation to say that the relevant kind of belief is just like truth-heeding belief except that it is not truth-heeding. I have been arguing that we lack a clear idea what it would be for a creature to take something *actually* to be true (as distinct from merely imagining it, or supposing it for the sake of argument, or …) while consciously ignoring evidence of its truth or untruth – that we simply do not have a grip on the attitude of belief which can survive its removal from the context of responsiveness to grounds-for-thinking-true. In general, I suspect that philosophers who are willing to contemplate the possibility of beliefs that are heedless of evidence nevertheless want to conceive of the relevant attitudes as reflecting the subject's take on the truth. My quarrel with such philosophers is not that they use the term "belief" in an impermissible way, but that they have not specified a clear use for the term at all.

5. PRINCIPLE (P) AND THE NATURE OF BELIEF

I hope the connection between these remarks and principle (P) is already palpable, if not yet clear. Principle (P) says that a fact with which I am not presented cannot guide my thinking, in the sense of "guide" defined earlier. The upshot of the last section is that an attitude which was not formed and maintained in a manner responsive to relevant evidence would not be a belief at all, or would at best be a special and marginal case of belief. It should be clear that the notion of responsiveness involved in the latter point is closely related to the notion of guidance involved in the former. To say that *S*'s thinking can only be guided by facts with which she is presented is to say that "Because *q*" can be a true answer to

the question "Why does *S* think that *p*?" (taken in our sense) only if *S* is aware that *q*.[11] To say that beliefs must be held in a manner responsive to relevant evidence is to say, among other things, that if a subject is conscious of having no satisfactory answer to the question "Why do you think that *p*?", then, except in special circumstances, her belief that *p* must cease (or her attitude toward *p* must cease to count as one of belief). The latter point implies that it is no accident that I am normally aware of the grounds for my beliefs in a way that permits them to guide my thinking. A conviction that arose in a subject despite the fact that she had no satisfactory answer to the question "Why do you think that?" would be *unstable* in an important way: it would be liable to collapse under the subject's own scrutiny.

"Liable to collapse" is obviously a vague formulation, and even in this vague form my claim may seem subject to counterexamples. Let me try to clarify what I mean by mentioning some things that I do not mean.

In the first place, I certainly do not mean to suggest that only beliefs which are the outcome of a process of conscious reasoning will be stable under scrutiny. Plainly, we arrive at many of our beliefs without reasoning, and indeed without occurrent awareness of having formed a belief.[12] Such unconsidered beliefs are not necessarily more unstable than those we form by conscious reasoning, for they may arise in us in a manner that, although not consciously directed, is nevertheless responsive to what are in fact good grounds for belief. But notice that, as I have defined the notion of a belief's being guided by a fact, a subject's formation of an unconsidered belief that *q will* count as guided by the fact that *p*, so long as (1) she is aware that *p* and (2) "Because *p*" would be a true answer to the question "Why does she think that *q*?" This is not to require that the subject should be conscious of the process

---

[11] Or at least: only if *S* was aware that *q* when she came to believe that *p*. It is certainly possible for a person to forget what convinced her of something; the important point is that a fact of which a person is not aware cannot be her reason for *forming* a belief. Nor, obviously, can a forgotten fact provide her with a reason for retaining a belief if the question whether to retain it arises.

[12] For many of our beliefs, the question when we formed them seems not even to the point. When did I form the belief that my car is still where I parked it yesterday? Surely not at the stroke of midnight, when I was fast asleep. Yet if I had

by which her belief is formed, only that she should be aware of the fact that *p* and be convinced that *q* on that account. A rational creature's capacity to be convinced of something on a certain account will certainly be connected with her *ability* to reflect on the grounds for her conviction. But this requirement that she be able to reflect on her grounds (which might be cashed out in terms of various counterfactuals concerning the ground she *would* give if asked) need not imply that, whenever she is convinced that *q* because *p*, she must *actually* have reasoned from *p* to *q*. Just as a person can act for the sake of a certain end even though she does not deliberate before acting, so a person can believe something for a certain reason even though she never reflects on her reasons for belief.

A subject's belief would be unstable, in the sense I have in mind, only if it were formed in the absence of any grounds on which she was *able* to reflect. What must be imagined is that a subject finds herself with the impulse to suppose that *p* although when she puts the question "Why should I think that *p*?" to herself, she can find no satisfactory reason for her conviction. In this case, I claim, her conviction will be liable to collapse: she will be unable, except in special circumstances, to see it as her take on the truth, and hence unable to believe it.

This may seem to imply that a rational subject who realizes that she cannot recall her grounds for believing that *p* necessarily will or should lose confidence in that belief. But again, this is not what I mean. Suppose it occurs to me that, although I believe that my friend Sam has a sister, I do not recall my grounds for believing this: whether he told me, or I heard it from someone else, or got the information in some yet other way. Recognizing this need not make me give up my belief: I may simply trust that I got the information somehow. Nor need my self-trust be irrational: I may have good reason to be confident in my ability to retain information, either in general or in this case. Indeed, it is hard to see how we could get by without constantly trusting ourselves in this way. We seem to hold very many of our beliefs (for instance, those arrived at by induction) without being able to recall in any detail what

---

for some reason been awoken at that hour and taken to the spot where I left my car, and on arriving had discovered that

convinced us, and it would plainly be inefficient for us to try always to keep track of our justification for everything we believe.[13]

These observations seem undeniable, but what I have claimed does not commit me to denying them. My point is that a belief that "just came to me," a belief my formation of which bore no relation to my being aware of considerations that would support it, would be in a special way in jeopardy: the thought that I had sound reasons for holding such a belief – even reasons I could not now recall – would be a thought for which I could not have good grounds. The difficulty, then, is to see how a subject who grasped that this was how matters stood could persist in holding such baseless beliefs – how she could continue to think of them as her take on the truth.

Finally, I do not mean to suggest that just any belief will be liable to collapse in the absence of grounds. There are classes of well-founded beliefs to which the question "Why do you think that?" has no straightforward application. If I am asked why I believe that 2 + 2 = 4, or that the earth is more than five minutes old, I do not know what to say, although I certainly do believe these things, and think myself reasonable to do so. Recognizing that I cannot produce, or even imagine producing, a specific reason for thinking these propositions true does not undermine my belief that they are true – nor should

---

the car was gone, I would have been greatly surprised.

[13] Gilbert Harman gives a persuasive defense of these claims in Chapter 4 of his *Change in View* (1986). But although I concede these points, I should emphasize that I do not concede the moral Harman draws from them, namely that a subject should stop believing that *p* only if she "positively believes that her reasons for believing *p* are no good." (Harman calls this "The Principle of Positive Undermining": see p. 39.) My objection to this principle is not so much that it is mistaken as that, in speaking as though a subject could follow a *policy* about when to stop believing things, it makes no sense. One could follow a policy about when to "stop believing," presumably, only if one could in general *decide* what to believe on grounds unrelated to one's assessment of the likelihood that what one believes is true. I hope by now my objections to this way of thinking about belief are clear. (Perhaps Harman means his Principle of Positive Undermining not as a norm of reasoning in the sense of a norm specifying what a rational believer *ought to do*, but rather as a norm describing what a rational believer's mental tendencies *ought to be* – that is, as a description of what patterns of belief-revision are generally adaptive in the task of coping with the problems of a human life on Earth, where this adaptiveness is supposed to figure, perhaps, in an evolutionary explanation of the persistence of the relevant tendencies. If this is his point, however, then the ground of the relevant "ought" plainly does not lie in the concept of belief, and his "principle of rational belief-revision" is not the kind of imperative that can give me epistemic reasons for belief. For the fact that a certain mental tendency is generally adaptive, and was selected by evolution for that reason, obviously does not imply that in a given case a belief formed in accordance with the tendency is more likely to be *true*.) I am grateful to Anton Ford and Lionel Shapiro for helpful conversations about Harman.

it. These convictions occupy as fundamental a position in my thinking as any beliefs I hold: I do not have a clear idea of what it would *be* for me to have some specific reason for thinking such things true, something more firm on which my confidence in these beliefs could rest. But nor do I (nor, I think, does anybody) have a clear idea of what would show them to be false. We may say, if we like, that such beliefs are "baseless," or that we have no specific reason for holding them. But they are not baseless in the sense relevant to my argument: they are not beliefs concerning which (1) it is clear what a satisfactory reason for holding them would be and (2) the subject in question has no such reason.[14] So although *these* beliefs are not liable to collapse when I recognize that I lack specific grounds for thinking them true, this observation has no tendency to show that just *any* belief could be sustained in the recognized absence of grounds.

6. JUSTIFYING PERCEPTUAL BELIEFS

It seems clear that the question "Why do you think that?" *does* normally have application to beliefs about one's environment acquired through perception. If I am asked why I think that the table in front of me is green, I am not at a loss for an answer. It would be odd for someone who knew that the table was in plain view and that I was looking at it to ask me such a question, but only because the answer would be obvious: I believe it is green because I *see* it, and can see that it is green. If my judgment is challenged, I may defend it by pointing out that the lighting conditions are normal, that I am by all accounts a competent judge of colors, that to the best of my knowledge there is nothing wrong with my eyes, and

[14] Something broadly similar applies to authoritative beliefs about one's present mental state (e.g., the belief that one is presently in pain): they are baseless in the sense that the question "Why do you think that?" does not have application to them, not in the sense that it has application but normally lacks an answer. Of course I know what it is for me to be in pain, and I know what would count, from a third-person standpoint, as evidence that I am in pain. But I do not know what it would be for me to think that I am in pain and yet not be in pain, or what it would be for me to need to look for

so on. But the primary answer to the question "Why do you think that?", put to me in this context, is that I have a capacity, sight, which is now being actualized and which gives me the information that is the basis for my belief – the information, namely, that the table is green.

Notice that it is perfectly natural to speak here of sight providing me with information and of my basing my belief on that information. This is precisely what it is not natural to say of a blindsighted person: such a person does not receive visual information in a way that allows her to base her beliefs *on that information*. Her perceptual beliefs simply come to her – or rather, inklings about how things stand in her environment come to her, and if she knows on other grounds that these inklings are reliable, these inklings may give her reasons for belief. A blindsighted person's justification for belief is thus *structurally* different from the justification available to an ordinary perceiver: whereas facts (or seeming facts) present themselves to ordinary perceivers, blindsighted persons are presented only with their own inklings, their own impulses to take things to be true. This observation shows that the phenomenological difference between blindsight and ordinary perception is not just superficial, not "merely a matter of phenomenology." It reflects a basic difference in the kind of answer to the question "Why do you think that?" made available by these two faculties: the difference between the answer "Because I see that it is green" and the answer "Because I have an inkling that it is green, and know that my inklings in these matters are reliable." Whereas the former answer treats my sense-impression of a green table in front of me as by itself a reason to believe that the table in front of me is green, the latter treats my inclination to suppose that there is a green table in front of me as a mere bit of data, which only becomes a reason for belief (a ground for taking the relevant content to be true) when conjoined with the additional premise that my inclinations in these matters are reliable.

My purpose here is not to refute the kind of externalist who likens our epistemic situation to that of a blindsighted person, but I do think that the argument of this chapter leaves this sort of position

---

reasons for thinking that I am in presently pain. I do not know what a reason for me to think that I am presently in pain

looking awkward. For suppose that I find myself under the perceptual impression that there is a green table in front of me. Being self-conscious, I can ask myself whether I have reason to believe what my senses seem to be telling me. If I cannot see my perceptual impressions as grounds for my belief, then, as I have been arguing, my very ability to believe my senses will be in jeopardy. But according to epistemic externalists, a crucial component in a subject's warrant for relying on her own perceptual impressions need not be accessible to her. The difficulty then is to see how a reflective subject, conscious of a justificatory gap here which she cannot fill, can continue to give her perceptual impressions the *pro tanto* weight that we in fact give them. It is common for externalists to brush aside this sort of worry, arguing that it is part of a skeptical problematic in which they simply have no interest. But this can only reflect a profound alienation from the perspective from which we ordinarily make up our minds about what is the case: for a subject capable of reflecting on her own grounds for belief, the question of whether her own grounds are any good is not optional.[15]

7. CONCLUSION: THE STATE OF THE ARGUMENT

Kant claims that a representation which I could not accompany with "I think" would be nothing to me as a knowing being. My aim in this chapter has been to bring out why this is so, at least for creatures that can reflect self-consciously on their beliefs. The basic thought is this: a creature that can reflect self-consciously on its beliefs can put to itself the question "Why do I think that *p*?" (where this question is

---

would be.

[15] A similar alienation from the perspective of self-conscious reflection is exhibited in a number of recent defenses of the thesis that "nonconceptual" perceptual states can justify belief. Critics of the idea of nonconceptual justifiers commonly claim that a nonconceptual state cannot provide *the perceiving subject* with a reason for belief (see McDowell 1994, Lecture III & Afterword, Part II, and Brewer 1999, Chapter 5). In an important discussion of this claim, Richard G. Heck professes mystification about what the demand that perceptual states provide the subject with reasons for belief can amount to, if it does not just mean that the contents of these states must entail the beliefs that they justify (see Heck 2000, §3). But a state whose content entails a given belief is of no help *to the subject* in justifying her belief to herself if the

taken in the sense described in §2 above), and whether it persists in believing that *p* will depend on whether it can justify its belief to itself. Now, obviously, the only considerations that can figure in such a justification will be considerations of which the creature is aware – considerations it can accompany with "I think." A representation that such a creature could not accompany with "I think" would thus be nothing to it *in such a process of self-justification*. But I have been arguing that, in general, the very sustainability of our beliefs requires that we should be able to justify them to ourselves in this way. If this is right, then the characteristic thesis of externalist epistemology – that a belief can count as knowledge in virtue of circumstances which a subject cannot cite in justification of that belief – is in a certain way beside the point. If a self-conscious subject is to hold attitudes that are even *candidates* to be knowledge, she must in general be able to justify those attitudes to herself. That, as I understand it, is the force of Kant's claim.

I have not tried to present my argument for this conclusion as an interpretation for Kant, but I do not think the points I have made are foreign to his way of thinking. We can see their Kantian provenance by reflecting on the passage from the *Grounding for the Metaphysics of Morals* quoted at the head of this chapter:

> [W]e cannot possibly think of a reason that consciously lets itself be directed from outside as regards its judgments; for in that case the subject would ascribe the determination of his faculty of judgment not to his reason, but to an impulse (G 4:448).

Kant goes on to draw consequences from this point for the understanding of practical reason in particular, but he clearly intends the point to apply to reason in all its forms. It is, I think, a fundamental tenet of his thinking about rationality, part of what he means in characterizing reason as a *spontaneous* faculty. Commentators often read the passage as asserting a fundamental incompatibility between our conception of ourselves as rational beings and the thought that we are part of the order of nature – as

---

content of that state is not accessible to the subject in self-conscious reflection on her grounds for belief, and the states whose contents Heck calls "nonconceptual" are precisely ones whose content is not accessible in this sort of reflection.

though we must suppose ourselves temporarily to depart from the natural order whenever we act for a reason. And then, rightly, they declare this doctrine to be mysterious. But on a better reading, I think, Kant's point is not so obscure.

To say that "we cannot possibly think of a reason that consciously lets itself be directed from outside as regards its judgments" is just to make the point on which I have been insisting: that a creature which can think about its reasons must be capable of seeing its judgments (and equally, as I have been emphasizing, its beliefs) as formed for reasons which it takes to justify those judgments. To become conscious that one's judgment has been "directed from outside" would be to find oneself inclined to think something in the recognized absence of a satisfactory reason for thinking it: in such a case, one could only regard the inclination to have the thought in question as something not subject to one's reason, something produced by an alien force acting on one's mind. Kant says that we "cannot possibly think of" a faculty of reason that consciously permits its judgments to be determined by such a force, and that, by my lights, is exactly the right thing to say: such determination is *inconceivable*, for an affirmation that a subject could not see as its take on the truth would not be a judgment, and an attitude that a subject could not see as its take on the truth would not be a belief.

I have been arguing that one must be able to see one's thoughts as responsive to facts of which one is aware if one is to be able to see them as reflecting one's take on the truth. It follows that if representations are to figure as part of our grounds for belief, or for that matter for any other act or attitude that involves taking a stand on what is true, they must not simply produce in us blind impulses to believe; they must, rather, supply us with (apparent) reasons for belief, reasons whose sufficiency we can consider. This is what I mean when I say that representations must present us with reasons: they must present (seeming) facts to our rational powers in such a way that we can consciously take them into account in thinking about what is the case.

In reaching this conclusion, we have taken a first step toward filling in the intuitive argument for

the Kantian thesis: we have clarified why a creature that can put to itself the question "Why do I think that *p*?" must be able to accompany its representations with "I think." We have not, however, clarified what is at stake in the ability to put this kind of question to oneself – what, in a word, is at stake in self-consciousness. Kant, of course, claims that a creature without self-consciousness would lack the faculties of understanding, judgment and reason, the distinctive faculties of a rational being. To see what might justify this claim, we will have to examine Kant's views on the connection between reason and self-consciousness. This task will occupy the next two chapters.

III. REASON, JUDGMENT, AND SPONTANEITY

> Man is a thinker, and is universal, but he is only a thinker inasmuch as the universal exists *for* him. The animal is also *in itself* universal, but the universal does not exist *for* it as such; for it only the singular exists. The animal sees something singular, for instance, its food, or a man… Nature does not bring *Nous* to consciousness of itself until man first doubles himself so as to be a universal for a universal. This first happens when man knows that he is 'I'.
>
> Hegel, *Encyclopedia Logic*, §24, Zusatz 1

1. INTRODUCTION

In the preceding chapter, I argued that a self-conscious creature can only *believe* a proposition if she can see that proposition as expressing her take on the truth, and that this requires that she be able to consider why she believes what she does. In short, a self-conscious creature must be aware of her own grounds for belief. This thought provides us with one component in an argument for the Kantian thesis: it shows that if a subject is self-conscious at all, then she must be able to accompany the representations that ground her beliefs with "I think." What it does not show is why a rational creature must be self-conscious. This, in effect, was the second difficulty for the intuitive argument: why is self-consciousness crucial to rationality?

To answer this question, we will need some account of rationality, on the one hand, and of self-consciousness, on the other. The purpose of this chapter is to address the first of these two needs; the second will be addressed in the next chapter. I should hasten to add that, since my aim is only to say enough to bring out the connection between rationality and self-consciousness, my account of each will be highly schematic and incomplete. For one thing, I will focus almost exclusively on *theoretical*

rationality – the kind of rationality exercised in thinking about what is the case – while giving scant attention to *practical* rationality – the kind exercised in thinking about what to do. Moreover, I will for the most part concentrate on just one aspect of theoretical rationality, the capacity to *judge*. In this I follow Kant, who holds that all actions of what he calls "the understanding" (his name for theoretical reason insofar as it is directed toward sensibly-given objects) can be traced back to judgments, so that "the *understanding* in general can be represented as a *faculty of judging*" (A69/B94). If Kant is right to regard judging as the understanding's paradigmatic activity, and to regard the understanding as one of reason's central powers, then presumably we should be able to learn something about the nature of reason generally by reflecting on this particular capacity.

The question what reason is and what powers are distinctive of it was, of course, one of Kant's main concerns, and I think that his answers to these questions are highly illuminating. I will therefore be following him more closely in this chapter than I did in the last. I will begin, in the next section, by considering some remarks he makes about the difference between rational and nonrational creatures. These remarks will start to bring out the sense in which a rational being is free in a way that a nonrational creature is not, and thus will prepare the way for an examination of Kant's famous but obscure suggestion that, metaphysically speaking, what is distinctive of a rational understanding is its *spontaneity*. The task of understanding this suggestion, and, in particular, of understanding the idea that judging is an exercise of such spontaneity, will occupy the remainder of the chapter. In the next chapter, I will argue that the relevant kind of spontaneity necessarily involves self-consciousness.

Before proceeding, I should say a few words about the shape my argument will take. The claim that a rational creature must be able to accompany its representations with "I think" would be obviously trivial if "rational creature" just meant "creature capable of accompanying its representations with 'I think'." And it would be more subtly trivial if it turned out that the abilities that are supposed to characterize a rational creature – the abilities to think discursively, to judge and to deliberate – require the

capacity to accompany one's representations with "I think" only because they are understood in some specially demanding sense for which there is no obvious rationale. If, for example, "judging" were simply defined as forming a belief that one can self-ascribe on the basis of representations that one can self-ascribe, then it would be easy enough to argue that a creature capable of judging must be capable of apperceiving its representations; but the result would be trivial, for it would only postpone the question why this capacity for self-ascription is important. We can if we like give the conjunctive capacity to form beliefs apperceptively a special name, and argue that the capacity for apperception is essential to it, but we might as well give a name to the capacity to form beliefs while juggling bowling pins, and argue that the capacity to juggle bowling pins is essential to it: in neither case is it clear why the new capacity *matters*.

The thesis that a rational creature must be able to accompany its representation with "I think" will be nontrivial only to the extent that we can see a significant and irreducible *kind* of cognitive capacity to which the ability to accompany one's representations with "I think" is essential – a capacity which does not simply resolve into some independently-intelligible capacity that we share with creatures that lack self-consciousness plus the capacity to self-ascribe one's representations. What we need, in short, is an account of why a creature that has the capacity to apperceive its representations has made a significant cognitive advance. The principal aim of this chapter, therefore, will be to argue that the capacity to judge is such a significant and irreducible kind of cognitive capacity. It will be to show, in the terms I introduced in the Introduction to this dissertation, that a mind with this capacity is *categorially different* from one without it. If this can be shown, and if it can be shown that only a creature that can accompany its representations with "I think" can possess this capacity, then we will have clarified the connection between the capacity for first-person thought and the capacity for rational thought generally, the connection on which Kant famously insists.

2. RATIONAL AND NONRATIONAL CREATURES

It will be useful to begin by considering what Kant says about the role representations play in the mental lives of nonrational creatures. Kant's views on brute mentality have been a source of puzzlement to commentators. For on the one hand, he remarks in many places that brutes *do* have representations and are not mere mechanisms, as the Cartesians held. But on the other hand, there are important passages in which he appears to deny that a creature lacking self-consciousness can have representations at all. For instance, at a crucial juncture in the B-Deduction, he suggests that a representation that could not be "accompanied with the 'I think'" would be "impossible, or at least would be nothing to me" (B132). Many commentators have taken this to imply that, on Kant's view, only a self-conscious creature can have conscious representations.[1] But if this were his view, then his admission that mere brutes can have representations would commit him either to contradicting himself or else to maintaining a very strange position. For to admit that brutes can have conscious representations would be to contradict the claim that conscious representation requires the capacity for self-consciousness; but it would be strange indeed to hold that although brutes can have representations, these representations are *unconscious*. Which of these unappealing alternatives to pin on Kant is then unclear: there are passages in which he seems to speak of brutes as lacking consciousness, and others in which he clearly says that they are conscious.[2] But in any case his position looks to be a catastrophic muddle.

---

[1] Commentators who read Kant as holding that consciousness requires self-consciousness include Norman Kemp Smith, *A Commentary to Kant's Critique of Pure Reason* (1918), p. xli; Jonathan Bennett, *Kant's Analytic* (1966), pp. 104ff.; Paul Guyer, "Kant on Apperception and *a priori* Synthesis" (1980); Patricia Kitcher, "Kant's Real Self" (1984); Susan Hurley, "Kant on Spontaneity and the Myth of the Giving" (1994). I am indebted for several of these references to Steve Naragon, "Kant on Descartes and the Brutes" (1990), which contains a very helpful survey of passages from Kant and the interpretative literature surrounding them, but which itself makes a version of the assumption that I am going to criticize.

[2] For a use of the word "consciousness" (*Bewußtsein*) in a way that suggests that creatures without the capacity for self-consciousness must also lack consciousness, see §V of the Introduction to the *Logic*, where Kant defines consciousness as "a representation that another representation is in me" (Ak. 9:33). For remarks suggesting that mere brutes are conscious, as least to the extent that they "perceive" (*wahrnehmen*) and hence represent things "with consciousness," see *Logic*, Introduction, §X (Ak. 9:64-65).

I want to suggest, however, that there only appears to be a muddle here because commentators fail to see that Kant is making a claim of categorial difference. His claims about brute mentality and rational mentality seem to be in tension only because they assume that the preconditions of rational representation must be no different from the preconditions of brute representation.[3] Once we reject this assumption, we can understand Kant as claiming that conscious representations *as predicated of rational creatures* must be accompaniable with the "I think." If this is Kant's view, then it makes perfect sense for him to hold that brutes are conscious, and have objectively-contentful representations, while saying on other occasions – in contexts where he is speaking about the representations of rational creatures – that brutes cannot have conscious representations in *this* sense. For on this view, the application of the language of conscious representation to rational beings presupposes a set of mental capacities that brutes, *qua* brutes, necessarily lack.

To see that this was Kant's view, however, we must turn to the details of his discussion.

*

Kant never gives a systematic account of brute mentality, but he touches on the topic often enough, usually to emphasize the contrast between a brute mind and the kind of mind we humans possess. A characteristic discussion occurs in a much-quoted letter to Marcus Herz, written shortly before the publication of the second edition of the *Critique of Pure Reason*. There, in a passage explaining why we can only have knowledge of objects insofar as our representations "reach the unity of consciousness," Kant suggests that mere brutes should not count as "knowing beings."[4] Were I a brute, Kant remarks, I might certainly have "representations connected according to empirical laws of

---

[3] This is a version of the Uniformity Assumption that we noted in §2 of the Introduction.

[4] Letter to Marcus Herz, May 26, 1789 (Ak. 11:52). The German word that is being translated as "knowledge" here is *Erkenntnis*, which Kemp Smith usually renders as "knowledge," but which more recent translators have tended to render as "cognition." The reasons for this change, which turn on the question whether *Erkenntnisse* must always be veridical, need not concern us here. What will be important for our purposes is the distinction between *Erkenntnis* and *Kenntnis*, which is also sometimes translated as "knowledge." As we shall soon see, Kant holds that brutes *kennen* but do not *erkennen*. This suggests that he recognizes two grades of knowing, or cognizing, or whatever we choose to call it. Our

association, carrying on their play in an orderly fashion, and thus even having an influence on my feeling and desire," but all this would occur "without my knowing the slightest thing thereby, not even what my own condition was." The suggestion that brute mentality operates according to "laws of association" but does not attain to knowledge is a ubiquitous feature of Kant's remarks on animal thought. But what does this mean? What does Kant admit brutes to be capable of, and what capacities does he take them to lack?

The idea that animals are capable of a kind of association which allows them to exhibit something akin to rationality is not new to Kant: it is present, for instance, in Leibniz, who remarks in his *Monadology* that

> [m]emory provides a kind of *connectedness* to souls which resembles reason but must be distinguished from it. For we see that animals which have a perception of something that strikes them and of which they have previously had a similar perception expect, from the representation in their memory, that which has been conjoined in that previous perception, and are thus led to sensations similar to those they have had before. For example, when one shows a stick to dogs, they recall the pain that it has caused them and whine and run off. (1991, §26)

A dog can have a visual representation of its master holding a stick, can be conditioned by past experience to associate such representations with pain, and so can be led on this occasion to expect pain and to cower in fear in consequence of that expectation. Kant agrees with Leibniz that, although this sort of ability to learn from experience resembles reason, it does not actually manifest reason. He refers to this sort of brute capacity for learning as an "*analogon rationis,*" an analogue of reason.[5] The point he emphasizes about such cases is that although a brute may be conditioned to form the right expectation on seeing a brandished stick, there is an important sense in which a brute does not *know* what to expect.

Kant's reason for holding that brutes lack knowledge seems to be that they have no control over the path their representations take, and hence cannot be said to make *judgments* about what is the case. A

---

task will be to understand what distinguishes these different grades.

[5] See *Critique of Judgment*, Ak. 5:464n, *Metaphysics Li*, Ak. 28:275-276, and *Metaphysik Volckmann*, Ak. 28:449-450.

dog does not see its master brandishing a stick and judge that it is about to be given a painful beating. Rather, it just has a certain representation, and – in accordance with a blind mental habit, an "empirical law of association" – expects pain. The dog's expectation of pain would count as a judgment only if it were related to its perception of the stick as a rational consequence to a ground. But if the dog's transition from perception to expectation is simply the result of a habit, simply an instance of an empirical law of association, then even if its expectation is correct, and is caused by a perception that is in fact a sound basis of that expectation, the perception does not cause the expectation in virtue of being *taken to be a sound reason* for it.[6] Rather, the connection between these representations goes by way of a nonrational tendency to associate one representation with another.

*

There is, I think, something immediately plausible about Kant's suggestion that to judge is not merely to associate one representation with another but rather to take some sort of active ("spontaneous") role in shaping how we represent things. Our sense that judging involves spontaneity comes out, for instance, in our idiomatic talk of judging as "making up one's mind." On reflection, however, the difference that this contrast is supposed to mark can seem elusive. For is it so obvious that brutes do not "take their representations as reasons" for belief? We can concede for the sake of argument that our understanding of the behavior of creatures that believe and act for reasons takes a different form from our understanding of the behavior of inanimate things – that it is an understanding by reference to a "constitutive ideal of rationality," a conception of what a creature *ought* to believe and how it *ought* to act, rather than a conception of how things generally *tend* to happen.[7] But if, as is usually

---

[6] Compare *Critique of Practical Reason,* Ak. 5:12, where Kant claims that although animals which have repeatedly experienced something in certain circumstances in the past may *expect* similar results in the future, they do not make *inferences* about the future on the basis of past experience. An inference is a transition from one representation to another that occurs precisely because the relevant transition is taken to be justified.

[7] The phrase "constitutive ideal of rationality" is, of course, due to Donald Davidson in his "Mental Events" (1980). For the contrast between how things ought to happen and how they tend to happen, see John McDowell's gloss on Davidson's claim in his "Functionalism and Anomalous Monism":

maintained, this special form of understanding applies wherever talk of propositional attitudes has application, then surely it applies not only to human beings but to at least some kinds of brutes as well.

For surely it is plain that even dogs, to say nothing of more sophisticated animals like gorillas and chimpanzees, respond to their surroundings in a way that reflects the interaction of propositional attitudes according to rational principles. When we are not in the grip of some philosophical theory, we are perfectly willing to explain a dog's belief that its leash is in a certain closet by saying that it believes that it is there because it has seen it put there; and we are equally ready to explain the dog's pawing at the closet door by noting that it has this belief and wants to its owner to get out the leash and take it for a walk. And surely our basis for positing a connection between what the dog has seen and what it believes, and between this belief and its pawing at the door, is not just that we have inductive evidence about the causes and effects of canine beliefs in general, or of the causes and effects of the beliefs of this dog in particular. Seeing the leash being put in the closet is *prima facie* a good *ground* for believing that it is in there, and believing it is in there is an intelligible reason to paw at the door, given the desire to go for a walk (and some experience with the usual procedure of going for walks, some knowledge of how it can get its owner's attention, etc.). Moreover, our basis for ascribing the belief in question to the dog plainly depends on our recognition that this is so.

In general, our ascriptions of beliefs to animals are constrained by our convictions about what their senses give them grounds for believing, and about what actions (among those in their repertoire) would be reasonable given such beliefs and whatever wants or purposes we take them to have – much as our ascriptions of beliefs to rational beings are subject to such constraints. If Kant's claim that animal thought proceeds according to "empirical laws of association" amounts to a denial of this point, then it

---

> [T]he concepts of the propositional attitudes have their proper home in explanations of a special sort: explanations in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be. This is to be contrasted with a style of explanation in which one makes things intelligible by representing their coming into being as a particular instance of how things generally tend to happen (McDowell 1985, p. 389).

is highly implausible. But if we admit that animal thought is intelligible in terms that essentially involve reference to what would be adequate grounds for belief and action, then why should we refuse to credit animals with the capacity to "take their representations as reasons," to "make judgments," and to acquire knowledge as a result?

I mention this challenge to Kant's claim about the difference between brute mentality and rational mentality not because I think it shows his view to be mistaken, but because I believe it helps to bring out the need to appeal to the idea of a categorial difference in making sense of his position. There are moments in Kant's discussion of the difference between rational and nonrational beings when he seems to stray close to Cartesianism. In the *Grounding for the Metaphysics of Morals*, for instance, he famously remarks that "[e]verything in nature acts according to laws. Only a rational being has the power to act according to its conception [*Vorstellung*] of laws" (Ak. 4:412). This contrast between rational beings and everything else tends to suggest (although it need not imply) that all nonrational things, from stones to the most sophisticated nonrational animals, share the same kind of subjection to laws. The challenge brings out that this simple picture is not satisfactory. Whereas the kind of subjection to laws characteristic of stones is non-normative (in that explanations of stone-behavior involve no reference to how it is correct, sensible, etc. for stones to behave), ordinary explanations of animal thought and behavior do involve implicit reference to norms. Animal life is to be understood as an attempt to pursue ends characteristic of the kind of creature in question, in light of information supplied by the sensory capacities possessed by such creatures; and one of the constraints on any explanation of an animal's thought or behavior is that it be intelligible as part of an attempt to negotiate these imperatives successfully.

None of this, however, need disturb Kant's main point: that the kind of subjection to laws characteristic of rational creatures differs *both* from that of inanimate things *and* from that of animate nonrational beings, and that *this* difference turns on the fact that rational beings can consider the laws

according to which they think and act. In insisting on this point, Kant need not deny that there is a perfectly good sense in which brutes can be said to respond to reasons, to acquire knowledge, and so on. He need only claim that there is a *kind* of responsiveness to reasons of which they are not capable, and a correlative kind of knowledge which they do not attain. He need only claim, in short, that rational knowledge differs categorially from the kind of knowledge attainable by mere brutes. And there are places where Kant himself comes close to putting the point this way. In an important passage from the *Logic*, for instance, he says of animals that they "are cognizant of [*kennen*]" objects but do not "cognize [*erkennen*]" them (Ak. 9:65).[8] The kind of knowledge he is speaking of when he denies that animals are knowing beings is *Erkenntnis* – knowledge that involves the application of concepts. But it would be perfectly natural in ordinary contexts to translate *kennen* as "know."

As I see it, then, Kant's insistence on denying that brutes are knowing beings is not essential to his position. We can concede that animals are in a sense knowers, and thus avoid needless controversy, while still preserving the crux of his point: that to ascribe knowledge to a rational being is to hold it to a different kind of standard from the one we apply when we ascribe knowledge to an animal. By the same token, however, there is no tension between Kant's denial that brutes are knowing beings and the observation that brutes respond intelligently to perceptions of their environment *provided that we recognize that he is speaking of the kind of knowledge distinctive of rational creatures* (namely, *Erkenntnis*).

*

The foregoing characterization of Kant's position is only preliminary, for we do not yet have a positive account of what the distinctive kind of knowledge possessed by rational creatures *is*, or how we are to understand the associated notions of "judging," "taking as a reason," and so on. Our task in §3

---

[8] There does not seem to be any very illuminating way to mark the contrast between *kennen* and *erkennen* in English. What Kant seems to intend is a contrast between having representations that resemble and differ from one another in ways that correspond to the similarities and differences between objects, on the one hand, and having some explicit awareness of the objective similarities and differences one is representing, on the other. Thus he glosses the contrast with the remark that animals "represent something in comparison with other things, both as to *sameness* and as to

will be to develop such an account. Already, however, we are in a position to see how a creature might exhibit an analogue of reason without being able to reflect on reasons as such. For consider again the example of the dog that expects pain from a brandished stick. This expectation can be thought of as the result of an associative disposition: the disposition to enter a state of pain-expectation (with whatever consequences this has, in combination with other mental states and dispositions, for behavior) when confronted with a perception of a characteristic kind: a perception of *stick being brandished*. Now, such a disposition might arise in various ways. It might be just an instinct that the dog has from birth. It might be a habit of expectation learned from experience – might, that is, be the product of a more general tendency, after having repeatedly experienced the conjunction of events of kind A and events of kind B, to acquire the disposition to move from perception of an A to the expectation of a B. Or it might be the product of some even more sophisticated kind of learning. The details of the disposition's genesis are not important to Kant's point; what is important is that such associative dispositions plainly can be acquired, and can operate, without the animal's understanding, or even being able to consider, their justifiability.

To acknowledge this point need not prevent us from recognizing that the transitions made by an animal may in general *be* justifiable, and that this may be crucial to the animal's success in its particular form of life. Responsiveness to what, in its normal environment, are in fact good grounds for belief and action is part of an animal's being as it ought to be. There is thus no special problem about how our understanding of animal thought and action can go by way of a recognition of its justifiability: no problem over and above the general problem – if it is a problem – of understanding how there can be a way that animals of a certain species ought to be at all. What we must now consider, however, is why we should suppose that a *rational* creature's mode of responsiveness to grounds for belief and action differs categorially from this sort of brute responsiveness.

---

difference" but do not do so "with consciousness" (Ak. 9:36T). I say more about this contrast below in §4.3.

3. RATIONALITY AND THE POWER OF JUDGMENT

What underlies Kant's claim that brutes are not knowing beings, I have suggested, is the thought that brutes cannot *judge*. If this is right, then a defense of Kant's claim must begin with an account of judgment. What is it to judge, and why does the presence of this capacity make for a new kind of representation, rather than merely enabling a creature to perform certain sophisticated operations on representations of the same basic kind that brutes possess? The purpose of this section is to answer these questions. I begin (§3.1) by offering an interpretation of Kant's famous but obscure claim that only a creature with a *spontaneous* power of representation can judge. I then argue (§3.2) that only a creature capable of reflecting on a certain kind of "Why?"-question can exhibit the relevant sort of spontaneity. Finally (§3.3), I argue that the representations of a creature capable of such reflection will take a special form, the form we are referring to when we call a creature's representations "conceptual." In reaching this conclusion, we will have clarified how the presence of the capacity for judgment makes for representations of a new kind, representations categorially different from those of mere brutes.

3.1 JUDGMENT AND SPONTANEITY

Kant's name for the cognitive faculty that enables us to judge, the faculty that sets us apart from brutes, is "understanding." He famously offers various characterizations of this faculty, all of which are supposed to be equivalent. The understanding is, he tells us, the faculty of judgments, of concepts, and of rules (A68-9/B93-4, A126). These characterizations he calls "logical." But he also offers what he describes as a "metaphysical" characterization of the understanding: it is the *spontaneous* faculty of

cognition (A51/B75, A126).[9] We will be in a position to see what judging is, and why the capacity to judge categorially transforms a creature's powers of representation, when we understand the import of this metaphysical characterization, and why it should imply that the representations of a judging creature will have a distinctive logical character. Our first task, then, is to reflect on the metaphysical point.

In general, to say that a thing possesses spontaneity is to say that it is capable of determining itself, rather than having its state determined by external causes. Thus Kant calls the understanding a spontaneous faculty of representation because it "brings forth representations itself" (see A51/B75, and compare A68-9/B93-4). This is what is supposed to distinguish understanding from sensibility, which is a "receptive" faculty, one that acquires representations by being affected by mind-independent objects (A19/B33, A50/B74). But what are we to make of this contrast? In what sense does the understanding "bring forth representations itself," and how is such spontaneity connected with the capacity to judge?

Kant's reason for holding that judging is an exercise of spontaneity seems to be that to judge is to apply a concept, and a concept is a kind of representation whose nature is only intelligible by reference to spontaneous activity. Whereas sensible intuitions "rest on affections," concepts, he tells us, "rest on functions," which is to say on "the unity of the action of ordering different representations under a common one" (A68/B93). There is much in this formulation that is obscure, but at least part of the thought seems to be that the kind of order that concepts bring to our representations is one that requires activity on our part: only insofar as a subject actively orders different representations under a common representation is that subject exercising concepts at all, and what a concept is must be understood in terms of the "unity" of such an active ordering.

It is helpful to see this thought in the context of Kant's opposition to the kind of account of mental content characteristic of classical empiricism. Classical empiricists typically suppose that we simply abstract our most basic concepts from what is given to us in experience (that, as Hume puts it,

---

[9] For the distinction between logical and metaphysical characterizations of the understanding, see Kant's *Logic* (Ak. 9:36).

our "ideas" are "copied from our impressions").[10] When Kant denies that we can make sense of concepts without appeal to spontaneous activity on our part, he is denying that this sort of account is adequate. On the empiricist view, the contents of our concepts can be accounted for simply in terms of our *being affected* in characteristic ways. Thus classical empiricists assume that if encounters with things that have a certain property produce a characteristic sensory impression in us, then nothing beyond familiarity with that impression is required for us to be in a position to form a concept (an "idea") of the relevant property of things. An empiricist account of how I form my concept of the color blue, for instance, would presumably begin from the assumption that visual encounters with blue things produce a characteristic impression in me, and would suggest that I can form the idea of a general property of things, blueness, simply by concentrating on this distinctive impression while abstracting from the other features of the various particular experiences in which it occurs. What is characteristic of this kind of account, in short, is an ambition to explain the sort of cognitive power we acquire when we acquire a concept in terms of acts of abstraction performed on conscious episodes whose characterization does not presuppose possession of concepts.

Kant's view, by contrast, is that although conscious episodes of the kind contemplated by the empiricists may be *precursors* of conceptually-informed experiences, conceptual capacities are not explicable by reference to such precursors. A way of putting his objection to empiricist accounts of conceptual capacities is to say that they ignore the distinction between sensation (*Empfindung*) and cognition (*Erkenntnis*). Kant defines a sensation as "the effect of an object on our capacity for representation" (A20/B34). Sensations, in other words, are the characteristic states of consciousness that result from things affecting us. The point Kant stresses about such episodes is that, although they

---

[10] When I speak of classical empiricists, I will have Hume foremost in mind: I follow Kant in taking him to be the most consistent and far-sighted of the classical empiricists. But although I will speak of the view of "empiricists," it is not of great importance to me whether the view I discuss was actually held by Hume or any other philosopher. What matters is only that there is a recognizable kind of explanatory ambition suggested by the claim that our ideas are copied from our impressions: the ambition to take the content of sensation as primary and explain the content of thought in terms of it.

are the "matter" of cognition, they do not in themselves have any objective significance. Considered in themselves, he maintains, sensations merely "refer to the subject as modifications of its state" (A320/B376, and compare A253/B309). A "cognition," by contrast, is an "objective perception" (*loc. cit.*) – a kind of mental episode that is grasped as *representing* something as thus-and-so, not just as a shaping of the subject's own conscious state.[11] We cannot acquire concepts by abstracting them from mere sensation, for a concept is a kind of cognition – a kind of mental content whose significance is to represent things as being a certain way – and nothing that is present in mere sensation has this sort of representational significance. Sensations only become cognitions when they acquire a "form" that mere receptivity does not provide.[12]

Now, if we reflect on the matter, I think we must acknowledge that this claim about the relationship between sensation and cognition embodies an important insight. It seems clear that we can, by an act of abstraction, make sense of the idea of a way things are with us when, e.g., we see something blue. And it seems possible, as well, to imagine a creature undergoing such sensations without grasping their objective significance. Indeed, it is natural to imagine a certain phase in human development as involving this sort of relationship to sensations. Consider a child who has just learned to respond consistently to visual presentations of blue things with the cry "Blue!" No doubt she has learned to produce a certain vocalization in response to a certain sensation, but does she yet experience blueness as a property of objective things? Or to put the question another way: Does her cry yet have the significance "Something blue"? Does it, in short, give expression to an experience as of an objective

---

The Kantian claim that I shall be defending is that this explanatory project cannot succeed.

[11] As my gloss should make clear, I take Kant's term "cognition" to apply to any mental episode that presents itself as representing something – any mental episode that *purports* to inform us that things are thus-and-so. When I claim that sensations are not by themselves cognitions, I do not mean merely that they cannot by themselves supply us with *knowledge*; I mean that they do not by themselves have objective purport at all.

[12] This is not to deny that we can acquire some kinds of concepts by abstracting them from experience. The important point (expressed very obscurely) is that the experience from which concepts are abstracted must not be understood to consist of a manifold of formless sensation, but rather of sensation already infused with the forms of objective understanding. What I mean by speaking of sensation "infused with the forms of objective understanding" should

property? Or is the content of her experience rather to be conceived as like the content of our pain-experiences – as exhausted by things being a certain way *for her*, so that it does not even make sense for the kind of circumstance she experiences to exist unperceived? Should we, that is, take her cry "Blue!" not yet to have the significance "Something blue," but rather to have a significance more like that of our cry "Ouch!"?

Surely we cannot simply take for granted that a child's ability to respond to visual presentations of blue things with the cry "Blue!" reflects an experience as of *something blue*. To have an experience as of something blue is to experience the apparent instantiation of a property of objects, but merely to have learned to produce a certain cry in response to a certain pattern of sensory stimulation is not to have acquired any conception of objects or their properties.[13] That these achievements are distinct will be particularly clear if we continue to concentrate on what it is to learn to use a vocalization to *express* a concept. It is obvious that (1) there are other aspects to the correct use of the general term "blue" besides its use in response to perceptual situations. To learn how to use the term correctly is to learn how to use it, e.g., in quantified judgments about all blue things, and in requests for something blue which is absent. In these kinds of employments, and in countless others that a competent user of the term "blue" must master, the correctness of the use of the term "blue" has no direct connection with the subject's current perceptual situation. Moreover, (2) even when a general term like "blue" is being used to identify a feature of a subject's current perceptual situation, it is not enough for the subject to use the term in a way correlated with the presence of that feature. Her use of the term must be intelligible as a *predicative* use: it must express the perception that some discernible *thing* in her perceptual situation is

---

become clearer shortly.

[13] This is not an *a priori* claim about the course of child development. I do not deny that children could acquire concepts of objective properties *as soon as* they acquired the ability to produce a certain cry in response to a certain kind of perceptual situation (although as a matter of fact this is surely not how things go). What I deny is that to have acquired the latter ability is itself to have acquired the former. Even if these abilities were acquired at once, they would not be identical, for – as I shall shortly argue – to have a conception of an objective property involves more than just being able to discriminate objects with that property.

blue. And surely her employment of the expression "blue" will only express this sort of perception if it is connected in complex ways with her use of expressions intelligible as singular terms.[14] But if to possess the concept *blue* is to have a general capacity to think of things as blue – an intellectual capacity that (partially) explains one's ability to understand the various kinds of judgments to which I have just drawn attention – then it seems clear that merely becoming attuned to a certain pattern of sensory stimulation does not amount to grasping the contours of a concept.[15]

When Kant claims that we cannot simply abstract concepts from the matter of receptivity, since sensations only become cognitions when they acquire a "form" that the manifold of sensation does not itself contain, this is the point I take him to have in mind. To characterize this form in more detail, we would need to investigate the kinds of structure that are involved in the representation of objects – the representation of things that can hold together a set of diverse properties in a kind of unity, that cannot possess two incompatible properties at the same time, that can endure through time and undergo changes, and so on. I take the discipline Kant calls "transcendental logic" to have precisely this project: to describe the general features of objective thought by identifying the forms of judgment that define the very idea of an object. To consider Kant's more detailed claims about the specific forms under which objective thought must fall would require more space than I have here. Nevertheless, I hope that the very abstract points I have made about the interdependence between grasping concepts and having a conception of objects already begin to bring out why Kant should hold that sensation can only be the

---

[14] To see this, consider what would be required for a child's sequential use of the expressions "blue" and "doggie" in response to a single perceptual situation to count as expressing the thought that a certain toy dog was blue, rather than just expressing a recognition of two unrelated features of her perceptual situation. It seems clear that the expression "blue" is not functioning as a genuine predicate until there is a distinction between its use in conjunction with expressions like "doggie" to pick out two unrelated features of the subject's perceptual situation (the presence of something blue and of a doggie) and its use to pick out something with a certain feature (the presence of a blue doggie). On this point, compare W. V. Quine, *Word and Object* (1960), §§19-21.

[15] The argument of this paragraph is a condensed presentation of points made by Peter Geach in criticism of a doctrine he calls "abstractionism" in §10 of his *Mental Acts* (1957). I should add that, if the points I have been stressing are obvious for terms like "blue," they are – as Geach brings out in adjacent sections – all the more obvious for the many other kinds of expressions (terms for primary qualities, count nouns, mathematical terms, logical expressions, etc.) the

matter of cognition, and must take on a form that is not given in mere receptivity if it is to supply us with objectively-significant perceptions.

These points should also begin to bring out why conceptual representation presupposes a kind of spontaneity. As we have seen, a subject will count as possessing concepts only to the extent that she exhibits an appreciation that what makes a given judgment (or any other episode involving the application of a concept) correct is not merely what is presently going on with her.[16] It follows that our use of concepts involves a kind of spontaneity *relative to the contents of sensory receptivity*: to the extent that we can make sense of the idea of a raw sensory episode – something that merely happens to us when we encounter things with a certain property, described in terms that do not presuppose the whole suite of conceptual capacities drawn on in our mature thought about objects and their properties – such an episode will not have the kind of content from which we could derive a concept. To form a concept we must bring to bear a framework that is not given in mere receptivity. This framework, which is presupposed in all concept-use, is something the understanding "brings forth itself."

3.2 SPONTANEITY AND REFLECTION

Our way of bringing out the spontaneous contribution of the understanding to conceptual representation has been to ask what is required for a child's cry to come to express an empirical judgment, a classification of something sensibly given as falling under a concept. Our conclusion was that it is necessary for her cry to come to express the perception of *an object* as having a certain property, which in turn requires that she should have mastered whatever conceptual tools are needed to represent objects whose states are, in general, independent of what is presently going on with her. In order to

---

principles of whose connection with perceptual circumstances are vastly more complex than those of color words.

[16] Except in the special case where her judgment concerns her own present state of consciousness – and even in this case there is clearly a distinction between undergoing an associative episode and judging that one is undergoing such an episode.

characterize the cognitive powers that are distinctive of a rational being, however, we must be more specific about the sense of representing objects that is at issue here. For surely there is a sense in which even brutes can be said to represent objects and their properties. After all, we ordinarily have no qualms about ascribing representations of objects to at least some animals that lack self-awareness. Why did the dog dig just here? An intelligible answer would be: "Because he remembered that this is where he buried his bone." And surely this explanation presupposes that the dog has a representation of a particular bone as having been buried in a certain place. But if we concede that brutes can have such representations of particular objects and their properties, then we need to say more to characterize the kind of power of objective representation that is distinctive of a rational being.

Our next question must therefore be: what kind of appreciation of objects *qua* objects is distinctive of rational creatures, creatures that can judge? And once again, we can make this question more concrete by asking when a child's cry becomes intelligible as *expressive* of such an appreciation. For presumably we might train a parrot to cry "Blue!" when something blue was set in front of it, no matter what other features the object possessed, in all sorts of lighting conditions, and so on. The parrot would then meet one plausible standard of objective representation, the standard that is typically applied in studies of animal perception: it would be capable of responding discriminatingly to a property of objects, rather than to more proximal similarities in its patterns of sensory stimulation.[17] It is clear, however, that its cry might meet this standard while still not expressing a genuine *judgment*, a conceptual classification of something as blue, for the parrot might still have no understanding of the significance of its own response. A creature that only exhibited this sort of uncomprehending responsiveness to properties of

---

[17] Animals that meet this standard are often said to possess a *concept* of the relevant property. One of my aims in this section, however, is to make it plausible that, in the special sense of "concept" that applies to rational creatures, mere brutes do not possess concepts. This is not to deny that we can *define* standards for ascribing concepts to brutes: we can of course use the *word* "concept" however we please. I would only urge that we not overlook the fundamental difference between the standards typically applied in assessing brute concepts and the standards that govern the rational case. For a critical review of literature on "animal concepts," see Nick Chater and Cecilia Heyes, "Animal Concepts: Content and Discontent" (1994).

objects would plainly not count as grasping objective properties in the sense that we are trying to characterize. What, then, do we require in a comprehending child that would be missing in such a parrot?

When the question is put in these terms, I think the general lines of the answer are clear: to count as expressing a judgment, a child's utterance must not just manifest an isolated disposition to make certain sounds in certain conditions (e.g., when she is exposed to certain sensory stimuli). For a child's cry expresses a judgment only if she *understands* what she is saying, and understanding dawns only when the child comes to recognize truth-relationships between this sentence and other sentences that she, or another person, could utter. She must, as we can put it, come to appreciate what she is doing when she utters the relevant sentence as *taking a stand on what is so*, a stand that bears relations of implication and exclusion to stands on other questions, a stand that other speakers could query or contradict. And this, in turn, requires that she understand the relevance of a certain kind of "Why?"-question to what she says. In general, a competent speaker who claims that *p* can be asked "Why do you think that *p*?"; and it will be a criterion of her understanding this question, and thus of her understanding what she has said, that she be able to answer by producing grounds of a certain kind: grounds that bear on the truth of the claim that *p*. She must, as it is sometimes put, be able to "play the game of giving and asking for reasons," where "reasons" here means considerations bearing on the truth of the claims she has made.[18]

The capacity to express one's representations in this comprehending way is, I want to suggest, the distinguishing mark of a rational creature. A mere brute can of course perceive that an object has a certain property, can respond to such perceptions in ways that are appropriate given its desires and aims, can retain traces of past perceptions in memory, and can even in some cases be trained to "express" its perceptions in the manner of our imagined parrot. But just as mere brutes do not express their

---

[18] I borrow the phrase from Robert Brandom's *Making It Explicit* (1994). Because I think the capacity to express conceptually-contentful assertions requires the capacity for self-consciousness, I do not accept Brandom's claim that the

representations *in comprehending assertions*, so likewise they do not form representations in the way that deserves to be called *judging*. For judging that *p*, like saying that *p*, is a special kind of *act*, which stands in a different kind of relationship to grounds from any act in the repertoire of a mere brute. A creature that can judge is one that can entertain the sort of "Why?"-question just mentioned – a "Why?" that demands grounds for thinking a proposition true – and the act of judging is the (particular, dateable) act of forming a belief for truth-related reasons grasped as such. The ability to make comprehending assertions is the mark of this capacity because a creature counts as able to make assertions just if it is able explicitly to answer this sort of "Why?"-question about its own utterances.

It should be clear, moreover, that there is a fundamental difference between a creature that merely acquires beliefs from perception, retains them in memory, and perhaps possesses various uncritical habits of "generalization" and "inference," on the one hand, and a creature that can ask itself why it believes what it does, on the other. The capacity to reflect self-consciously on one's beliefs might seem like an inessential addition to one's cognitive capacities if it were merely a capacity to *monitor* what one believes – merely a capacity to know what one believes, which left one's mechanisms of belief-formation and belief-alteration unchanged. But the capacity to judge is more than this: it is the capacity to reflect critically on the *grounds* for one's beliefs, which leaves no mere tendency to form or alter beliefs untouched. A creature that can ask itself why it believes what it does stands in a free and distanced relation to all of its beliefs: it not only holds them, but can reflect on their soundness, and alter them in light of this reflection.[19] Of course, even a rational creature will actually subject only a fraction of its beliefs to such scrutiny, but all of its beliefs are *potentially* subject to it, and this implies that no mere

---

normativity characteristic of conceptual thought can be reduced to more primitive forms of implicit (i.e., non-self-conscious) normativity. Nevertheless, the foregoing paragraph should indicate that my thinking owes a debt to his work.

[19] The "and" in this formulation should not be taken to suggest that the ability to reflect on the soundness of one's beliefs and the ability to alter one's beliefs in light of such reflection are wholly distinct. A creature will only count as able to reflect on the soundness of its beliefs if, in the normal case, what it believes just is what it takes itself to have sound reasons for believing. The capacity to alter one's beliefs in light of one's reflection on grounds is thus an inextricable part of the capacity to reflect on grounds at all.

tendency to form or alter representations is sacrosanct for it.

A recognition of this potential for critical reflection informs the whole apparatus we use in characterizing, explaining and assessing a rational creature's thought and action. Thus, in the fundamental kind of case, when we explain a rational creature's believing something or doing something by reference to something else it believes, perceives, or remembers, the creature's rational powers are also implicitly involved in the explanation: to say that a rational creature is (e.g.) doing *A* because it believes that *p* is to say that it does this because it *takes it to be the case* that *p* and *takes this to be a reason* for doing *A*. This is so even if the action in question is not the outcome of a process of conscious deliberation. For even if the creature does not *actually* reflect on its grounds for doing *A*, to say that it is doing *A* because it believes that *p* is to presuppose that it is *capable* of reflecting on grounds for action as such, and that if the creature *were* asked why it is doing *A*, it would be able to produce the relevant ground, and would assent to it. If these conditions were not met, we would withdraw the explanation.[20] We can thus say that, even where there is no conscious deliberation, a rational creature's normal transitions from belief to action, or from one belief to another, proceed with that creature's tacit consent. In the case of mere brutes, by contrast, there is simply no question of the creature's consenting to the transitions that it makes in thought and action.

3.3 REFLECTION AND CONCEPTS

Someone who accepted these points might still wonder why this admittedly significant transformation in a creature's power to *reflect* on its representations should count as transforming the nature of those

---

[20] I do not mean to deny that there are intelligible kinds of explanation which advert to beliefs but which are not subject to these conditions. Perhaps psychoanalysis can sometimes produce true explanations of behavior which advert to beliefs of which the subject is not conscious, and which he would not endorse even if he were made conscious of them. Be that as it may, it seems clear that there is a *kind* of explanation of the behavior of a rational creature whose application requires that the creature know why he is doing what he is doing: the mark of this is that there is a use of the claim "*S* is doing *A* because she believes that *p*" on which the claim would be withdrawn if it were discovered that *S* did

representations themselves. Our project is to understand why ascriptions of representational states such as believing that …, perceiving that …, and so on, have a categorially different significance in application to rational creatures from the one they have in application to mere brutes. But why should this follow from the fact that only rational creatures can reflect on their representations in the sense just specified?

Kant's answer to this question is encapsulated in his oft-repeated claim that only a rational creature is capable of bringing its representations to concepts.[21] To express the point in contemporary terms: only rational creatures are capable of *conceptual* representation. I now want to argue that to represent conceptually is to represent in a quite distinctive way, a way that has no counterpart in the cognitive life of a mere brute. If this is right, then the ascription of representational contents to a rational, concept-using creature will have a wholly different significance from the ascription of a similar content to a nonrational creature. And this, of course, is what we are seeking to establish.

We can begin to see what distinguishes conceptual from nonconceptual representation by noting the close connection between concepts and inference. In general, to represent a subject as falling under a predicate is to represent it as having a characteristic that other subjects might also possess. A predicate is thus an essentially *general* kind of representation, one that connects any particular to which it is applied with other actual or possible particulars that also fall under this predicate.[22] In virtue of this generality, a representation of a subject as falling under a predicate will stand in relationships of deductive entailment and inductive support to other actual or possible representations. Now, to say that a creature's representation of a subject as falling under a certain predicate is *conceptual* is to say that it represents that state of affairs in a way that involves *grasp* of the place of the content in question in such a system of

---

not regard *p* as any reason to do *A*.

[21] See for instance *Anthropology from a Pragmatic Point of View*, §40 (Ak. 7: 196-197), and compare *Metaphysik Li* (Ak. 28: 286), *Metaphysik Mrongovius* (Ak. 29: 888-889).

[22] Kant lays great stress on the fact that concepts are general representations, representations of characteristics that are at least potentially common to many particular representations (see e.g. A68-9/B93-4, A103, B133-4n, and *Logic* §§1-2 [Ak. 9:91]). The next few paragraphs spell out what I take to be at stake in this point. My understanding of what it means to characterize a representation as conceptual owes a great deal to the work of Gareth Evans. For Evans' views on

inferentially-related contents. This grasp will manifest itself in an ability to recognize inferences that such relationships warrant: a creature that conceptually represents, e.g., a certain knife as sharp will be in a position to recognize its usefulness in tasks where something sharp is required, to take account of experiences with it in drawing inferences about the potentialities of sharp things, etc. To just the extent that animals cannot be said genuinely to infer, therefore, they cannot be said to represent conceptually either, for to represent conceptually just is to represent in a way that reflects some (perhaps partial) grasp of the location of one's representation in an inferentially-articulated network of other representations.

This is not to deny that we can legitimately ascribe representations of subjects-as-falling-under-predicates to mere brutes. Nor is it to deny that such ascriptions can figure in genuine (not merely "as if") explanations of brute behavior. It is only to insist that ascriptions of representations with subject-predicate structure have a different significance, and can figure in a different *kind* of explanation of thought and action, when the subject is a rational creature. When we ascribe a (personal-level) propositional representation to a rational creature, we characterize its representation as the product of the joint exercise of several distinct conceptual capacities, each of which is capable of being exercised in quite different contexts, and in quite different logical connections, from the present one. If, for instance, we ascribe to a rational creature the thought *This table is red*, we commit ourselves to supposing that it possesses general capacities to think of *anything* of the appropriate kind as red, or as a table, and to attend to all kinds of implications of such classifications, as dictated by the entailment- and exclusion-relations of the relevant concepts and the logical structure of the propositions in which they figure. When we ascribe such a representation to an animal – so the Kantian thought would go – we undertake no such general commitment.

This contrast reflects a fundamental difference between the role we expect representations to play in the lives of rational creatures and the role we expect them to play in the lives of mere brutes. We

---

concepts, see his "Semantic Theory and Tacit Knowledge" (1981), §III and *The Varieties of Reference* (1982), Chapter 4, §3.

expect a rational creature to be able to take account of its representations in all kinds of projects, in a way mediated by indefinitely many other propositional attitudes. There is no particular behavior we expect from a rational creature who represents a certain table as red: whether the creature covets the table, or attempts to destroy it, or looks on top of it for the key to a certain apartment, will depend on what it wants and what else it believes – not to mention a whole range of other kinds of attitudes. By comparison, the dispositions associated with a given representation in a mere animal will be more or less inflexible. "More or less" here must obviously admit a very wide range of possibilities, corresponding to the enormous differences between the behavioral flexibility of, say, a bumblebee and a baboon. Nevertheless, if it is true that animals cannot subject the fundamental tendencies that govern their thinking to critical reflection, then there will be a sense in which *all* animals, however sophisticated, have inflexible dispositions: their most basic cognitive and conative dispositions will not be subject to alteration by reasoning, but only by processes like conditioning, or being conked on the head.

A nonrational creature may have objectively-significant representations in the following sense: it may respond to its environment in a way that reflects an ability to discriminate (certain kinds of) objects and their properties. Very roughly, a nonrational creature's belief that *p* can be thought of as a kind of disposition: namely, the disposition to take the means to its ends that would be rational given that *p*. Thus we typically count a nonrational creature as capable of recognizing a certain property of objects just if it can respond to the perceptual presence of objects with that property with a distinctive kind of behavior, one that is intelligible given its aims. But to meet this minimal condition of objective perceptual discrimination – to exhibit at least *one* distinctive kind of response to things with a certain property in *one* type of circumstance – is not by itself to exhibit grasp of the *concept* of that property. To represent conceptually is precisely to grasp information in a form that makes it available to whatever interests or purposes one may have. A creature shows a conceptual understanding of the fact that an object has a certain property only by showing that it has grasped the significance of its discrimination,

not only for a given purpose in a particular context, but in a form that makes the information available for *whatever* purposes the creature may have or acquire.

The mark of this sort of understanding is precisely the ability to articulate the relevant fact in speech, for to do this is to signify one's recognition of the relevant fact as a truth in its own right, independent of any particular context or purpose. For, except in special cases, to *say* that *p* is not to *act in a way made practically rational by* the belief that *p*.[23] Rather, saying that *p expresses* the belief that *p* in a way that no other kind of action does: whereas nonlinguistic actions may reflect the belief that *p* by being *manifestations* of the relevant disposition to act (in something like the way that catching fire in certain conditions is a manifestation of flammability), to say that *p* is to perform an act whose significance is to represent the speaker *as* having the relevant disposition. The ability to express beliefs in this context- and purpose-independent way is the mark of a rational understanding because a creature counts as rational only if the contents of its representations are available to it in this distinctively abstract and general form.

I take this to be the point that Hegel has in mind when he says, in the remark quoted at the head of this chapter, that although an animal's awareness is "*in itself* universal," only man is "a universal for a universal." Mere animals can respond discriminatingly to what the philosophical tradition has called "universals" – that is, to properties which can be instantiated by any number of objects. But this capacity to *respond discriminatingly* to what are in fact universals is not itself a capacity to *recognize* them in their universality – that is, to appreciate their significance in abstraction from any particular context or purpose. We can express this point by saying that, although the representations of a nonrational animal

[23] This is not to deny that a subject's saying that *p* on a given occasion can, like any action, be understood as the result of the joint efficacy of a belief and a desire. But the belief that motivates *saying* that *p* obviously need not be the belief that *p*: it may serve my purpose to represent myself as believing that *p* even though I do not in fact believe it. What we can say about the connection between saying that *p* and believing that *p* is this: to assert that *p* is necessarily to *represent oneself as* believing that *p*.

may be objectively valid *in themselves*, they are not objectively valid *for their subject*.[24] Only a creature whose representing achieves the kind of spontaneity characteristic of a rational intellect is in a position to recognize universals as universals – i.e., to exercise concepts. Hegel follows Kant in claiming that man's capacity to "know that he is 'I'" is the fundamental condition of such recognition. In another, more lyrical passage, he explicitly credits this thought to Kant:

> The 'I' is as it were the crucible and the fire which consumes the undifferentiated manifold of sense and reduces it to unity. This is what Kant calls *pure apperception*. (Hegel 1975, §42, *Zusatz* 1)

4. CONCLUSION: OPEN QUESTIONS

The general aim of this chapter has been to consider the kinds of powers that are distinctive of a rational understanding, with a view to asking why only a self-conscious creature can exhibit these powers. We have seen that a rational understanding is characterized by a distinctive kind of spontaneity in its representing, which makes possible a distinctive kind of generality in its thought and a corresponding kind of flexibility in a rational creature's responses to circumstances. Kant's claim, of course, is that our capacity for apperception is what makes it possible for us to think in a way that has these characteristics. We can now consider why this is so.

I have emphasized that a rational creature must be able to subject its beliefs to the kind of scrutiny implied by the question "Why do I think that?" One way of bringing out the work that remains to be done is to ask what the word "I" is doing here. In a sense, of course, the answer is clear: the word marks the subject's recognition that it is his own belief under consideration. But in the first place, it is

---

[24] Although I have expressed this point using Hegel's terms, I think it really describes Kant's considered view of the matter as well. We can see this in the passage from the *Logic* in which he tells us that animals represent things "in comparison with others as to identity and difference" but that they do not do so "with consciousness" (L 9:65T). This is presumably to say that animals can differentiate between things on the basis of their properties, but cannot represent the

not yet clear what "his own" means here. To say that I know that the belief that *p* is my own seems to be just a convoluted way of saying that I know that *I* believe that *p* – but what knowledge is this? And in any case, it is not clear why the question need contain the expression "I think" at all. After all, as I have repeatedly emphasized, the relevant "Why?"-question is really a question about what grounds there are for taking the proposition under consideration to be true. It may be natural to express this question in English using the words "Why do I think that *p*?", but the mere naturalness of this form of words does not by itself show that each of the contained expressions corresponds to a distinct conceptual capacity that someone asking the question must exercise. Perhaps the presence of the expression "I think" here is merely a conventional flourish, an artifact of English usage with no deep significance. Perhaps not, of course, but we do not yet have any clear account of what its significance is.

To address these concerns, we will need some account of the sense of the expression "I", so that we can gauge what knowledge a person who accompanies a proposition with "I think" is expressing; and we will need an account of why the occurrence of this expression in the question "Why do I think that *p*?" is significant. The task of the next chapter will be to answer these questions, and thus to clarify the connection between rationality and self-consciousness.

---

points of identity and difference *to themselves* in the manner of a concept-user.

IV. REASON, OBJECTIVITY, AND SELF-CONSCIOUSNESS

> [T]he proposition *I think* (taken problematically) contains the form of every judgment of understanding whatsoever.
>
> Kant, *Critique of Pure Reason*, A348/B406

1. INTRODUCTION

In the last chapter, we examined the powers that are distinctive of a rational creature – in particular, the power of conceptual thought, whose paradigmatic exercise occurs in judging. I argued that these powers differ in kind from the cognitive abilities of a nonrational creature, and I tried to bring out the character of this difference by describing the distinctive kind of spontaneity that a rational intellect must possess and the distinctively abstract and general sort of relationship to facts that such spontaneity makes possible. Having made these observations about rationality, we are finally in a position to ask why a rational creature must be self-conscious. Or, given that our characterization of rationality has centered on the power of conceptual representation, we can also express this question by asking: Why is the representation *I* an essential element in anything intelligible as a repertoire of concepts? For this, as we have seen, is Kant's view. He claims that the representation *I* is "the vehicle of all concepts whatsoever" (A341/B399), "a mere consciousness that accompanies every concept" (A346/B404). Our question in this chapter will be: What kind of consciousness is this, and why is it necessary for conceptual thought?[1]

We have already seen one indication of the connection between the capacity for conceptual

[1] On Kant's usage, all thought is conceptual thought, for to think (*denken*) just is to bring representations under concepts (see A50/B74, A69/B90). In this chapter, however, I will sometimes include the redundant adjective to emphasize the contrast with another mode of representation that might be called "thought" in a more permissive sense (*sc.*, "animal thought").

thought and the capacity for self-consciousness: namely, that the basic question a concept-user must be able to put to itself – "Why do I think that *p*?" – involves the first-person pronoun. As I emphasized at the end of the last chapter, however, merely to note this connection is not to explain it. We still need to ask: What is the word "I" doing here? Why is its presence not merely a grammatical accident, in something like the way that the "It" in "It is raining" is a grammatical accident? English grammar requires that a complete sentence have a subject, so we report rain by ascribing it to an unspecified "It," but no one would argue on this basis that a creature capable of thinking thoughts about rain must have a special concept *It*. For the word "It" in "It is raining" is not a functioning part, as becomes clear if we try to answer the question "*What* is raining?" It seems clear that a creature can understand this kind of sentence without drawing on any conception of the subject of which raining is being predicated. Similarly, it is natural to wonder whether a creature might not learn to play the game of giving reasons for assertions, the game whose commencement is signaled in English by the question "Why do I think that *p*?", without drawing on any conception of the subject designated by "I."[2] For, as I have emphasized, this question is really a request for considerations that speak in favor of *p*'s being *true*. But then what, apart from mere grammatical convention, requires us to mention the thinking subject in framing the question?

As I have posed it, this is a question about the significance of a certain kind of linguistic expression, but I hope it is clear that this topic remains closely related to questions on which Kant explicitly takes a position, questions that concern not language but thought. Before proceeding, it will be useful to recall the path that has led us from Kant's texts to this point. Recall that the Kantian Thesis is:

(KT) Any creature that possesses the faculty of understanding must be able to accompany all of its representations with "I think."

---

[2] One author who explicitly claims that this *is* possible is Robert Brandom, who holds that

> there need be nothing incoherent in descriptions of communities of judging and perceiving agents, attributing and undertaking propositionally contentful commitments, giving and asking for reasons, who do not yet have available the expressive resources 'I' provides (1994, p. 559).

I think many philosophers of mind and language hold views that commit them to agreeing with Brandom about this. See §2 of the Introduction to the dissertation for some cases in point.

If we could explain why the occurrence of "I think" in "Why do I think that *p*?" is not merely a grammatical accident, we would have an argument for (KT). For we have already concluded that

(1) to count as possessing the faculty of understanding, a creature must understand the question "Why do I think that *p*?"; and that

(2) a representation that a subject could not produce in response to this question could not figure among her grounds for belief.

Point (1) is a lesson of Chapter III: it follows from our conclusion that the defining activity of the understanding, namely judging, can only occur in a creature that understands this sort of "Why?"-question. Point (2) is a lesson of Chapter II: it follows from our conclusion that a representation that a subject could not produce in response to the question "Why do I think that *p*?" would be "nothing to her" in the sense that it could not guide her thinking about what is the case. Now, if we can show that the occurrence of "I think" in "Why do I think that *p*?" is not just a grammatical accident, then it will follow from (1) that a creature that possesses the faculty of understanding must possess the concept *I*. And given (2), it will also follow that a creature that possesses the faculty of understanding must be able to bring *all* of its representations within the scope of apperceptive reflection, on pain of those representations being "nothing to it." This would complete our argument for (KT).

The aim of this chapter is to supply the missing link in this argument by showing that the occurrence of "I think" in "Why do I think that *p*?" is no accident. My approach will be, not to dispute that the question "Why do I think that *p*?" is in the first instance a request for considerations that speak in favor of *p*'s being true, but to argue that the reference to the thinking subject in this question is not merely something additional to and unconnected with its reference to a proposition capable of truth or falsity. Rather, the "I think" simply makes explicit a contrast that is essential to any judgment on a proposition. Kant's way of putting this is to say that the *I think* "contains the form of all judgments of understanding" (A348/B406).[3] His point is that accompanying a judgment with "I think" does not add a

[3] He says this of the proposition *I think* "taken problematically." A problematic judgment is one in which a certain combination of representations is not actually asserted but acknowledged as possible. The point of the phrase here, I

further element to the *content* of the judgment; it simply makes explicit that in virtue of which the relevant representational matter constitutes a single, unified judgment.[4] A creature capable of judgment must be capable of accompanying its representations with "I think" because, as we shall see, it must be capable of understanding this form if it is to be capable of judging at all.

My strategy in arguing for these conclusions will be to begin by taking a step away from Kant: in §§2-3, I consider some more recent attempts to connect the capacity for rational thought with self-consciousness. These recent discussions have the advantage that they are plainer and more readily intelligible than Kant's own. Seeing their strengths will help us to grasp what Kant was driving at; seeing their weaknesses will help us to understand why he put things as he did. This will then equip us to turn, in §4, to Kant's own account of the connection between conceptual representation and apperception.

2. THE RATIONALITY ROUTE

When we considered the intuitive motivation for (KT) in Chapter I, we were not yet in a position to focus on the significance of the "I think" in the question "Why do I think that *p*?" Instead we asked more vaguely: Why does rationality require the capacity for apperception? Our preliminary answer turned on the thought that what distinguishes a conscious reason from a blind impulse to believe is that a conscious reason is the sort of thing on which the subject can reflect, and that the ability to accompany one's representations with "I think" is precisely what transforms those representations from sources of

---

take it, is just to abbreviate the claim that is stated more fully at B131-2: that although not all of my representation must actually be accompanied with "I think," it must be possible for the "I think" to accompany any of my representations. So what Kant is saying, rather awkwardly, at A348/B406 is that the most general form of any judgment is determined by the requirement that it must consist of representations capable of being accompanied with the "I think." *Actually* to accompany a judgment with "I think" is to make this formal determination explicit.

[4] Compare how "It is true that" functions in "It is true that *p*." Prefixing "It is true that" to "*p*" does not add to the content that is asserted by simply saying "*p*"; rather, as Frege famously remarks, it "seems to make the impossible possible," allowing what is involved in a judgment's being asserted to enter into the content judged without altering that content (Frege 1997, p. 323). Another way to put this is to say that accompanying a judgment with "It is true that" does not alter the content of the judgment, but makes explicit something about the form in virtue of which it is a judgment.

blind impulses into considerations open to reflection. We can call this the *rationality route* to the conclusion that the "I think" must be able to accompany all of my representations, for it depicts the capacity for self-conscious awareness of one's representations as a precondition of rational reflection on the significance of those representations. A number of recent authors have argued along these lines for the importance of self-consciousness. My aim in this section is to consider the problems and prospects of such arguments. Although I think the rationality route has considerable intuitive appeal, I shall suggest that existing accounts of this appeal are unsatisfactory. It will emerge that a satisfactory "rationality route" defense of the Kantian thesis must meet two conditions, the joint satisfaction of which is no simple matter: it must explain why the capacity for self-consciousness is essential to rationality without defining "rationality" in a way that makes the connection trivial.

The basic idea of the rationality route is forcefully expressed in a passage of Christine Korsgaard's *The Sources of Normativity*. According to Korsgaard,

> the human mind *is* self-conscious in the sense that it is essentially reflective. I'm not talking about being *thoughtful*, which of course is an individual property, but about the structure of our minds that makes thoughtfulness possible. A lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will… But we human animals turn our attention on to our perceptions and desires themselves, on to our own mental activities, and we are conscious *of* them. That is why we can think *about* them… For our capacity to turn our attention on to our own mental activities is also a capacity to distance ourselves from them, to call them into question. I perceive, and I find myself with a powerful impulse to believe. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I believe? Is this perception really a *reason* to believe? (1996, pp. 92-93)

Korsgaard's view is thus that the capacity to reflect self-consciously on our own perceptions and desires necessarily involves an ability to "distance ourselves" from such "impulses," and that this ability to "back up" makes us capable of entertaining questions about reasons, and at the same time forces such questions upon us.

When we considered this line of thought in Chapter I, our main complaint about it was that it is

more metaphor than argument. The intuition underlying such claims about the importance of self-consciousness is that a rational creature must be able to "back up" from its representations, and that the capacity to take the relevant sort of step back requires self-consciousness. But although the image of backing up is evocative, its content is far from clear. Why should rationality require the capacity for this apparently quite sophisticated kind of reflection? No doubt a rational creature must in some sense be able to reflect on *its representations* – that is, on their contents. But the capacity to accompany one's representations with "I think" seems to involve more than this: it seems to involve the capacity to reflect, not just on the content of one's representations, but on *the fact that one has* these representations. Why should this count as a precondition of rationality, rather than just a special and sophisticated application of it?[5]

Korsgaard herself says little about this question – arguing for the connection between rationality and self-consciousness is not her central concern – but a number of other philosophers have recently tried to answer it. It will be useful to consider a few of these answers. A good place to begin is with a frequently-cited passage from Colin McGinn's *The Character of Mind*. According to McGinn, a rational being must be aware of its own beliefs for the following reason:

> If a person were not aware of his beliefs, then he could not be aware of their inconsistency; but awareness of inconsistency is (primarily) what allows normative considerations to get purchase on beliefs; so the rational adjustment of beliefs one to another seems to involve self-consciousness, that is, knowledge of what you believe. Without such self-consciousness the control of logic over thought would be deprived of its compelling force; rationality as we know it requires knowledge of the contents of one's own mind (1997, pp. 21-2).

Similarly, Sydney Shoemaker writes that

---

[5] In view of our discussion in Chapter II, we should also feel another, vaguer misgiving about Korsgaard's description: namely, that there is something odd about the idea that the representations of a rational creature confront it as "impulses," even impulses from which it can "back up." For if I see what looks like a bent stick standing in a glass of water, but do not form the belief that the stick is bent, have I resisted an impulse? This may be an evocative metaphor, but taken literally it surely misdescribes the situation: if I know that the stick's bent appearance is illusory, then although the appearance remains, it does not have *any* power to lure or press me into belief. Talk of impulses suggests that perceptions and desires are forces that impinge on our reason, so to speak, "from without," pushing it this way and that. But in fact, as we have seen, reason is not merely the power to *resist* nonrational forces in the mind: insofar as a subject is rational, no mere impulse has any power – even a resistible one – over its thought.

> [a]n essential part of rationality, for creatures with the conceptual capacities of human beings, is the appropriate adjustment of beliefs and desires in the light of new information about the world, and, as a necessary means of such adjustment, the conducting of appropriate tests and reasoning. For someone to know what sorts of tests and reasoning are called for it is essential that he know what his current beliefs are—only so can he know which of them are called into question by new information, and so what questions about the world his tests and reasoning should be focused on (1991, p. 207).

The common theme of these two passages is that our capacity to "adjust" our beliefs and other attitudes in light of new information depends on our knowing what beliefs we hold, and hence depends on self-consciousness. Logic demands consistency in our beliefs, but only if we know what we believe can we heed these demands. Norms of sound epistemic practice demand that we "conduct appropriate tests and reasoning" to check whether our beliefs are sound, but only if we know what our current beliefs are can we see what tests will serve. Rationality requires self-consciousness because only a self-creature creature can comply with these requirements.

Although this line of thought has an initial attraction, I think it does not get to the heart of the matter. What McGinn and Shoemaker's remarks have in common is the suggestion that knowledge of one's own representational states makes possible the consideration of whether they accord with rational norms. Self-consciousness, then, is necessary to rationality as something like a means to an end: it is what allows us to know what representations we have, which is necessary if we are to reflect on their consistency, well-foundedness, and so on. But if self-consciousness is merely a means to the implementation of rational norms, then it makes sense to ask why conformity to the relevant norms needs to be achieved by just this means. Why couldn't such conformity be achieved by the operation of unreflective mechanisms? And in any case, how does self-consciousness facilitate the end? If it is one thing to know *what* one's current beliefs are and quite another to consider whether they are *reasonable* – if we might discover that something was, in fact, "our current belief" without yet considering the question of its soundness – .then it is hard to see why the capacity for self-consciousness by itself should give us any more control over our representations than is possessed by creatures that lack self-consciousness. For suppose I know that I hold two beliefs which are inconsistent, and determine to perform certain

"tests and reasoning" to determine which should be retained. If these tests and reasoning are conducted at the level of self-conscious reflection, and if one's first-order belief is a state distinct from one's second-order belief about what one believes and whether it is sound, then we still need an account of the means by which the tests and reasoning effect a change in the first-order belief. In fact, of course, to suppose that, having followed a course of self-conscious reasoning, we might in general still face a question of how to make ourselves believe the conclusion of that reasoning, would make the very idea of self-conscious reasoning unintelligible. But the deficiency of the McGinn-Shoemaker account of the importance of self-consciousness is precisely that it supplies no account of this unintelligibility: given their conception of what self-consciousness does for us, it is hard to see why there should not be an intelligible question here.[6]

What we need is an account of what such things as "tests and reasoning" are, and how they depend on self-consciousness. Merely to say that self-consciousness opens our representations to such processes, making possible "the control of logic over thought," is to presuppose the very connection of which we are seeking an account: the connection between the capacity to know one's own mind and the capacity to reason. Our question should be: what sort of control is this, and why would it be impossible in a mind not equipped with the capacity for self-consciousness? For we want to show, not merely that self-consciousness *facilitates* rational thinking, but that it is an *essential precondition* of rational thinking. This, then, is one condition that a "rationality route" defense of the Kantian thesis must meet: it must explain why self-consciousness is, not merely an *aid* to rational thought, but a capacity *essential* to rationality. Neither McGinn nor Shoemaker really clarifies why this should be.

By the same token, however, a defense of the Kantian thesis must not simply stipulate an understanding of "rationality," or some capacity crucial to rationality, in a way that makes its connection with self-consciousness trivial. We can see how this might be a problem by reflecting on Tyler Burge's

---

[6] McGinn and Shoemaker are criticized along similar lines in Moran 2001, Chapter 4, a discussion to which I am indebted.

recent (and in other respects illuminating) work on the connection between rationality and self-consciousness. Burge is interested in the link between knowing one's own mind and being able to think rationally. To clarify the particular sort of rationality that he will be concerned with, he introduces the notion of "critical reasoning," which he explains as follows:

> Critical reasoning is reasoning that involves an ability to recognize and effectively employ reasonable criticism or support for reasons and reasoning. It is reasoning guided by an appreciation, use, and assessment of reasons and reasoning as such. As a critical reasoner, one not only reasons. One recognizes reasons as reasons. One evaluates, checks, weighs, criticizes, supplements one's reasons and reasoning. Clearly, this requires a second-order ability to think about thought contents or propositions, and rational relations among them (1996, p. 98).

Burge then seeks to show that the ability to reason critically requires knowledge of one's own mind (Burge 1996), and that such knowledge requires grasp of the concept *I* (Burge 1998).

There is, however, a danger that this argument will collapse into triviality. For what is it to recognize reasons "as reasons"? When Burge says that a critical reasoner appreciates "reasons and reasoning as such," he must mean that a critical reasoner is capable of thinking thoughts that explicitly involve such concepts as *reason*, *justification*, *proposition*, and so on. Otherwise it would not be clear that this ability requires "a second-order ability to think about thought contents or propositions, and the rational relations among them." If critical reasoning requires the ability explicitly to deploy these sophisticated sorts of concepts, however, then it seems unsurprising that this should also require the ability to think of attitudes as *mine*. As Burge in effect argues: a creature that did not grasp that a proposition was something toward which it might have an attitude, and a reason something that might require it to alter its attitudes, would not have understood the concepts *proposition* and *reason*. But what is the interest of this point? Why is mastery of these concepts so important? Certainly the observation that "critical reasoning," so defined, requires the capacity for self-consciousness does not provide much support for the Kantian thesis; it merely leaves us with the question why the capacity for *critical* reasoning should matter, why its attainment should count as effecting a wholesale transformation in a creature's power of representation, rather than merely as adding some special and rarified representations to it.

It may seem unfair to criticize Burge for failing to vindicate the Kantian thesis, since this is not his aim.[7] But even if we concede that Burge has fulfilled his own aims, we can question the interest of his project. The capacity to "think critically about reasons as reasons" sounds significant, but if the phrase "as reasons" just builds in the demand that a possessor of this capacity should be able to employ concepts of propositions and propositional attitudes, then an examination of the preconditions of this capacity will turn out to make less contact with our naïve interest in the nature of rationality than at first appears. Our naïve interest in rationality is an interest in an ability to think in a certain manner ("rationally") about *any* topic, not in our ability to think thoughts about certain *particular* topics ("about reasons as reasons"). An argument which establishes that the latter ability requires the capacity for apperception is of interest only if we can also produce an argument connecting the ability to think about reasons as reasons with the ability to think rationally. Burge, then, has merely equipped us with another way of posing the question that really interests us. We can now ask: "Why does the capacity to 'reason critically' make a categorial difference to the nature of a creature's thought?" But it should be clear that this reformulation does not represent real progress.

A successful "rationality route" defense of the Kantian thesis must thus explain why self-consciousness is essential, but not trivially essential, to rationality. How could this be done? Well, we have seen that it is unpromising to treat the ability to know *what* one's current beliefs are as one thing and the ability to consider whether they are *reasonable* as quite another. This suggests that we should investigate the connection between self-conscious awareness of one's beliefs and knowledge of whether they are reasonable. And to avoid the triviality problem, we must interpret the phrase "knowledge of whether they are reasonable" in the most minimal and uncontentious way. We should not, in particular, make it a stipulative requirement on "reasoning" that it deal with certain sophisticated sorts of contents – e.g., that it make explicit use of such concepts as *belief*, *proposition*, *reason*, or *justification*. For, as we have

---

[7] In fact he specifically dissociates himself from Kant's view. See Burge 1996, p. 99, note 5.

just observed, what we are trying to get at, when we speak of a creature as "rational," is not its ability to entertain certain specific representational *contents*, but its capacity to represent in a certain special *manner*. Only if we can characterize rationality as involving a special manner of representation will we be able to avoid the lingering question: why should *that* ability make a fundamental difference to a creature's power of representation?

3. THE OBJECTIVITY ROUTE

In seeking such a characterization of rationality, it will be useful to consider another, apparently distinct sort of argument for the claim that a capacity for self-consciousness is an essential precondition of our kind of thought. A number of authors have argued that a creature capable of entertaining thought-contents that bear on a world of mind-independent objects must be capable of self-consciousness. We can call this the *objectivity route* to the conclusion that the "I think" must be able to accompany all of my representations, for it portrays the capacity for self-consciousness as a precondition of our even having representations that bear on mind-independent objects. And this line of thought should seem promising, for if the very capacity for objective representation presupposes the capacity for self-consciousness, then self-consciousness does indeed make available to us, not just certain specific representational contents, but a whole new *manner* of representing.

The thought that the capacity for objective representation presupposes the capacity for self-consciousness is also promising for another reason: it seems to have a Kantian provenance. Thus, in §§18-19 of the B-Deduction, Kant claims that there is a connection between the "transcendental unity of apperception" and the "objective validity" of thought. Only insofar as our representations have the kind of unity that permits unified apperceptive awareness, he argues, "does there arise from this relation *a judgment*, i.e., a relation that is *objectively valid*, and that is sufficiently distinguished from the relation of these same representations in which there would be only subjective validity, e.g., in accordance with the

laws of association" (B142). I understand this as follows. To judge is to bring together representations in a way that bears on how things are objectively. To judge "This body is heavy," for instance, is to combine the representations *this body* and *heavy* in such a way that an object is specified and held to have a certain property.[8] This sort of "objectively valid" combination of representations cannot consist in the obtaining of any mere "law of association" connecting tokenings of the one representation (a representation of a certain body) with tokenings of the other (a representation of heaviness). For any tendency I have to associate from the representation of this body to the representation of heaviness will have mere "subjective validity": it will be a connection of representations that holds only for me, and carries no implication about what any other subject (or indeed about what I) *should* associate with what (compare B140). By contrast, when I judge that a certain body is heavy, I represent in a way that subjects itself to standards of truth or falsity, standards which imply that if my representation here and now is correct, then *any* representation of that object must agree with it. In this case, I represent the body in question and the attribute of heaviness as combined in what Kant calls "an objective unity": I represent them as combined "in the object, that is, no matter what the state of the subject may be" (B142). This is what the copula "is" signifies: that two or more representations stand together not merely by some accident of the subject's psychology, but with a kind of necessity – the kind captured in the remark that if my judgment that this body is heavy is correct, then *any* other correct judgment about this body *must* agree with mine on this point. Objective validity is thus *necessary universal* validity (compare P §18-19, A104).

Now, Kant claims that a creature incapable of apperception would not be capable of objectively valid representation either. If this were true, it would have radical implications for how we must think of the representations of creatures that lack self-consciousness. For a creature that cannot combine the representations *body* and *heavy* in objectively-significant judgments such as "This body is heavy"

---

[8] This is Kant's example: see B142-3.

presumably cannot have the representations *body* and *heavy* separately either: these representations plainly are what they are only in virtue of the fact that they can enter into such combinations. More generally, Kant's claim would entail that creatures which lack self-consciousness are incapable of representing objects and their properties – at least in the sense in which we self-conscious creatures are capable of representing objects and their properties. We might still leave room for another, less demanding sense in which creatures that lack self-consciousness can be said to have objectively significant representations: this, in effect, was the strategy of the last chapter. But if we could explain why, in one significant sense, a creature that is incapable of having the representation *I* is also incapable of objective representation, we would have shown that a self-conscious creature's powers of representation differ categorially from those of a creature that lacks self-consciousness.

I think we can find an argument for this conclusion in Kant's own texts, and I will turn to that argument shortly. Kant's reasoning will stand out more clearly, however, if we first consider a more recent attempt to supply such an argument.

*

To anyone familiar with the recent history of anglophone Kant scholarship, the claim that only a creature capable of having the representation *I* can represent mind-independent objects should sound familiar. P. F. Strawson famously attributes such a claim to Kant in his seminal book *The Bounds of Sense* (1966). Strawson calls the claim that the "I think" must be capable of accompanying all of my representations the "necessary self-reflexiveness" of experience, and he argues that

> [w]hat is meant by the necessary self-reflexiveness of a possible experience in general could be otherwise expressed by saying that experience must be such as to provide room for the thought of experience itself… What is necessary is that there be a *distinction*, though not (usually) an *opposition*, implicit in the concepts employed in experience, between how things are in the world which experience is of and how they are experienced as being, between the order of the world and the order of experience (1966, p. 107).

Were there no such distinction, then, according to Strawson, experience could not present us with a world of mind-independent objects at all. But, he argues, a creature's concepts only contain this crucial

opposition if it grasps the contrast between "This is how things are" and "This is how things are experienced as being" – which grasp seems to require the capacity for self-consciousness, or anyway the crux of it.

Strawson's own defense of these claims is convoluted, and I do not think it will be worth our while to try to puzzle through it. But Strawson has better thoughts on this topic than the ones he presents in *The Bounds of Sense*. In the second chapter of his earlier book *Individuals* (1959), he offers a simple and intuitively compelling reason for thinking that the capacity for objective representation requires the capacity for self-consciousness. The basic thought is this: to represent an object is to represent something that can exist unperceived, and the very idea of existence unperceived requires a certain surrounding. Specifically, a subject can only represent things capable of existing unperceived if he can distinguish between "himself and his own states on the one hand, and other particulars of which he has knowledge or experience on the other" (Strawson 1959, p. 64). A creature that does not grasp this distinction cannot understand the objective significance of its own sensory states.

In Strawson's discussion this line of thought is only sketched, but it is discussed in greater detail in an illuminating commentary by Gareth Evans. To represent a mind-independent object, Evans suggest, is to represent something that can exist unperceived. And, according to Evans,

> the idea of unperceived existence, or rather the idea of existence now perceived, now unperceived, is not an idea that can stand on its own, stand without any surrounding theory. How is it possible that phenomena *of the very same kind as* those of which he has experience should occur in the absence of any experience? Such phenomena are evidently *perceptible*; why should they not be perceived? To answer this question, some rudimentary theory, or form of a theory of perception is required. This is the indispensable surrounding for the idea of existence unperceived, and so, of existence perceived (1985c, pp. 261-2).

Not just any state of sensory consciousness is intelligible as bearing on a world independent of the subject. Pain, for instance, is not an awareness of something that exists independently of the subject's awareness of it: there does not seem to be any such thing as a pain's existing unperceived. What, then, is required for a state of consciousness to amount to a representation of a mind-independent object? Evans' answer is that the subject must conceive of the object of the state as one that can exist

unperceived, and that this requires that the subject have some grasp of herself and her faculties as a perceiver. She must grasp such facts as that she can only see the object in question if there is adequate light, that her perceptions of its color will be distorted if the light is of the wrong kind, that she will be prevented from seeing the object if it is occluded by another non-transparent object, and so on. Hence she must be able to think of her own perceptual states as such. And this seems in turn to require that she possess the concept *I* and the concept of a representational state.

Evans illustrates this point by asking us to consider a child who is trained to make a certain utterance in response to a certain sensory stimulus. We might, for instance, teach her to utter the words "That's red!" when something red is put in front of her. Now, when does this cry come to express an awareness of a circumstance understood as capable of obtaining independently of her awareness of it? We have already seen, in the last chapter, that it need not express such an awareness at the outset: merely to learn to respond differentially to a certain characteristic pattern of sensory stimulation need not involve understanding that some enduring object or circumstance is the cause of that pattern of stimulation. The thought Evans takes from Strawson, and that Strawson claims to find in Kant, is that we are not entitled to interpret her cry as expressing an awareness that the thing in front of her is red (rather than merely as expressing a state of consciousness without objective purport) until she is intelligible as grasping that "That's red" can be true when she is not aware of it. And this, Evans argues, requires that she understand what is occurring when she *does* have the state of consciousness to which she gives expression with the cry "That's red!" as connected with the fact that the thing in question is red by some condition that is sometimes but not always satisfied (e.g., that the lighting is right, that the object is not occluded, etc.). In short, she must have a conception of herself and her capacities as a perceiver, such that she can understand how the very same circumstance that gives rise to her experience can exist without her experiencing it:

> The proposition [e.g., "That's red"] will then be understood to entail that, if that condition is satisfied, it may be perceived to be true. In the formulation of the condition there lies a theory, or the form of a theory, of perception (Evans 1985c, p. 262).

This line of thought is obviously more suggestive than conclusive. For one thing, it is vague about exactly what concepts a creature capable of representing mind-independent objects must possess. It is possible to imagine objections to the claim that a creature capable of representing objects as such must have the concept *I*, or the concept *representation*. Perhaps an understanding of unperceived existence can be constructed from simpler conceptual materials.[9] Moreover, it is possible to imagine objections which deny that a creature capable of representing mind-independent objects must be capable of *conceptualizing* any of these ideas or distinctions. Perhaps it is true that, for a creature to count as representing mind-independent objects, it must in some sense *be sensitive* to the distinction between its own states and the objects on which those states bear. Still, it might be asked, why should this sensitivity take the form of a capacity explicitly to conceptualize this distinction?[10] Our willingness to attribute representations of objects to dumb brutes might be taken to show that a creature can display whatever sensitivity is required without being able to conceptualize itself or its own representational states at all.

Perhaps the most serious objection to the Strawson-Evans argument, however, concerns its reliance on the claim that to represent an object is to represent something that can exist unperceived. If this claim is not to beg the question, it had better not mean that to represent an object is to represent something *as* capable of existing unperceived – at least not if this involves the requirement that the subject must grasp the concept *exists unperceived*. For this would not be obvious, and certainly would not follow from the uncontroversial point that garden-variety objects *are in fact* things that can exist unperceived. That a creature capable of representing objects must be capable of framing the concept *exists unperceived* (a concept which presumably can only be framed by a creature who also grasps the idea of a subject's perceiving something) is part of what the argument should prove, not something it can

---

[9] It might, for instance, be argued that a creature can possess the idea of unperceived existence without having the idea of existence unperceived *by me*; or again, that a creature can possess the idea of *unperceived* existence without the more general idea of a contrast between objects and representations of objects. Quassim Cassam entertains several such objections (not to the exact argument considered here, but to closely related arguments) in Chapter 3 of his *Self and World* (1997).

take as a premise.

When Strawson and Evans claim that a creature capable of representing objects must grasp "the idea of unperceived existence," then, the idea of unperceived existence they have in mind must already be somehow present in the ground-level thought "*a* is *F*," which does not mention the subject or her representational state. For their intended conclusion is not the nearly trivial claim that a creature lacking the concepts *I* and *representation* cannot frame the concept *unperceived existence*, but rather the substantive and interesting claim that a creature lacking the concepts *I* and *representation* cannot represent things that are potentially *instances* of the concept *unperceived existence*. But now it is far from clear how the idea of unperceived existence might be implicitly involved in ground-level thought about objects (thoughts like "*a* is *F*," rather than thoughts such as "I see that *a* is *F*" or "*a* exists unperceived," which explicitly involve concepts of representation). The claim that to represent an object is to represent something that can exist unperceived follows from the uncontroversial point that ordinary objects can in fact exist unperceived, it seems, only if the phrase "something that can exist unperceived" is read *de re*. But if the phrase is read in this way, then it is not clear how the claim can underwrite the conclusion that the capacity for objective representation requires grasp of the concepts *I* and *representation*. For to say that a subject who represents an object must represent something *which can in fact* exist unperceived is not obviously to set *any* requirements on how the subject must represent the thing in question. Thus, on closer scrutiny, the argument propounded by Strawson and Evans seems simply to fall apart.

Nevertheless, I think the Strawson-Evans argument has strong intuitive appeal. Focusing on the question when we count a creature's utterances as expressing empirical judgments is a useful way of making this appeal vivid. We have already agreed that the capacity to judge, e.g., that things are red, must involve more than the mere disposition reliably to utter the words "That's red!" when in the presence of a red thing. A child that has learned to do this has learned to use the sentence "That's red"

---

[10] Robert Brandom has often put this sort of objection to me in conversation. I take the whole approach of his *Making It Explicit* (1994) to be predicated on a rejection of the sorts of claims made by Strawson and Evans.

correctly in *one* sense, but the sense in question is one that might apply to a trained parrot or perhaps even to a color-detecting device. It is, one wants to say, a correctness by *our* standards, but not yet a correctness by standards grasped by the speaker herself. We have gotten the child to produce a sentence of ours in appropriate circumstances, but nothing we have yet required ensures that the child *understands* the sentence she produces – understands that in producing these noises she is saying of something she perceives that it has a certain property.

What, then, would entitle us to attribute such an understanding to her? Without attempting to specify conditions that would satisfy a behaviorist, we can surely say this: her speech must come more generally to reflect an understanding of the separate significance of the words that make up the sentence "That's red." Her "that" must become intelligible as a term which refers demonstratively to objects, her "red" as a predicate of objects. Her vocalizations must, in short, acquire whatever complexity is required for speech to count as expressing thoughts about how things are objectively. This, in effect, is just the point we noted in the last chapter: that a creature's vocalizations in response to circumstances only come to express empirical *judgments* when they begin to be intelligible as reflecting a grasp of a general framework for thinking about objects and their properties.[11] And really this point bears not just on speech but on "mental representation" as well, for we can equally ask when we should count the sensory state to which the child's cry "That's red!" gives vent as having a content like *Something red over there*, rather than a merely subjective kind of content comparable to the content of an experience of pain. And here again, the right answer seems to be that the sensory state in question should count as possessing an objective content only once it comes to stand at the intersection of two ranges of representational contents, one about objects and another about properties of objects. (I mean, of course, that it must stand at the intersection of two ranges of contents which the subject has the standing *capacity* to

---

[11] For a highly illuminating discussion of what this might involve, see Ernst Tugendhat, *Traditional and Analytical Philosophy* (1982), esp. Lectures 19-27. And see also Tugendhat 1986, where it is argued that the points made in Tugendhat 1982 entail that only a subject who has mastered the first person can make empirical judgments (see pp. 63-67).

entertain. She need not actually think any particular thought, but her mental life must have enough complexity that it makes sense to attribute the relevant representational capacities to her.)

Now, we can think of Strawson and Evans as urging a certain view about what must belong to this range of representational contents: namely, that it must include contents relating to the subject herself and her own representational states. This, I think, is the best way to understand their claims about the conditions under which a subject can represent things capable of existing unperceived: as urging that a sensory state only becomes intelligible as having a content bearing on things capable of existing unperceived insofar as it becomes part of a set of systematically-related states, including ones whose content bears on the subject herself and her own representations. For the moment, we can take this simply as a challenge: objects are things that can exist unperceived, and we have granted that mere reliable responsiveness to what are in fact objects and their properties does not constitute judging about them. But then if we do not suppose that a creature capable of representing objects must distinguish between its own states and the states of things independent of it – and must be able to mark this distinction by making two sorts of judgments, ones that predicate properties of objects and ones that ascribe representational states to the judging subject – what account *can* we give of the preconditions of objective representation?

*

On first hearing, I think the argument propounded by Strawson and Evans sounds remote from anything that can be found in Kant. I hope, however, that the foregoing discussion suggests that the connection is not so distant. For consider again the Kantian claim that I mentioned at the beginning of this section: that only a subject capable of apperception can have "objectively valid" representations, representations whose content bears on how things are "in the object, regardless of what the state of the subject may be." The best way to understand Strawson and Evans's claims about the conditions under which a subject can grasp mind-independent objects, I have suggested, is as addressing just this topic: the conditions under which comprehending ground-level judgments about objects are possible. Moreover,

the conclusion that Strawson and Evans draw, that such judgments are possible only in the context of a capacity to make another kind of judgment, one about the perceiver and her own perceptual states, amounts to a special case of Kant's conclusion, that objectively valid representation requires the capacity for apperceptive awareness of oneself and one's own representations. Finally, the way that Strawson and Evans seek to make this conclusion plausible – namely, by focusing on the fact that the mere establishment of a reliable correlation between how things are objectively and what vocalizations the subject makes does not suffice to constitute the relevant vocalizations as assertions – has its counterpart in such Kantian remarks as the following:

> If I remove from empirical knowledge all thought (through categories), no knowledge of any object remains. For through mere intuition nothing at all is thought, and the fact that this affection of sensibility is in me does not amount to a relation of such representation to any object (A253/B309).

To be sure, Kant casts his observation as one about the relation between sensation and objectively contentful experience, rather than as one about the relation between reliable dispositions to vocalize in response to circumstances and utterances that express empirical judgments. I hope, however, that I have said enough to bring out the common structure of the underlying point.

4. KANT ON CONCEPTS AND APPERCEPTION

We have been exploring two sorts of arguments for the Kantian thesis: one that derives the requirement that the "I think" should be able to accompany all of my representations from the nature of rationality, and another that derives the requirement from a reflection on the preconditions of objective representation. We have seen that there is something attractive in each of these arguments, but that each is beset by difficulties. The basic difficulty for the former approach is to find a clear and non-question-begging interpretation of the thought that a rational creature must be able to "back up" from its representations. The basic difficulty for the latter is to explain why and in what sense a creature capable

of objective representation must grasp the distinction between its own states and the states of independent objects. Now, I have been discussing the two approaches as if they were distinct, but in fact I think that a successful defense of the Kantian thesis must reveal the two routes as just different angles on the same point. This will become clear if we consider Kant's own account of the connection between the capacity for conceptual representation and the capacity to accompany representations with the "I think." As we shall see, this account involves a synthesis of two ideas about the character of conceptual representation, ideas that correspond to the two lines of thought we have been considering.

*

Kant famously insists that our intellect is a *finite* power, one that does not produce its own object, one that can only cognize an object given to it through another power, sensibility. He also insists that ours is a *discursive* intellect, one that cognizes by bringing given representations under general representations of the kind he calls "concepts." To understand the connection between these two characterizations of our intellect is to understand why the rationality route and the objectivity route must be different angles on the same point.

To say that our intellect is finite is to say that it works on material that it does not itself produce. It is not the cause of its objects, not the source of the truths that it knows. Rather, its knowing consists, at least in the fundamental case, of bringing order to impressions that result – as the word "impressions" itself suggests – from our senses being affected by independently-existing objects. Depending on one's frame of mind, these remarks will sound either platitudinous or controversial. No one but a radical skeptic or a thoroughgoing idealist will deny that our knowledge of the world around us depends on our senses being affected by objects that we do not create. This is a platitude.[12] What will perhaps seem controversial is the transition from this point to the claim that we know those objects by "bringing order

[12] It is an interesting question exactly what sort of platitude it is. It seems that it could not be an empirical proposition: for how could we know on the basis of experience that what we have is in fact *experience* – that is, a receptive awareness of independently-existing objects? The character of our knowledge of our finitude deserves attention, but I cannot take up this topic here.

to sense impressions" – that, to put it in more Kantian language, our understanding must "unify a manifold of receptivity" in order to cognize any object (compare, e.g., A120, B129-30, B136-7). For this talk of unifying a receptive manifold may sound as though it presupposes something like a sense data theory of knowledge – a theory on which our knowledge of the world around us is founded on an epistemologically prior body of knowledge about our own sense impressions, on the basis of which we construct or hypothesize a world of independently-existing objects. And this sort of view of the foundations of empirical knowledge has been the object of forceful attacks.

At this point, however, we are in a position to make a different sort of sense of the thought that our understanding cognizes by unifying a manifold of impressions. For an important theme of this and the preceding chapter has been that there is a kind of responsiveness to impressions that does not yet involve any understanding of their objective import, and that any given impression only comes to have an objectively-significant content insofar as it is grasped as having a determinate location in a system of possible contents with a certain sort of structure – a structure that discriminates, on the one hand, enduring objects, and on the other hand, their various and changeable properties. We have not investigated the required structure; we have merely noted that, trivially, to count as objectively significant, a subject's representing must acquire *whatever* complexity is necessary for representing objects and their properties. But it is clear that the required complexity will involve grasp of relations among different contents: to grasp, e.g., a given impression as a representation of a certain enduring object's being green will involve understanding such things as that (1) if the representation in question is correct, then something which is correctly represented as not-green at the moment in question must be another object; and that (2) if the representation in question is correct, then a correct representation of that very object as not green must be a representation of it at a different moment in its history. Such principles merely make explicit distinctions that are contained in any ground-level representation of some perceptually presented thing as having a certain property: to have a perceptual impression a certain thing's being thus-and-so is to represent in a way that stands in such relations, and such representations

are possible only in a creature whose representing is intelligible as taking account of relations of these kinds.

What this "taking account" can come to is a question I will turn to presently. But however we fill in this requirement, we already have in its schematic form a basis for interpreting the thought that objectively significant representation requires a synthesis of the receptive manifold. For what we have seen is precisely that any given impression only attains to objective significance insofar as our grasp of it relates it to a manifold of other possible representations, including other possible impressions: to discern in a given impression a representation of an enduring object with various changeable properties just is to grasp the impression in question as standing in such relations. This is not to suggest that this synthetic grasp is the product of cogitation upon an "unsynthesized manifold" of impressions, supposing that there could be such a thing. It is rather to claim that, *qua* knowing beings, our *basic* grasp of our own impressions is one that involves a synthesis in the sense just indicated.

The finitude of our intellect consists in the fact that our cognition depends on our receiving a manifold of impressions, which is to say, an ever-changing series of contents whose primary form might be represented as

this *F*'s being *G*

where "this" marks the fact that what is in question is a perceptual presentation of an object, *F* is a sortal concept which contains principles of individuation and identity for objects of the relevant kind, *G* is a property of such objects, and the connective word "being" (as opposed to "is") marks the fact that what is in question is merely an impression, not necessarily something the subject judges to be the case. We see what qualifies such representations as impressions in seeing how they might arise as *successors* of modes of affection with merely subjective contents, in understanding how the capacity to have "this"-representations depends on perceptual capacities, and in seeing the contrast between the mode of combination involved here and the mode of combination involved when the subject combines the relevant representations in a judgment ("this *F* is *G*"). But there is no commitment in this account of

impressions to the idea that an unsynthesized manifold of sensation has any epistemic role to play in our cognition.[13]

If our intellect is finite in this sense, however, then it must also be discursive. For we have seen that impressions only take on objective significance insofar as they are grasped as bearing on enduring objects and their changeable properties, and that this requires that any given impression should be grasped as standing in determinate relations to various other possible impressions. But it is precisely the discursiveness of our cognition that effects this grasp – or it might be better to say: our grasp of the relatedness of our representations just is the discursiveness of our cognition. For to bring one's impressions to discursive consciousness is precisely to discern in them instantiations of features that might also be instantiated in other impressions. It is, so to speak, to grasp the here and now as a token of what might be elsewhere or at another time. And this, as we have seen, is the *sine qua non* of objective representation.

Earlier I postponed the question what it can come to for a subject's representing to take account of the relation of a given impression to other possible representations. We must now return to this issue. It is by now a familiar point that there can be a nonrational, nonconceptual analogue of objective representation – or, as we can also say, a lower grade of objective representation. In this lower grade of representation, we said, impressions are connected with one another, and with behavior, according to empirical laws of association: laws that allow the representer to track what are in fact enduring objects and to discriminate what are in fact their changeable properties, but that do not reflect an exercise of capacities to think abstractly about such topics. What is special about discursive representation comes out precisely in its contrast with this more primitive mode of being onto the objective. An associative intellect can be said to be in touch with general properties of things in the following sense: its

---

[13] Thus Kant describes sensation as the *matter* of knowledge, which receives the *form* of thought about objects and their properties (A86/B118). Famously, the contrast of matter and form, which Kant inherits from the Scholastics and the Scholastics inherit from Aristotle, is not supposed to distinguish two potentially separable *elements* in the constitution of a thing. It is supposed to distinguish two *aspects* of what it is for a thing to exist.

representing can be governed by laws which lead it to have a certain association whenever it has an impression of (what is in fact) a thing with certain property. For instance – to return to our Leibnizian example from the last chapter – it may associate impressions of brandished sticks with pain. To say that it associates according to such a principle is to attribute a triple generality to its representing: it is capable of associating impressions of brandished sticks with one another, of associating pains with one another, and lastly of associating impressions of brandished sticks with pain. But in each case, the generality is merely a matter of the creature's moving from one representation to others that *we* recognize as related to it. And this clearly can occur in without the creature's having a capacity *itself* to reflect on the relevant commonality: to think such thoughts as "Here's another stick being brandished" or "Brandished sticks tend to be followed by something painful." The generality of an associative intellect's representing is not *for it*: it makes transitions according to general principles, but not in virtue of a cognizance of those principles. When we say of such a creature that it represents something as having a certain property, the significance of our ascription is to be understood in accordance with these constraints.

By contrast, a creature with a discursive intellect is one whose representing reflects a grasp of the general as such. When we say that such a creature represents something as having a certain property, we imply that it brings the object in question under a predicate that *it understands* as potentially capable of applying to other objects. It is this grasped generality that Kant regards as the defining feature of conceptual representation: thus he remarks that "every concept *must be thought as* a representation which is contained in an infinite number of different possible representations (as their common character)" (A25/B40, emphasis added). And this grasped generality informs all of the explanations in which ascriptions of conceptually contentful representations figure: to say of a creature with a discursive intellect that, e.g., it believes *a* is *G* because it sees that *a* is *F*, is to imply that it accepts some general proposition that connects being *F* with being *G*. If we offer a superficially similar explanation of a belief held by a creature with an associative intellect, by contrast, the "because" carries no such implication: the transition that is explained, in the case of a discursive intellect, by the *representation* of a general

proposition, is explained in the case of an associative intellect by the operation of a law whose effect does not depend on its being represented.

In making these observations, we have arrived, in effect, at an interpretation of Kant's claim that a rational creature can only acquire knowledge from a receptive manifold by "uniting" the manifold in "a concept of the object" (B139). What Kant means by saying that the manifold must be "united" is just this: that particular impressions must be brought under (united by) representations with the kind of grasped generality we have just been discussing. Failing this, the creature will just be subject to a play of impressions in which it *recognizes* nothing, not even what its impressions are – for even this kind of recognition, *qua* recognition, would require that the relevant impressions be brought under concepts. This is just what Kant says in the letter to Marcus Herz quoted in Chapter I:

> [W]ithout those conditions [viz., "the specific character of our kind of intuition" and – what is relevant for our purposes here – "the uniting of the manifold in a consciousness"], all sense data for a possible cognition would never represent objects. They would not even reach that unity of consciousness that is necessary for knowledge of myself (as an object of inner sense). I would not even be able to know that I have sense data; consequently for me, as a knowing being, they would be absolutely nothing.[14]

Of course this passage also contains a further thought of which we do not yet have an interpretation, namely that the relevant unity is the unity *of consciousness*. I take this to be the same point that Kant expresses at the beginning of §18 of the B-Deduction, in a sentence of which I have already quoted a part: "The transcendental unity of apperception is that unity through which all the manifold given in an intuition is united in a concept of the object" (B139). This claim connects the thoughts we have been working through with the topic of apperception. Our final task in this chapter will be to make sense of this connection.

*

Why should the unity of representations in a concept be identified with the unity of apperception? Briefly, the answer is this. (1) A creature can bring its representations to the unity of

concepts only if it can represent general propositions. But (2) such representations are only ascribable to creatures that are capable of putting to themselves the sort of truth-oriented "Why?"-question that has been the object of our attention over the last few chapters. And (3) the capacity to put this question to oneself is already tantamount to the capacity for apperception. I will discuss these points in turn.[15]

Point (1) is something we have just noted: to ascribe a representation of an object's having a certain property to a creature with a discursive intellect is to presuppose that it has the power to reflect on general propositions concerning that property, and to make transitions from one particular representation to others on the basis of such reflection. This thought was developed at length in the last chapter, so I will not spend more time elaborating it now. I will simply note that, although the distinction between a power of representation informed by a capacity to reflect on general propositions and a power of representation not thus informed provides us with a schema for understanding the difference between rational and brute intellect, this schema obviously needs filling in at a crucial point: we need to know what the capacity to *reflect* on general propositions amounts to. After all, the capacity for such reflection is not supposed to manifest itself in some one thing a creature does, but to transform our understanding of everything it does. But in what does this general transformation consist?

Point (2) contains our answer to this question. A creature should count as capable of reflecting on general propositions just if it can consider, with regard to any given proposition, the question whether that proposition is true and why. To say this is not to give a particular behavioral criterion for discursive thought: recognizing the power to entertain this sort of question will obviously be part and parcel with recognizing a certain kind of intelligibility in a creature's life as a whole. Nevertheless, in making this

---

[14] Letter to Marcus Herz, May 26, 1789 (Ak. 11:51-52).

[15] Note that if we can demonstrate these points, we will have achieved the promised synthesis of the rationality and objectivity routes to the necessity of self-consciousness. For we will have argued that it is a condition of acquiring *discursive* knowledge about *a world of independent objects* that we be self-conscious. The objectivity route sought to show that self-consciousness is a precondition of knowledge of a world of independent objects, the rationality route that, in effect, it is a precondition of discursive knowledge (inasmuch as discursive knowledge requires the capacity to "step back" and consider the grounds of one's own beliefs). We will have shown that the necessity of self-consciousness in fact reflects the interrelationship between these two aspects of our intellect.

connection, we give some content to the idea of a capacity "to reflect on general propositions." For note, first of all, that the capacity to answer such a why question will necessarily involve the capacity to reflect on general propositions, since any conclusion that rests on a particular ground will always also rest on some general principle connecting that ground with the conclusion, and a creature capable of making its reasons explicit must be capable of articulating this principle. And note, secondly, that a creature capable of asking why a certain proposition should be believed will be, as we noted in the last chapter, one that can *make up its mind* about what to believe. For even to count as capable of entertaining the relevant question, a creature must in general be capable of having its answer to the question determine what it believes. If this condition were not satisfied, then the creature's supposed answer to the question would not reflect its belief about what is true, and a creature that could not in general have its reflection determine its belief in this way would not count as capable of considering the truth-oriented "Why?"-question at all.

A creature capable of entertaining this "Why?"-question is thus, in a clear sense, a creature whose holding-true is potentially mediated by reflection on general propositions. Moreover, this potentiality informs all of its beliefs, even those that are not actually the outcome of reflection. For if we say of such a creature that, e.g., it believes that *a* is *G* because it believes that *a* is *F*, we presuppose that it could ask itself why this connection holds, and our explanation implies something about the answer it is committed to giving: namely that it should consist in some proposition relating being *F* to being *G*. To apply this form of explanation to a creature capable of reflecting on reasons for thinking a proposition true is thus to imply that it is at least tacitly committed to some general proposition.

But what does this have to do with the capacity for apperception? At the beginning of this chapter, we asked why the occurrence of the word "I" in the question "Why do I think that *p*?" is not merely a grammatical accident – why the capacity to reflect on the grounds for thinking a proposition true should have to involve the capacity to think of oneself. How has the foregoing discussion helped us with this question? —It will not help us so long as we leave the content of the representation *I* itself

unexamined. We might express the spirit of the worry that is troubling us as follows: "I know what it is to think about the truth of propositions, but why should the capacity to do *that* require that I be able to think about *myself*?" – where this is asked, perhaps, looking down at the sorry heap of flesh and bone that it is my lot to be. But what does it mean to call this heap of flesh and bone "myself"? What is the self of self-consciousness?

Kant says that the "I" of apperception is "a representation of the spontaneity of the thinking subject" (B278). It will help us to see what he means, I think, if we first approach the topic of self-consciousness from a quite un-Kantian angle. Let us ask: How would we identify an expression in an unfamiliar language as a form of the first person? What role or roles would that expression have to play? A familiar and widely accepted answer – associated above all, perhaps, with David Kaplan – is that the relevant expression would have to function as an indexical that refers to whoever utters it.[16] But if what qualifies a term as an indexical is just that it refers, in any given context, to a certain element of that context, then this account of the function of "I" turns out to be inadequate. For if an expression is to count as a form of the first person, it must present its object in a certain way, namely as the *subject* of thought and action. But it is possible to describe indexicals that refer to the speaker but that plainly do not meet this condition.

This can be shown by slightly altering G. E. M. Anscombe's well-known example of the "A"-language, a language in which each speaker uses the expression "A" as a name for himself.[17] In Anscombe's example, each speaker learns to report observations about the person who is in fact himself (e.g., that his arms are crossed) using the name stamped on his wrist – which is in every case "A" – and to report observations about others using names stamped, at points invisible to their bearers, on their chests and backs. In a recent paper, John McDowell proposes that we vary the example to make "A"

---

[16] See Kaplan's "Demonstratives" (1989).

[17] See Anscombe, "The First Person" (1975), pp. 143-144.

function as an indexical.[18] This can be accomplished by supposing that the speakers are taught to use "A" to report observations concerning the body at the front of whose head is the point from which the scene is observed. Given well-known facts of human physiology, this will ensure that "A" refers to the person who is doing the observing, and, in fact, to the person who utters "A."[19] But so long as the only application of "A" is one that involves bringing it under predicates known to apply through observation, the term clearly does not express self-consciousness, and hence is not a true first person. McDowell shows this by noting that a user of the modified "A"-language would have only observational knowledge of the actions of the person he called "A," even though that person was in fact himself. We can bring out the same point in a different way by considering how a user of this language could know the beliefs of the person he called "A." Suppose the person in question thinks that *p* and is prepared to say so. Given that he applies "A" only on the basis of observation, it is perfectly coherent for him to say in the same breath: "*p*. But does A believe that *p*?" Of course he will straightaway have his answer, at least if he does not think too hard about what is entailed by A's having made this odd remark. But the fact that this can be even momentarily an open question already shows that "A" is not a true first person. For, as G. E. Moore famously pointed out, there is an incoherence involved in saying "*p*, but I don't believe that *p*," and this is an incoherence that anyone who understands the first person must be in a position to recognize.

This observation begins to answer our question about what would qualify an expression in an unfamiliar language as a form of the first person. To understand the use of "I" involves understanding that, if one is prepared sincerely to assert that *p*, then one is automatically entitled to assert "I believe that

---

[18] John McDowell, "Referring to Oneself" (1998a), pp. 138-140.

[19] If someone wants to argue that this term should not count as referring indexically to the speaker since it is possible to imagine strange circumstances in which it would not refer to the person who utters it, I have no objection; but I would then ask how we would go about telling, of an expression in an unfamiliar language, that it is an indexical referring to *the speaker*. Thinking this through would, I think, simply lead us by another path to the conclusions I am going to recommend. The term "the speaker" (which really means: the *subject* who is speaking) actually contains our whole problem, a problem on which we make no real progress by appealing to the general notion of an indexical.

p." Any expression that was not governed by such a rule would not be a form of the first person.[20] This is a way of getting at what is correct in the traditional idea that the essential nature of the I is to be a thing that thinks: the truth in this idea is that to recognize that *I* am a certain person is to recognize that that person is the one who thinks what I think (i.e., believes what I believe) – including this very thought. To say that I am this heap of flesh and bone is to identify this corporeal person as the one that has my thoughts (an identification whose possibility depends, no doubt, on my ability to identify it also as the one whose actions implement my *practical* thinking about what to do). For the crux of self-consciousness is just this: to know one's own thoughts without observation.

But now what is involved in knowing one's own thoughts without observation? How can one tell when one is prepared sincerely to assert that *p*? If one merely knew by observation that one was disposed to blurt out certain sentences, then, obviously, one's having the further disposition also to blurt out those same sentences preceded by "I believe that" would not make this "I" into an expression of self-consciousness. An utterance of a sentence only expresses what *I* think if it is an expression, not of some mere disposition I find in myself, but rather of my reflection on what is true. And this point connects the capacity to know one's own thoughts without observation with the capacity to consider the truth-oriented "Why?"-question. For the capacity to reflect on this question just is the ability to determine whether to affirm a given proposition on the basis of a reflection on grounds for holding it true. What explains our entitlement to accompany any sentence we are prepared sincerely to assert with "I believe" is that, when we are being sincere, our assertions express our assessment of what is true – which is to say, our beliefs. We can thus restate our requirement for counting an expression as a form of the first person in a more informative way: an expression E is a form of the first person only if it is governed by the rule that a person who takes *p* to be true is automatically entitled to say "E believes that

---

[20] For the moment I will only claim that this is a necessary condition for being a genuine first person, not that it is a sufficient one. As I have noted, McDowell argues that a genuine first person must be connected with nonobservational knowledge of one's own actions. I think this is right, but I am not sure that it constitutes a further condition distinct

p." This is the "transparency" of the question what I believe to the question what *to* believe that has recently been emphasized by a number of authors.[21]

We can now return to Kant's claim that *I* is a representation of the spontaneity of the thinking subject. Our interpretation of the kind of spontaneity characteristic of a rational creature has been just this: it is the spontaneity involved in the capacity to reflect on the truth-oriented "Why?"-question. And we have just seen that the expression "I" is indeed a representation of just this spontaneity. For its defining role is in marking exercises of this capacity: to accompany a proposition one holds true with "I believe" is just to make explicit that the proposition in question reflects one's view about what is true, and to identify a person as *me* is to see that person as the locus of my capacity to do this.

If this is right, then our initial question about what the expression "I" is doing in the question "Why do I think that *p*?" was misleading. The expression "I" would not be the expression that it is – one that expresses self-consciousness – if it did not figure in this question (and, of course, in questions related to this one, such as whether I *should* think that *p*). And this is the key to understanding Kant's claim about the connection between concepts and self-consciousness. We have seen that the capacity to reflect on the truth-oriented "Why?"-question is what makes discursive consciousness possible. What we can now see is that this is also the crux of self-consciousness – of the knowledge of one's own spontaneity that becomes explicit when one learns to say "I think." This is not to claim that a creature which has not mastered an expression for the first person cannot have discursive consciousness; but it is to claim that a creature cannot have discursive consciousness without having the kind of knowledge to which it is the function of the first person to give expression. A creature that has not yet mastered the first person lacks a term that would give explicit expression to self-consciousness, but it cannot lack the awareness in which self-consciousness consists – a capacity not merely to have representations but to

---

from the one I have been describing, as opposed to a condition that would turn out, on further examination, to be contained in the condition I have described. I will say more about this shortly.

[21] The most important of these authors is Richard Moran. I say more about Moran's work, and about the connection between belief, assertion, and the capacity to reflect on grounds, in the next chapter.

reflect on them – without lacking discursive consciousness altogether. And this, in effect, is what Kant says in the passage from his *Anthopology* quoted at the head of Chapter I:

> That man can have the representation *I* raises him infinitely above all the other beings living on earth… This holds even if he cannot yet say 'I'; for he still has it in mind… For this power (the ability to think) is understanding. (A I, §1; Ak. 7:129)

What I have just been giving is an account of why it makes sense to say a creature with the power of understanding, i.e., with discursive consciousness, "still has the representation *I* in mind" even if he does not yet possess an expression for it. To repeat: the capacity not merely to have representations but to reflect on them is the crux of self-consciousness, what becomes articulate in our use of "I". But any concept-user must have this power, and know that it has it, even if it doesn't yet have a special expression with which to mark this recognition. For it must be capable of facing the truth-oriented "Why?"-question, and to be able to answer this question is already to know everything it needs to know to answer the question what it thinks. This is the third and final step in our account of the connection between concepts and apperception: (3) the capacity to put to oneself the truth-oriented "Why?"-question is already tantamount to the capacity for apperception.

What led us to wonder what the expression "I" is doing in the question that inquires after grounds for thinking a proposition true was, no doubt, our sense that self-consciousness involves *more* than knowledge of our own thoughts, so that to know that such-and-such is what *I think* is to know something substantive. In ascribing a thought to myself, am I not thinking of it as belonging to a particular embodied creature, one possessed of all sort of properties, liable to all sorts of affections, and capable of all sorts of actions? No doubt I am, but two points must be noted. First, as we have already seen, what qualifies my awareness of any of these other facts as *self*-consciousness is that they are ascribed to the very thing whose thoughts I know without observation: for any expression *e*, my knowing that *e* is *F* will count as self-conscious awareness only if in knowing this I know that the thinker of my thoughts – including this very thought that *e* is *F* – is *F*. Secondly, if knowledge of any of these remaining facts is a *necessary* part of our self-consciousness, this must be because we cannot have self-

conscious awareness of our own thoughts without also having this other knowledge.[22] For suppose we could have self-consciousness awareness of our own thoughts without having this other knowledge. Then it would not be necessary that the thinking subject, which as we have seen is the fundamental object of self-consciousness, is identical to the object of this other knowledge. And this is to say that the "I" to which we ascribed our thoughts would not necessarily be identical to the heading, whatever it was, under which we had this other knowledge. But then this other heading would not as such express self-consciousness. In other words, the relevant knowledge would not be a necessary part of self-consciousness – which contradicts our assumption.

It should be possible, then, to investigate the essential content of our self-consciousness by reflecting on what is required for our nonobservational knowledge of our own thoughts. I will not undertake such a project here, however. Like Kant, I will be content to have shown something about "the I of pure apperception," without showing – what may be true – that knowledge of this I must include all that actual human self-consciousness essentially involves. At least with regard to this restricted idea of self-consciousness, we can endorse Kant's claim that the I is "a mere consciousness that accompanies every concept" (A346/B404). For to apply any concept is to exercise a power of reflection which is equally the source of this consciousness. In reaching this conclusion, we have completed our argument for the Kantian thesis.

[22] The "we" and "our" here are potentially substantive, for perhaps our human self-consciousness has conditions that the self-consciousness of an intuitive intellect would not.

V. TWO KINDS OF SELF-KNOWLEDGE

> The lower cognitive power is characterized by the passivity of the inner sense of sensations; the higher, by the *spontaneity* of apperception – that is, of pure consciousness of the activity that constitutes thinking… *[T]he 'I' of reflection* contains no manifold and is always the same in every judgment, because it is merely the formal element of consciousness. Inner experience, on the other hand, contains the matter of consciousness and a manifold of empirical inner intuition, *the 'I' of apprehension*…
>
> Kant, *Anthropology*, I, §7 (Ak. 7:140-141)

1. INTRODUCTION

The project of the four preceding chapters was to explain the meaning of the Kantian thesis and to say why it embodies an important truth about the minds of rational beings. Although there are many points where more could be said, I have at this point given the essentials of my interpretation and defense of Kant's thesis. My purpose in the two remaining chapters is not to continue along the same path, but to explore some paths that cross it: on the one hand, to show the bearing of the preceding discussion on some controversies in contemporary philosophy of mind; and on the other hand, to touch on some aspects of Kant's thinking about self-knowledge that I have so far neglected. I hope this will help both to reinforce the points made in the preceding chapters and to clarify their significance.

The present chapter is concerned with two topics. In the first place, any responsible discussion of Kant's views on self-knowledge must say something about the distinction he draws between the knowledge we have of ourselves through "pure apperception," on the one hand, and the knowledge we

have of ourselves through "inner sense," on the other.[1] In the foregoing discussion, I have been concerned only with the former kind of self-awareness, the kind manifested when a creature represents itself as *I*. If my interpretation is adequate, however, it should presumably help us to understand why Kant thought this sort of self-awareness had to be distinguished from another, different sort.

Secondly, although I have been stressing the importance of a certain kind of self-awareness, I have said little about the relationship between my topic and the kinds of problems about self-knowledge that are discussed by contemporary philosophers of mind. The principal sort of problem about self-knowledge that has interested recent philosophers is the problem of accounting for our "first-person authority" – our ability to speak, without recourse to self-observation and with some sort of special privilege, about our present mental states and attitudes. What bearing, if any, does my argument have on this problem?

At first glance, these two topics may look quite disparate. My aim, however, is to show that they are not. To make sense of our first-person authority, I will argue, we must begin by drawing the sort of distinction Kant drew: a distinction between two kinds of self-knowledge.

*

Kant famously held that we possess two fundamentally different kinds of self-knowledge: knowledge of ourselves through "inner sense" and knowledge of ourselves through "pure apperception." The former faculty, he claimed, gives us knowledge of our own sensations, knowledge of ourselves as passive beings, while the latter gives us knowledge of what we think and judge, knowledge of our own "spontaneity" (B67-8, A107, B132, B153, B278). Nevertheless, although he held that these two faculties are distinct, he also thought there was a relation of dependency between them: he argued that knowledge of ourselves through inner sense would be impossible in the absence of a capacity for pure apperception

---

[1] For the distinction, see *Critique of Pure Reason*, A107, B132 and B152-9. Kant sometimes calls inner sense "empirical apperception," distinguishing it from the "pure apperception" expressed by the representation "I think" (see A107, B132). At other times, however, he simply uses the term "apperception" to name the nonempirical form of self-

(at, e.g., B140). For, he claimed, the latter capacity, the one manifested in our ability to think of ourselves as "I" (A117n, B132, B157n), is the capacity that makes us "knowing beings" at all.

These Kantian claims are, of course, highly obscure, and the interpretative literature that has grown up around them is vast.[2] Without worrying about how to interpret Kant's actual views, however, we can think of the claims just mentioned as constituting the schema for a possible view about self-knowledge. Such a view would distinguish two kinds of knowledge of our own minds, an active and a passive. But although it would hold that these kinds of self-knowledge are distinguishable, it would also maintain that there is a relation of dependency between them – that one of the two kinds could not exist without the other.

I mention this schema for a view about self-knowledge not with the aim of doing further Kant exegesis, but because I think that seeing the possibility of such a view can help us to resolve a dispute in the contemporary literature on self-knowledge.[3] There have recently been a number of important attempts to account for our ability to know our own minds by connecting this ability with our capacity for some kind of agency. The most developed of these accounts is Richard Moran's recent *Authority and Estrangement* (2001), which argues that our ability to know our own current beliefs, desires, and other attitudes can on at least some occasions be understood as reflecting an ability to "make up our minds":

---

awareness (see, e.g., B153). From this point on, I will follow the latter practice: whenever I speak of "apperception," I will mean *pure* apperception, the kind distinct from inner sense.

[2] Recent books on this topic include Karl Ameriks, *Kant's Theory of Mind* (Second Edition, 2000); Andrew Brook, *Kant and the Mind* (1994); Pierre Keller, *Kant and the Demands of Self-Consciousness* (1998); Patricia Kitcher, *Kant's Transcendental Psychology* (1990); and C. Thomas Powell, *Kant's Theory of Self-Consciousness* (1990). This is only to mention book-length works in English whose main topic is Kant's theory of mind. To list all the discussions of these issues in articles and chapters would be a major research project in its own right.

[3] Following a widespread practice, I will use the term "self-knowledge" to refer to the awareness expressed in a subject's ability to speak in the first person, without self-observation and with apparent authority, about her own present mental states. Arguably, we also have this sort of knowledge of certain facts about, e.g., the current position of our own limbs and about what we are doing. Exactly what sorts of facts can be known in this special way, and what sort of privilege or "authority" belongs to such knowledge, are questions that will be discussed below. For the moment, I simply use "self-knowledge" as a label for whatever sort of knowledge it is that has interested philosophers under such headings as "privileged access," "first-person authority," and so on. Although little in my argument depends on the point, I take it for granted that our ability to speak authoritatively about our own present mental states does reflect *knowledge*. This has been denied, but I do not know of any convincing argument for denying it, and surely it deserves to be the default position.

an ability to know our minds by actively shaping their contents.[4] This sort of view has come under criticism, however, for its inability to account for our immediate, authoritative knowledge of other kinds of mental states that do not seem to be subject to our active control. I want to argue that this criticism rests on a false assumption about the uniformity of self-knowledge, an assumption that overlooks the possibility of a view like Kant's. And indeed, once we have this possibility in mind, I think its attractions will be obvious.

I begin, in §2, by giving a brief sketch of Moran's account and how an appeal to the notion of agency figures in it. Moran admits that there are kinds of authoritative self-knowledge to which his account does not apply, but he claims that the kind of self-knowledge it does describe is fundamental. Given his admission about the limited scope of his account, however, it is hard to see what grounds he can have for this claim of fundamentality. Indeed, critics of Moran have taken the fact that there are species of self-knowledge to which his account does not apply as itself reason for thinking that we must look for some more basic and embracing account. In §3, I consider this sort of criticism, focusing on the version of it presented by David Finkelstein in his recent *Expression and the Inner* (2003). It will emerge that this criticism depends on the assumption that a satisfactory account of self-knowledge should be fundamentally uniform, explaining all such knowledge in the same basic way. And we shall see that this assumption is accepted uncritically by many writers on self-knowledge, writers whose views are otherwise quite dissimilar.

The remainder of the chapter criticizes this assumption and lays the groundwork for a different kind of view. I begin, in §4, by considering Finkelstein's own attempt to give a uniform account of self-knowledge. Finkelstein, in the company of other recent authors, takes the phenomenon described by

---

[4] Other writers on self-knowledge who give a central role to the notion of agency include Burge 1996 and 1998 and Bilgrami 1998. There has also been considerable recent interest in another kind of connection between agency and self-awareness: interest, namely, in how our capacity to think of ourselves in the first person is connected with our capacity for embodied agency – our capacity, not to make up our minds, but actually to intervene bodily in the larger world in which we live. This is not my principal topic here, although I will touch on it at the end of §5. For recent discussion of these issues, see the essays collected in Naomi Eilan and Johannes Roessler, eds., *Agency and Self-Awareness* (2003).

Moran to be just one manifestation of the fact that self-ascriptions of mental states can *express* the mental states they ascribe. This view is worth considering both for its own interest and because the difficulties it faces will turn out to be representative: it will emerge that similar difficulties confront *any* attempt to deny the distinctness and fundamentality of Moran's topic. I show this by first arguing that Finkelstein's "expressivist" account of self-knowledge conflates two different kinds of expression, and that only an account which recognizes Moran's topic as distinct and fundamental can give a satisfactory account of how these kinds of expression fit together. I then argue, in §5, that *any* account of self-knowledge which does not recognize the distinctness and fundamentality of Moran's topic will be unable to account for the fact that self-ascriptions of mental states express knowledge, and specifically knowledge of oneself. The upshot is that we must recognize two kinds of self-knowledge, one exemplified in our knowledge of our own sensations, and the other exemplified in the kind of knowledge Moran investigates – our knowledge of our own thoughts and judgments. I conclude, in §6, with some remarks about the Kantian character of this outlook, and about the sense in which our knowledge of our own beliefs reflects a capacity for a kind of agency.

2. MORAN ON SELF-KNOWLEDGE AND AGENCY

There are familiar reasons to be puzzled by our capacity for self-knowledge. It is widely recognized that we can know our own present thoughts, attitudes, and sensations in a way that is fundamentally different from the way we know of the mental states of other persons. The precise character of this difference is a matter of dispute, but it is generally agreed that (1) a person is normally in a position knowledgeably to ascribe various kinds of mental states to himself without needing the sorts of *evidence* that would be required for his ascription of such states to another person, and that (2) self-ascriptions of these kinds of mental states are not normally liable to the same kinds of *error* that afflict ascriptions of such states to other people. The former feature of the relevant ascriptions is commonly referred to as their *immediacy*,

while the latter is one manifestation of their *authority* – their apparent entitlement to some sort of deference not accorded to third-person, evidence-based ascriptions of the same kinds of states. The *general problem of self-knowledge* is simply to explain how we can be in a position to speak about our own minds in such an immediate and authoritative manner, while still counting as speaking about the very same states that can be known to others only on the basis of observation or inference.

Moran's orientation, however, is not primarily toward this general problem but toward a particular subdivision of it. Early in his book, he remarks that

> [t]here are two basic categories of psychological state to which the ordinary assumption of 'privileged access' is meant to apply: occurrent states such as sensations and passing thoughts, and various standing attitudes of the person, such as beliefs, emotional attitudes, and intentions. (I will have comparatively little to say here about the case of sensations, which I believe raises issues for self-knowledge quite different from the case of attitudes of various kinds.) (2001, pp. 9-10)

Moran's view thus seems to be that the general problem of self-knowledge really comprises two different problems, one having to do with "attitudes" and the other with "sensations." Moreover, his decision to focus on our knowledge of our own attitudes, while leaving aside our knowledge of our own sensations, suggests that he regards these two problems as to a significant extent independent of one another.

What Moran finds striking about our knowledge of our own attitudes is this: we often seem to be able to know whether we hold them *by deliberating about the topics they concern*. If I want to know whether I believe that *p*, it seems that I can normally answer this question by considering whether there is *reason* to believe that *p* – whether there are persuasive grounds for thinking that *p* is true. And analogously, at least for so-called "motivated desires," it seems that I can normally determine whether I want X by considering whether there is reason to want X – whether there are persuasive grounds for thinking that X would be desirable. Likewise for intending to do something, for hoping for something, for fearing something, and so on: although there are certainly cases in which such attitudes prove recalcitrant in the face of our reflection on reasons for and against, still it is striking that often enough we can simply say what our attitude is by deliberating about the topic in question. Moran has popularized the term "transparency" as a label for this kind of relationship between a question about whether I hold a certain attitude and a

question about the object of that attitude.[5] Thus the question whether I believe that *p* is said to be "transparent" to the question whether *p* because it seems that I can settle the former question by settling the latter.

Such transparency can seem strange: how can I justifiably answer a question about what I believe by answering what seems to be a quite different question about a state of affairs independent of my belief? Moran's master thought is that the way to explain this seemingly paradoxical situation is to understand such self-knowledge as involving a kind of agency: I can know whether I believe that *p* by deliberating about whether *p* because my deliberation about *p* can constitute my *making up my mind* to believe that *p*. Thus, in a recent article summarizing his position, he explains his view as follows:

> What right have I to think that my reflection on the reasons in favor of P (which is one subject-matter) has anything to do with the question of what my actual *belief* about P is (which is quite a different subject-matter)? Without a reply to this challenge, I don't have any right to answer the question that asks what my belief [about, e.g., whether it will rain] is by reflection on the reasons in favor of an answer concerning the state of the weather. And then my thought at this point is: I *would* have a right to assume that my reflection on the reasons in favor of rain provided me with an answer to the question of what my belief about the rain is, if I could assume that *what* my belief here is was something determined by the conclusion of my reflection on those reasons. (Moran 2003, p. 405)

Moran's thought, in short, is that our ability to speak authoritatively about our own beliefs without looking for signs of belief in our behavior becomes intelligible if we suppose that my concluding that *p* on the basis of deliberation can *constitute* my coming to believe that *p*.[6] If I were entitled to assume that my deliberation determines what I believe, then my justification for thinking that I *did* believe that *p* could be simply whatever reasons I saw *for* believing that *p* – whatever grounds I saw for taking *p* to be true. And it would be appropriate to describe such knowledge as reflecting a kind of agency, for my deliberately-

---

[5] See Moran 2001, Chapter 2, §6. Moran cites Edgley 1969 as the original source of the term, and credits Evans 1982 with giving definitive expression to the point.

[6] This is not to suggest that my merely taking myself to have a certain belief must make it so. As Moran emphasizes, his view does not demand that a subject be incorrigible about her own attitudes, or that her beliefs about her own attitudes have a special authority no matter what their basis. What is important is that the question of what attitude I hold *can* often enough be settled by me on the basis of deliberation about whether *p*, and that – according to Moran – it is fundamental to the very possibility of thought about such attitudes that this should be so.

formed judgment that a certain proposition was true would be what made it the case that I believed the relevant proposition.

One kind of worry about this account would focus on the notion of agency at its center. It seems clear that we cannot choose to believe something in the same sense in which we can choose to do something: I cannot simply believe something because it seems to me a *desirable* thing to believe. But then in what sense, exactly, is my deliberating about what to believe an exercise of agency?[7] –For the moment, however, I want to table this worry. It seems to me plain that Moran is onto *something* important here; I propose to grant that he has given us at least the outline of an attractive account of how the transparency of questions about our own attitudes to questions about the objects of those attitudes might be explained. The question I want to raise concerns the sense in which these observations about how we know our own deliberated attitudes bear on the character of our knowledge of our own minds in general. Reflecting on this question will eventually bring us back to the question of what sort of agency is involved in our knowledge of our own deliberated attitudes.

A general point that Moran emphasizes about our knowledge of our own minds is that it is not normally "theoretical" in character: to know whether I believe that it will rain, whether I want to go for a walk, whether I am happy or hungry or suffering from toothache, I do not normally have to observe the kinds of signs in my behavior that I would have to observe to know such things about another person.[8] Now, one way of posing the general problem of self-knowledge is to ask: if our knowledge of our own minds is not founded on the kind of access we have to the mind of another person, what *is* it founded on?

---

[7] For questions about the kind of agency at issue in Moran's account, see O'Brien 2003 and Shoemaker 2003. On the impossibility of believing something simply because it would be desirable to believe it, see Williams 1973.

[8] Although I will follow Moran in contrasting a theoretical way of knowing about mental states with a non-theoretical one, I am not sure that the word "theoretical" is a happy one here. The intuitive point is that, if I am not in a position to know whether I believe that *p* by deliberating about whether *p*, then my epistemic position with respect to my belief that *p* is like my epistemic position with respect to the beliefs of another person. But is my knowledge of the beliefs of another person a kind of *theoretical* knowledge? Moran's own "Interpretation Theory and the First Person" (1994) gives reasons for answering "No" (at, e.g., pp. 168, 172). It seems preferable to say that although I know the minds of other persons *by observation*, I normally know my own mind *without observation*. For this distinction, see Anscombe 1963, which Moran himself cites in his discussion of the difference between theoretical and deliberative self-knowledge.

And Moran's invocation of the notion of agency can look like the beginning of an answer to this question: what makes non-observational knowledge of our own mental states possible, we might suppose, is that these states are in some sense brought into being by our deliberation. As we have seen, however, Moran does not mean to offer a general account of non-observational self-knowledge, for he explicitly sets aside our knowledge of our own sensations. And it is a good thing that he doesn't mean his account to apply to non-observational self-knowledge generally, for there seem plainly to be kinds of mental states of which our knowledge is both non-observational *and* non-deliberative: not just sensations but, for instance, appetites (i.e., brute, unreasoned desires for things of a certain kind) and what might be called "recalcitrant attitudes" (e.g., feelings of anger that I know to be unjustified but cannot overcome).

Indeed, given that we can very often speak authoritatively about our own attitudes without going through any conscious process of deliberation, it is not clear how much light Moran's account sheds even on our knowledge of the kinds of attitudes that are his principal topic. If I am asked such questions as whether I believe that Washington crossed the Delaware or whether I want to come along to the beach, I can often just answer straightaway, without any reflection on grounds for and against. In such cases, although I am surely expressing knowledge, the knowledge does not seem to reflect my now exercising the power to make up my mind. My mind, it is natural to say, is already made up. The question what I believe or desire is still, of course, transparent for me to a question about what is so or what is desirable, but the relevant convictions of fact or desirability are not being formed in the present, and so it is hard to see how an appeal to agency can help to explain my present knowledge of them. This sort of observation has led some philosophers to distinguish between the transparency of the question whether *to* believe that *p* to the question whether *p* and the transparency of the question whether I *already* believe that *p* to the question whether *p*. The explanation of the relevant transparency, according to these philosophers, is quite different in the two cases: in the former case, it is a matter of my being able to make up my mind by thinking about whether *p*; in the latter, a matter of my putting the question whether *p* to myself "as a

stimulus applied to [my]self for the empirical purpose of eliciting a response."[9] If this distinction is sound, however, then it seems that the application of Moran's agency-based account of self-knowledge is in fact quite limited.

Moran makes no secret of the fact that there are attitudes we can know ourselves to hold without reflecting on reasons or even in spite of our reflection on reasons.[10] Given this fact, however, and given the various kinds of non-attitudinal mental states of which we have immediate, authoritative knowledge although there seems in their case to be no question of reasons for and against, it is not clear how much light his account sheds on the general problem of self-knowledge. Moran clearly intends his book as a contribution to the question of how, in general, the striking features of self-knowledge can be accounted for. Indeed, he does not just claim to have accounted for our authoritative knowledge of some kinds of mental states; he claims that the kind of self-knowledge he describes is somehow basic. At one point, for instance, he characterizes himself as having

> argued the case for seeing the ability to avow one's belief as the fundamental form of self-knowledge, one that gives proper place to the immediacy of first-person awareness and the authority with which its claims are delivered. (2001, p. 150)

And in another passage he claims that the capacity to make "transparent" attitude-ascriptions is "what makes the difference between genuine first-person awareness and a purely theoretical or attributional knowledge of one's own states" (2001, p. 107). But in view of the apparent limitations on the scope of his account, it is hard to see how this sort of claim can be defended.

To justify his claim to have described the fundamental form of self-knowledge, Moran presumably owes us an account of why the cases of authoritative self-knowledge that he leaves aside are not as fundamental as the cases he does address, or a story about how the kind of self-knowledge he describes is a precondition of the kind he does not discuss. In fact, however, he says little more than that

---

[9] See Shah and Velleman, "Doxastic Deliberation" (2004), p. 13.

[10] See for instance his remarks on "unmotivated desires" which are not "an expression of one's reasons" at Moran 2001, p. 115, his remarks on unconquerable jealousy and fear at pp. 58 and 63, and his discussion of recalcitrant belief at pp. 131-132.

the kind of self-knowledge he describes *would* be an immediate and authoritative kind of knowledge, and that a subject must take herself to be capable of knowing at least *some* of her attitudes in this way, on pain of not possessing the relevant sorts of attitudes at all.[11] But even if we grant this, it is not clear why it should show that this sort of self-knowledge is more fundamental than other forms, or what its immediacy and authority has to do with the immediacy and authority with which we know various other kinds of mental states. This much, however, is clear: if there is truth in Moran's claim that the form of self-knowledge he describes is the fundamental form, it must be because "the fundamental form" does *not* mean "the form an account of which can serve as the model for an account of all species of immediate, authoritative self-knowledge." My own view is that there is a sense of "fundamental" on which Moran's claim is true. In §5, I will try to take some steps toward clarifying the relevant sense of "fundamental." First, though, I want to consider the kind of reaction to Moran's account that results if his claim about the centrality of deliberative self-knowledge is taken in the way that, I have suggested, it should not be taken. This kind of reaction can be found in Finkelstein's book.

---

[11] This is argued in the provocative but difficult fourth chapter of Moran's book. At the conclusion of this chapter, Moran remarks that

> [t]he problem with the idea of generalizing the theoretical stance toward mental phenomena is that a person cannot treat his mental goings-on as just so much data or evidence about his state of mind all the way down, and still be credited with a mental life (including beliefs, judgments, etc.) to treat as data in the first place. (2001, p. 150)

At least part of the argument for this seems to be that, even when I take a "theoretical stance" toward some aspect of my mental life – even when in a given case I look for evidence that I hold a certain attitude – still in doing so I necessarily presume that I can in general make up my mind on the basis of evidence. Hence, although I can treat the existence of a given attitude as a "mere datum," I cannot in general doubt my power to make up my mind on the basis of grounds without implying, absurdly, that I am incapable of even entertaining the question what my attitudes are (compare Moran 2001, p. 148). There is something compelling about this, but it is difficult to see how to get from here to the conclusion that a creature which did not treat its attitudes as open to deliberation could not be "credited with a mental life." Animals and infants presumably do not treat their attitudes as open to deliberation in the relevant sense, but they can hardly be denied to have mental lives, and when we are not concerned to defend some special philosophical doctrine, we are comfortable enough in ascribing at least some kinds of beliefs and desires to them. My own sense is that Moran would not want to deny the appropriateness of such ascriptions, and that his point should be read as restricted to beings capable of reflecting on their own mental states. Thus understood, I think Moran's claim gets at something important; but both his formulation of the point and his argument for it are less than perspicuous. I try to give a clearer account of the point in §5 below.

3. AN ASSUMPTION UNDERLYING CRITICISMS OF MORAN

Many of the questions I have raised about Moran's account are also raised by Finkelstein. Finkelstein quotes the passage in which Moran claims that our deliberative knowledge of what we believe is "the fundamental form of self-knowledge," and observes, as I have, that there seem plainly to be mental states concerning which we have immediate, authoritative self-knowledge, but to which Moran's style of explanation does not apply. Much of his discussion is devoted to giving examples of such states. But, as I have emphasized, the existence of such states only speaks against Moran's claim to have characterized the fundamental form of self-knowledge if "fundamental" means "capable of serving as the model for an account of all species of immediate, authoritative self-knowledge."

Finkelstein seems to assume that this is what "fundamental" must mean. He distinguishes two main kinds of reaction that a philosopher might have to the observation that the question whether I believe that *p* is normally transparent to the question whether *p*: the "Very Impressed" reaction, which concludes that this case provides a model that can explain "how we manage to speak with authority about a wide range of our own inner states and goings on"; and the "Not So Impressed" reaction, which holds that this case will not "generalize in such a way as to provide the key to understanding the self-ascription of very much besides belief" (2003, p. 155). Finkelstein's own reaction is more nuanced than either of these crude alternatives, of course: he thinks that Moran's model does generalize beyond the case of belief, but not so widely that we should be Very Impressed. As far as its actual content goes, this reaction seems judicious: we too have concluded that no straightforward generalization of Moran's account of deliberative self-knowledge can explain all forms of non-theoretical self-knowledge. But Finkelstein's use of the labels "Very Impressed" and "Not So Impressed" to name our alternatives seems tendentious. Why should we be very impressed by Moran's linking of self-knowledge with deliberative agency only if we think that Moran's model can account for non-theoretical self-knowledge in general? This labeling seems to reflect the idea that an account of self-knowledge is impressive only to the extent that it applies to all the various kinds of mental states about which we can speak with authority. We could call this the

Uniformity Assumption (UA), for it amounts to the assumption that a satisfactory account of our knowledge of our own minds should be fundamentally uniform, explaining all cases of immediate, authoritative knowledge of our own mental states in the same basic way.[12]

This assumption is not special to Finkelstein; it shapes the thinking of a broad spectrum of contemporary writers on self-knowledge. It is discernible, for instance, in Shaun Nichols and Stephen Stich's recent *Mindreading* (2003), which defends an account of self-knowledge quite different from Finkelstein's – one that appeals, not to the idea that first-person psychological ascriptions "express" the states they ascribe, but rather to hypothesized "monitoring mechanisms" which supply us with appropriate second-order beliefs about our own first-order mental states. Nichols and Stich do not address Moran's view specifically, but they do criticize what they call "ascent routine strategies" of accounting for self-knowledge: accounts that explain our ability to say whether we believe that *p* in terms of our having mastered an "ascent routine" which tells us to answer to the question whether we believe that *p* by determining whether *p*. Nichols and Stich's objection to this sort of strategy is that it is "clearly inadequate as a general theory of self-awareness" (2003, p. 194), since there are many kinds of self-knowledge which could not be arrived at by such an ascent routine. But like Finkelstein, Nichols and Stich never consider whether a theory of self-awareness *should* be general; they simply assume this without argument.

Nor are these authors atypical. The assumption that I have called (UA) is arguably present wherever philosophers are content to speak of *the* way in which we know our own minds. For this is to

---

[12] Moran is criticized on similar grounds by Dorit Bar-On and Douglas C. Long in their recent "Avowals and First-Person Privilege" (2001). Bar-On and Long call the sort of account that emphasizes the transparency of questions about what I believe and desire to a question about what is the case and what is desirable "the 'world-based' account," and they remark that

> [w]hereas introspectionist accounts have difficulty explaining the presumed truth of intentional avowals, the "world-based" account faces a complementary problem concerning non-intentional states. For, phenomenal avowals of pain, hunger, thirst, non-specific rage, for example, do not issue from an examination of the world outside the subject. (2001, p. 317)

imply that, insofar as there is something special about the epistemology of self-knowledge, this specialness admits of a more-or-less uniform characterization. And then, presumably, we should expect this special epistemology to be accounted for in a more-or-less uniform way. Anyone familiar with the literature on "first-person authority" will recognize that this sort of outlook is widespread.[13] To be sure, (UA) is rarely endorsed explicitly, but it is evident in the common tendency to argue about whether our immediate, authoritative knowledge of our own mental states is to be accounted for by appeal to some sort of quasi-perceptual faculty of "inner sense,"[14] by a reliable but non-perceptual tendency of our second-order beliefs about our own mental states to track our first-order mental states,[15] by a linguistic convention that simply grants a person's psychological self-ascriptions some sort of default authority,[16] or by the fact that such ascriptions normally "express" the states that they report.[17] What defenders of all these views have in common, despite their profound differences, is the conviction that *some* single basic strategy of explanation will account for all cases of immediate, authoritative self-knowledge. This is the conviction I want to question.

Here, as in Finkelstein's case, the underlying assumption is plainly that we should demand a uniform account of our authority in speaking about both deliberated attitudes and other kinds of mental states. Otherwise, why would it be a *problem* that the "world-based" account does not explain our authority in avowals of non-intentional states?

[13] Widespread but certainly not universal. The suggestion that we possess different kinds of self-knowledge which need to be accounted for differently appears, for instance, in Davidson 1984, Shoemaker 1990, Burge 1996, Bilgrami 1998, and Falvey 2000 – and, of course, in Moran. But these interventions have not altered the shape of the mainstream debate.

[14] This sort of view is more often discussed than defended. It is often attributed to some giant of modern philosophy, such as Descartes, Locke, or Kant. My own view is that, at least in Kant's case, although he does speak of a faculty of "inner sense," it is a travesty to attribute to him the view that the knowledge of our own mental states supplied by this faculty is quasi-perceptual. To defend this interpretative claim, however, would require a discussion of topics that fall outside the scope of this dissertation.

[15] Defenses of this sort of view include Armstrong 1968, Lewis 1972, and Lycan 1998. See Fricker 1998 for good discussion of choices to be made in formulating such a view.

[16] The most influential version of this approach is presented in a series of papers by Crispin Wright: see his 1987, 1991, and 1998.

[17] Early gestures toward this approach can be found in Ryle 1949 (e.g., at p. 102) and Shoemaker 1963, Chapter 6. (Shoemaker has subsequently adopted a more complex position, which shares Moran's emphasis on transparency. See his 1988 and 1990.) More recent and systematic defenses of the expressivist approach include Bar-On and Long 2001 and Finkelstein 2003, which I discuss shortly.

4. FINKELSTEIN ON SELF-KNOWLEDGE AND EXPRESSION

It will be useful to begin by considering the problems faced by Finkelstein's own attempt to give a uniform account of self-knowledge. This attempt is of interest, in the first place, because it is the most recent and sophisticated defense of the expressivist approach to self-knowledge, which represents a major strand in the literature on self-knowledge and which has points of genuine plausibility. But the expressivist attempt to give a uniform account of self-knowledge is also of interest because the difficulties it faces are representative. I shall argue that expressivism founders because its assimilation of the sort of expression involved in transparent self-ascriptions of deliberated attitudes to the sort involved in self-ascriptions of phenomenal states like pain obscures a fundamental difference in kinds of expression, one associated with a fundamental difference in kinds of self-knowledge. And it will emerge in the next section that similar difficulties confront *any* account of self-knowledge that does not recognize a distinct kind of knowledge associated with deliberated attitudes. Our discussion of expressivism will thus lay the groundwork for a general case against (UA).

The strategy of accounting for first-person authority by appeal to the notion of expression is not new. Expressivist accounts generally take their inspiration from remarks made by the later Wittgenstein, remarks like the following:

> [H]ow does a human being learn the meaning of the names of sensations? of the word "pain" for example. Here is one possibility: words are connected with the primitive, the natural, expressions of the sensation and used in their place. A child has hurt himself and he cries; and then adults talk to him and teach him exclamations and, later, sentences. They teach the child new pain-behavior.
>
> "So you are saying that the word 'pain' really means crying?" —On the contrary: the verbal expression of pain replaces crying and does not describe it. (Wittgenstein 1953, §244)

The state of pain clearly has certain natural expressions in human behavior: there are certain characteristic signs that human beings suffering from pain are naturally disposed to exhibit, signs whose exhibition is not in the first instance a matter of reasoned choice but of instinctive response. Crying is presumably one such natural expression. Passages like the one just quoted have been read as suggesting that our

knowledge that we are in pain (if it should be called "knowledge") is to be understood as a matter of our having learned to express pain in another way, namely by uttering sentences such as "I'm in pain." The attraction of this suggestion is that it avoids questions about our grounds for supposing that we are in pain – for to be an expression is to be, not a report made on the basis of grounds, but (at least in the fundamental case) a manifestation of an unreflective disposition. Such a view thus promises to account for our "privileged access" to our own pains without ascribing to us any weird epistemic powers. Encouraged by this promise, "expressivists" about first-person authority seek to generalize this model, understanding all authoritative self-ascriptions of mental states as expressions, not reports – and hence as groundless without being mysterious.[18]

Finkelstein's distinctive contributions to this approach come in two areas. In the first place, he stresses that the question whether mental state self-ascriptions are truth-evaluable is independent of the question whether they are normally used to make reports (see Finkelstein 2003, Chapter 4). It is simply a mistake, according to Finkelstein, to suppose that a non-reporting use of a sentence – one not made on

---

[18] As I am using the term, any philosopher who takes some univocal concept of expression to hold the key to an account of all kinds of first-person authority is an expressivist. This differs from Finkelstein's usage, on which philosophers only count as expressivists if they take the idea that avowals of mental states are expressive to exclude their being truth-evaluable (see Finkelstein 2003, Chapter 4, §3). It is worth remarking that, although expressivist approaches are inspired by Wittgenstein, it is not clear that Wittgenstein himself would have countenanced the application of the relevant notion of expression to self-ascriptions of thought and belief. In another passage from the *Philosophical Investigations*, he writes:

> Misleading parallel: the expression of pain is a cry – the expression of thought, a proposition. As if the purpose of the proposition were to convey to one person how it is with another: only, so to speak, in his thinking part and not in his stomach. (1953, §317)

The parallel that Wittgenstein questions here is between the *primary* expressions of pain and of thought, namely crying on the one hand and asserting a proposition on the other. But I think he would have regarded it as equally misleading to treat the *secondary* expressions of pain and of thought – saying "I'm in pain" and saying "I believe that *p*" – as parallel. For the point of his remark, seemingly, is that the primary expressions of pain and thought are not expressive in the same sense, since the purpose of "expressing" a thought is not, like the purpose of crying out in pain, primarily "to convey to one person how it is with another." If the primary expressions of pain and thought differ in this way, and if the sense in which the secondary expressions are tied to the states they concern is to be understood in terms of their connection with their respective primary expressions, then the contrast Wittgenstein points to should carry over to their secondary expressions. And indeed, it is one of the main themes of his discussion of belief in *Philosophical Investigations*, Part II, §x that self-ascriptions of belief are not primarily ways of conveying how things stand "in our thinking parts" either.

In support of his attribution of a generalized expressivism to Wittgenstein, Finkelstein cites the "Plan for the treatment of psychological concepts," which appears in different versions in several places in Wittgenstein's late writings on the philosophy of psychology (Finkelstein 2003, p. 97; cf. Wittgenstein 1964, §§472, 488; 1980, Vol. II, §§63, 148).

the basis of evidence – cannot put forward a truth-evaluable proposition. Avowals such as "I'm in pain," self-ascriptions of intention like "I intend to pursue a career in advertising," and so on, plainly put forward truth-evaluable propositions, but normally they do not put them forward as reports made on the basis of evidence. And expressivists must insist on this possibility in answering stock objections to their position. For if, as some earlier expressivists maintained, the expressive use of "I'm in pain" is just the verbal equivalent of a moan, then it is hard to see how there can be semantical relations between these uses and uses of those same sentences in contexts (for instance in the antecedents of conditionals) where they clearly must be capable of truth or falsity.[19] If denying that psychological self-ascriptions are reporting need not entail denying that they are truth-evaluable assertions, however, then this sort of objection is no threat.

Secondly, Finkelstein gives a novel account of how the expressivist proposal can be applied not only to avowals of states like pain, but also to cases such as saying what one thinks or what one means. This is important because, on the face of it, the project of extending the expressivist strategy beyond the case of simple sensations and appetites looks problematic. We can certainly speak of "expressing" our thoughts and attitudes in self-ascriptions, but the mere fact that this *word* seems apt here obviously does not show that self-ascriptions of thought, belief, and so on can be understood as expressive *in the sense identified by expressivists*. And in fact it is not at all clear how the expressivist model can be extended to such cases. After all, it hardly seems plausible that there are *natural* expressive behaviors associated with all the various thoughts a person might have and attitudes she might hold: acquiring the language needed to avow complex thought and attitudes seems part and parcel with coming to be capable of them.

Before we can judge whether the expressivist strategy can be extended to such cases, we need to

---

But it is worth noting that, although this plan enumerates several kinds of psychological concepts whose first-person present-tense use is "akin to an expression," it makes no mention of thought or belief.

[19] The most well-known version of this objection is due to Peter Geach, in his "Ascriptivism" (1960) and "Assertion" (1965). For further discussion see Wright 1998, §IX. Another defense of expressivism that responds to such objections by emphasizing the distinction between the question of whether a sentence is truth-evaluable and the question whether it is normally used to make reports is Bar-On and Long 2001.

get clearer about what the relevant sense of "expressive" is. Plainly, if the desired epistemological payoff is to be achieved, expressive uses of language will need to contrast with reporting ones. Merely to say that self-ascriptions of thought or belief are not reports, however, is not to make their authority intelligible, for it is just to reiterate the observation that was supposed to present a problem, namely that first-person present-tense ascriptions of various kinds of mental states seem not to rest on evidence. But then in what sense are such self-ascriptions expressive? Under what single concept of expression do these various cases fall?

By way of an answer, Finkelstein says only that self-ascriptions of thoughts and attitudes can be expressive inasmuch as

> our psychological self-ascriptions contextualize that which they ascribe and so aren't mere ascriptions. When someone ascribes, e.g., an expectation to himself, the ascription is part of the situation in which the expectation participates and from which it, as it were, draws its life. (2003, p. 111)

The general idea, which Finkelstein conveys mainly by means of examples, seems to be that the question of exactly what I believe, desire, expect, etc. is a matter of what attitude ascriptions best fit the overall pattern of my behavior in its context, and that at least some *self*-ascriptions should count not merely as comments on this pattern but constituents of it. The aim, in short, is to identify a broad counterpart to the idea that paradigmatic self-ascriptions of pain "give vent to" pain, and thus relate to the presence of pain in a way that differs from the way in which an evidence-based report on pain relates to the state on which it reports. Thus, according to Finkelstein, when a person says what she means, or explains how she reacted to a particular event by saying what she expected, or makes a psychological self-ascription in any number of other ordinary situations, her statements are to be understood as expressive in that they are parts of the complex syndrome in which her being in the relevant mental state consists.

Part of what makes this analogy between avowals of pain and self-ascriptions of attitudes attractive is that we do find it natural to speak of "expressing" attitudes like belief in speech. I can express my belief that *p* in one way simply by saying "*p*"; and in learning to say "I believe that *p*," I learn to express my belief that *p* in a more sophisticated way, one that involves making a self-ascription. This

advance seems analogous to the advance I make in learning to express pain not just by crying but by saying "I'm in pain." But from the fact that we find it natural to speak of "expression" in both of these cases, and the fact that we can in each case distinguish a primary (non-self-ascriptive) and a secondary (self-ascriptive) expression, it does not follow that the same *kind* of expression is in question in the two cases. And on further consideration, it should be clear that the two cases do not involve the same kind of expression. For the relation between the *primary* expression of belief and the state it expresses differs from the relation between crying and pain. Crying expresses pain in that it is a characteristic, natural response to being in this state. The disposition to cry when in pain precedes the development of the capacity to reflect on whether one is in pain, and this is what makes it intelligible that a creature whose disposition to say "I'm in pain" is brought into connection with its disposition to cry should be able to say whether it is in pain without reflection on grounds. But the ability to express one's beliefs in articulate speech is not in this sense a natural, unreflective ability. I do not have the ability to "express" my belief in speech – even in the primary, non-self-ascriptive way – until I have the ability to reflect on what I am saying and my grounds for saying it. And this implies that an account of my ability to say what I believe cannot take the same shape as an account of my ability to say whether I am in pain.

We must, then, distinguish two species of expression, one that can exist in a "prereflective" form and one that cannot. To draw such a distinction is not to deny that these different species of expression belong to a common genus. If two different species fall under this genus, however, and if I am right that the expressivist strategy of accounting for first-person authority is plausible for only one of the two species, then the existence of a common genus provides no support for a generalization of the expressivist story about pain to attitudinal states. Moreover, the failure to distinguish between the different species prevents the question of the nature of this common genus from emerging in a clear way. A satisfactory account of how behavior can express mental states must explain what kind of thing "expression" is such that these two species are both instances of it, and must make it intelligible how a reflective, verbal expression can "replace" a prereflective, nonverbal one. Finkelstein's undiscriminating use of the notion

of expression prevents him from seeing these questions.

It will sharpen our sense of the problem here if we reflect a bit on Finkelstein's other theme, that self-ascriptions of mental states can be truth-evaluable without being reports. It is certainly true that a sentence capable of bearing a truth-value can occur in contexts where it is not used to make a report (e.g., when it is used to state a conjecture or – a very different kind of "use" – when it is sung in a song); but intuitively, we still distinguish between comprehending uses in such contexts and uncomprehending uses. Suppose, for instance, that I train a parrot to cry "I'm in pain!" just when it is disposed to whatever behaviors are the natural expressions of pain in a parrot. Then it will be disposed to utter a sentence with a truth-evaluable content on just those occasions when the sentence is, in fact, true, and surely it is at least as plausible to say in this case as in the case of a person that we are not dealing with a reporting use of the relevant sentence. But we also want to say something else, namely that *the parrot doesn't know what it is saying* – and presumably this might also be true at a certain point in a human being's initiation into the practice of using the sentence "I'm in pain" to express pain.

What is missing in the parrot's case, it seems, is comprehension of the sentence it utters. But if the question of the *truth-evaluability* of an utterance of the sentence "I'm in pain" is distinct from the question of the *comprehendingness* of such an utterance, then merely to say "psychological self-ascriptions are expressive replacements for natural behaviors, which are, however, truth-evaluable sentences in a public language" is not yet to account for their comprehending use – for a person might meet these conditions and yet still be merely "parroting" the relevant ascriptions. And surely avowals of mental states are comprehending uses, even if they are in some sense not reports. If this is right, however, then there is a major lacuna in Finkelstein's account even of those self-ascriptions to which his view most plausibly applies, namely avowals of simple sensations and appetites: he has not given an account of what the subject's *comprehension* of such avowals consists in, and how it is compatible with her non-evidential use of them. This difficulty is slurred over by Finkelstein's undiscriminating use of the word "expression." For when *a self-ascriptive utterance* is said to "express" a mental state, it is overwhelmingly natural to hear this as

meaning that the utterance *asserts* the existence of the relevant mental state – and assertion requires comprehension. But all that Finkelstein's account actually provides for is expression in a weaker sense, the sense in which crying expresses pain. The fact that a creature has learned to use the truth-evaluable sentence "I am in pain" to express its pain in this latter sense does not yet show that its utterance expresses its pain in the former sense.[20]

What has emerged, in short, is that (1) Finkelstein lacks a clear account of the sense of "expression" that supposedly underlies all the various kinds of psychological self-ascriptions to which he wants his account to apply, and that (2) even in the cases where the Wittgenstein-inspired notion of expression has a relatively clear application, invocation of this notion cannot by itself constitute an account of the comprehending use of sentences that self-ascribe psychological states. Reflection on the latter point has led us to distinguish two senses of expression, which we might call "the assertion sense" (expression_A) and "the manifestation sense" (expression_M). A pre-linguistic infant's cry of pain is an expression only in the latter sense, whereas a paradigmatic *avowal* of pain is an expression of pain in both senses. A self-ascription of belief that meets Moran's Transparency Condition is clearly an expression in the former sense, but not so clearly an expression in the latter. An adequate account of our authority in speaking about our own mental states must, it seems, explain both kinds of expression, and how they fit together. In the next section, I argue that, once we have rejected (UA), we can see Moran and Finkelstein as providing materials for just such an account.

---

[20] Could Finkelstein avoid this difficulty by simply denying that "parroted" utterances are truth-evaluable? Could he just insist on calling utterances "truth-evaluable" only if they are intelligible as genuine assertions, not mere parrotings? I think this would only postpone the difficulty. For if Finkelstein wishes to insist that parroted utterances do not "express truth-evaluable claims," he owes us an account of what *is* required for this special sort of expression. Reflecting on the possibility of parroting is useful precisely because it shows that the mere presence of a reliable disposition to produce a certain utterance-type in response to a certain mental state does not ensure that tokenings of the relevant utterance-type will be expressive *in the sense that implies comprehension*. But surely it is *this* sort of expression that we need to understand if we are to understand our ability to speak authoritatively about our own mental states. Merely noting that linguistic utterances can come to replace natural expressive behaviors does not address this need.

5. JUDGMENT, REASONS, AND SELF-KNOWLEDGE

Our complaint about Moran was that he gives us no clear explanation of the scope of his account or why we should regard it as identifying the fundamental form of self-knowledge. Our complaint about Finkelstein was that he does not really clarify how Wittgenstein's suggestion about pain can be applied to thoughts and attitudes, and that in any case this suggestion seems as though it can be at best *part* of an account of the authority of avowals, even avowals of simple sensations and appetites. I hope that these problems are beginning to look complementary, and that it is beginning to seem plausible that an account of self-knowledge will need to draw on more than one kind of resource. I have not so far argued that no uniform account of self-knowledge can succeed; I have only pointed out some difficulties for Finkelstein's attempt to provide such an account. I now want to argue, however, that Finkelstein's account falls into these difficulties through its failure to recognize a distinct kind of knowledge associated with deliberated attitudes. Once we see the source of these difficulties, moreover, it will be clear that *any* satisfactory account of self-knowledge would have to recognize this kind of knowledge as distinct.

The basic problem with Finkelstein's account is that it lacks a story about what constitutes utterances of sentences like "I'm in pain" or "I'm hungry" as assertions, not just expressions$_M$ but expressions$_A$ of the relevant states. Expressivism seems right to this extent: it is plausible that learning to speak about simple sensations and appetites involves learning to use such sentences to express$_M$ conditions which one was formerly in a position to express$_M$ only through various instinctive behaviors. In order for such verbal expressions$_M$ to come to express$_A$ the existence of the relevant states, however, they must be brought into connection with a general understanding of language, an understanding that will involve recognizing semantically significant articulation in the relevant sentences – in particular, their articulation into a first-personal subject and a psychological predicate. This, in effect, is just the moral of our discussion of pain-avowing parrots in the last section.

What is involved in recognizing such articulation is a large question, but I suppose that the following point, at least, is uncontroversial: a speaker does not understand a sentence so long as she just

has an isolated disposition to utter it in certain conditions (e.g., when she is exposed to certain sensory stimuli, or when she is herself in a certain internal state). Understanding dawns only when she comes to recognize truth-relationships between this sentence and other sentences that she, or another person, could utter. She must, in short, come to appreciate what she is doing when she utters the relevant sentence as taking a stand on what is true, a stand that bears relations of implication and exclusion to stands on other questions, a stand that other speakers could query or contradict.

Now, it is obviously beyond the scope of this chapter to give any real account of what is involved in coming to appreciate that one is taking a stand on what is true. One preliminary observation that seems helpful, however, is that full-fledged understanding of the significance of assertoric speech will involve being able to respond to a certain kind of "Why?"-question. In general, a competent speaker who claims that *p* can be asked why she thinks that *p* is true, and it will be a criterion of her understanding this question, and thus of her understanding what she has said, that she be able to answer by producing grounds of a certain kind: grounds that bear on the truth of the claim that *p*. She must, as it is sometimes put, be able to "play the game of giving and asking for reasons," where "reasons" here means considerations bearing on the truth of the claims she has made.[21] And now the thing to notice is that a subject capable of uttering sentences in a way that reflects an appreciation of this "Why?"-question – a subject capable of using language assertorically – will necessarily be a subject who can, in general, ascribe beliefs to herself in a way that conforms to Moran's Transparency Condition. For a subject who can say that *p* just when she takes there to be adequate grounds for supposing that *p* is true is a subject whose speech already expresses$_A$ her beliefs: when she says that *p*, she will be saying something she takes to be true, and since to take something to be true just is to believe it, she will also be entitled to say "I believe that *p*."[22]

---

[21] I borrow the phrase from Brandom 1994. Brandom credits the idea to Wilfrid Sellars.

[22] I anticipate the objection that a person might master the game of giving and asking for reasons even though her answers in this game did not bear the right kind of relation to her beliefs. For mightn't a person be able to answer the

Moreover, although I have framed this as a point about the connection between the capacity to make transparent self-ascriptions of belief and the capacity for assertoric *speech*, it can equally be put as a point about the intellectual power that philosophers have traditionally called *judgment*: the power to make up one's mind on the basis of a deliberation. For the reasons I have given for holding that a creature capable of assertoric speech must be capable of *responding verbally* to the truth-oriented "Why?"-question are equally grounds for holding that any creature capable of judging must at least be capable of *thinking* about this question. The crux of my argument is simply that (1) a creature understands itself as taking a stand on what is true only if it understands the relevance of a certain kind of "Why?"-question to the stands it takes, and that (2) a creature capable of reflecting on this question will necessarily be capable of knowing its own beliefs by making it up its mind. I have focused on the capacity to express judgments in assertoric speech as a way of making these points tangible, but nothing in my argument depends on the assumption that the capacity to think requires the capacity for assertoric speech, or mastery of a public language. If somebody thinks a languageless creature might possess the kind of understanding I am describing – an understanding that involves a capacity to reflect explicitly on grounds for and against a given proposition – then that person should simply read the foregoing argument as demonstrating a connection between this kind of understanding and the capacity to know one's own beliefs. Having noted this point, I will continue, for vividness, to focus on the capacity to express judgments in assertoric speech; but the power of judgment itself is my real topic throughout.

If the foregoing account of the preconditions of the ability to self-ascribe beliefs is correct, then a

---

question whether *p* by giving grounds for or against and yet act in a way that belied any belief in the conclusions she reached on this basis? In that case, she presumably would *not* be entitled to say "I believe that *p*" just because her consideration of grounds led her to say that *p*. –The conceivability of such alienation, however, only brings out that mastering the game of giving and asking for reasons cannot just be a matter of coming to exhibit a certain order in one's vocalizations that is unconnected with the rest of one's behavior. A person who answers the question whether *p* with considerations that bear no relation to her action is not giving *her* grounds for taking *p* to be true or false. This may be a matter of deliberate insincerity, in which case it does not make problems for my claim as a point about subjects who are sincere. If it is not a matter of insincerity, but is an aberration, then it will count as some sort of breakdown of reflexive rationality – perhaps, if it is motivated, a case of wishful thinking or self-deception. But if it is not a matter of insincerity, and also is not intelligible as a local breakdown in a normal capacity to express beliefs in speech, then the subject in

subject's ability to ascribe beliefs to herself is not to be understood as a matter of getting her speech-dispositions keyed to the presence of isolable, naturally-existing dispositions, in the way that the disposition to say "I'm in pain" gets keyed to the naturally-existing disposition to cry out in pain. It is to be understood, rather, as a matter of her speech-behavior's coming in general to have the kind of connection with reasons that makes (certain of) her utterances intelligible as assertions: she must say things in a way that reflects an understanding of her utterances as claims subject to the kind of "Why?"-question just indicated. Once her speech acquires this sort of intelligibility, she acquires the entitlement to accompany her sincere assertions with "I believe," so to speak, for free. And although it seems plausible that our ability to say whether we are hungry or in pain *does* involve getting our speech-dispositions keyed to the presence of isolable, naturally-existing dispositions, still the relevant speech-dispositions only become intelligible as expressions$_A$ of the relevant states insofar as they become part of a larger ability to make, support and criticize assertions; for only insofar as we can do this do we count as capable of using language to make claims about what is true.[23]

If this is right, then we are in a position to say why the kind of self-knowledge that Moran characterizes is fundamental. It is fundamental because the ability to say what one believes in the way that Moran specifies is intimately connected with the capacity to *judge* – the capacity that, according to

---

question will not count as capable of making assertions at all. At best, it will be as if an alien voice is speaking from her mouth.

[23] A defense of the expressivist approach to sensation and appetite would therefore have to include a story about how comprehension of what is involved in self-ascribing sensory and appetitive states can be compatible with the continued disposition to make such self-ascriptions without observation. After all, if I understand what I am saying in saying "I'm in pain" or "I'm hungry," then I understand that I am making a claim with certain truth-conditions, and there is plainly a difference between these truth-conditions actually being fulfilled and my merely asserting that they are. We need, then, to understand how it is possible for comprehension of what I am saying not to undermine the disposition groundlessly to make such claims about myself. And the problem here is not just to explain how it is *psychologically* possible for such a disposition to persist, but how it can be *rational* for a subject to persist in it. For surely we do not think that there is any irrationality in avowing one's pain or one's hunger without looking for grounds. Indeed, it seems that a person who did look for such grounds would not understand what pain and hunger are.

I am inclined to think that a sophisticated expressivist account *can* meet this demand. The thought would be that understanding what (e.g.) pain is involves understanding that my spontaneous inclination to aver, or think, "I'm in pain" is *itself* reason to think I am in pain – provided that I have been trained in the way the expressivist account specifies. But my aim, in any case, is not to defend expressivism. My aim is to show that, whatever the true story about our knowledge of our own sensations and appetites, it will presuppose a distinct story about our knowledge of our own deliberated attitudes.

philosophical tradition, is the mark of a rational understanding.[24] For, as we have already noted, the capacity to judge is the capacity to make up one's mind about what is the case on the basis of a reflection on grounds for and against, and the capacity to make utterances in a way that reflects a grasp of the truth-oriented "Why?"-question is the outward mark of this capacity. Moreover, the capacity to answer this question is intimately connected with the capacity to make transparent self-ascriptions of belief. For, as we have seen, if a creature can express judgments in speech at all, then it will necessarily be entitled to accompany the judgments thus expressed with "I believe." And equally, if it cannot express judgments in speech, then none of its apparently self-ascriptive utterances will count as genuine assertions about its own states. Its speech-behavior will be, at best, like that of our pain-avowing parrot, which utters true self-ascriptive sentences but does not understand them. The kind of agency a subject exercises in thinking about what to believe – the kind of agency that is Moran's topic – is thus a kind of agency of which a subject must be capable if she is to be capable of expressing judgments at all.

Indeed, there is an even closer connection between being able to self-ascribe beliefs and being able to self-ascribe other kinds of mental states. Any explicit self-ascription of a mental state will involve a form of the first person: this is what makes it a *self*-ascription. But it should be clear on reflection that a creature does not understand the first person if it does not understand its use in present-tense self-ascriptions of belief. For to understand that an expression is a form of the first person is to understand that, when that expression is accompanied with a predicate, the subject to which the predicate is being applied is *the very subject who is claiming that this predicate applies to this subject*.[25] But surely a creature

---

[24] Thus Kant holds that the faculty of "understanding" (by which he means, the faculty of *rational* understanding, as distinguished from the kind of faculty of representation possessed by mere brutes) is "the faculty of judging" (A69/B94).

[25] This, in effect, is just a language-oriented way of expressing the traditional thought that the referent of the expression "I" is *the thinker*. The point of the traditional thought is this: to understand that the subject of a certain predication is *oneself* is to understand that the subject of the relevant predication is *the very subject who is thinking that this predicate applies to this subject*. Compare Evans' remark that

> the essence of 'I' is *self*-reference. This means that 'I'-thoughts are thoughts in which a subject of thought and action is thinking about him*self*–i.e. about a *subject* of thought and action… I do not merely have knowledge of myself, as I might have knowledge of a place: I have knowledge of myself

understands this only if it understands that, when it is entitled to say that *p*, it is also entitled to ascribe the belief that *p* to itself. For to understand oneself as making a certain claim involves understanding that one is representing oneself as *believing* that claim, and recognizing the entitlement to accompany one's sincere assertoric utterances with "I believe" is obviously a crucial step in coming to understand this.[26] Coming to understand the entitlement captured in Moran's Transparency Condition is thus an essential part of coming to understand the first person at all. Lacking this understanding, a creature can of course make utterances involving the expression "I"; but until it knows how to use this expression in transparent belief-ascriptions, and recognizes the relationship between the use of "I" in this context and its use in other kinds of self-ascriptions, it does not understand the significance of this expression, and thus does not understand any utterance of which it is a part.[27]

> *as* someone who has knowledge and makes judgments, including those judgments I make about myself. (1982, p. 207; and see pp. 258-261 for further interesting discussion)

Note Evans' suggestion that understanding the first person involves understanding oneself as the subject of thought *and action*. I think this is exactly right: although I have only been discussing knowledge of one's own beliefs and its connection with the ability to answer a certain kind of "Why?"-question, I think that closely similar points could be made about knowledge of one's own actions and its connection with the ability to answer a certain (different) kind of "Why?"-question. I say more about this shortly.

[26] If this does not seem obvious, try imagining a subject who has learned to use the expression "I" in various sorts of contexts, but who has not yet learned that she can automatically accompany her sincere assertoric utterances with "I believe." What should we say that "I" means in such a subject's idiolect? This much seems clear: her "I" does not yet express the idea of the thinking subject, the subject whose beliefs are expressed in sincere assertoric speech. For it would be perfectly coherent for her to sincerely assert that *p* and yet wonder "Do I believe that *p*?"

[27] I do not, of course, mean to suggest that recognizing the entitlement to utter the words "I believe that *p*" when one is entitled to say that *p* is by itself a *sufficient* condition for understanding the first person, or its use in belief-ascriptions. Understanding the first person arguably involves understanding its use not just in belief-ascriptions but also (at least) in ascriptions of actions and attitudes toward actions. Moreover, understanding the use of the first person in the context of a *belief*-ascription requires possession of the general concept of belief. As Evans puts it:

> [W]henever you are in a position to assert that *p*, you are *ipso facto* in a position to assert 'I believe that *p*'. But it seems pretty clear that mastery of this procedure cannot constitute a full understanding of the content of the judgment 'I believe that *p*'. Understanding of the content of the judgment must involve possession of the psychological concept 'ξ believes that *p*', which the subject must conceive as capable of being instantiated otherwise than by himself… Without this background, we might say, we secure no genuine 'I think' ('think that *p*') to accompany his thought ('*p*'): the 'I think' which accompanies all his thoughts is purely formal. (1982, pp. 225-226)

I do not mean to downplay these further requirements on understanding the first person and its use in psychological talk. My point is only that it is a *necessary* condition for understanding the first person that one have learned to use it in combination with a psychological verb to make transparent self-ascriptions of belief. (The expression used to signify the first person need not, of course, be the English word "I," and the verb in question need not be "believe." As Evans indicates, English itself has another verb that is used for this purpose, namely "think." What is necessary is that a subject have mastered a language in which *some* expression fills each of these roles.)

This is not to say that a creature can only make genuine assertions if it can self-ascribe beliefs in accordance with Moran's Transparency Condition. It is merely to say that (1) if a creature's language-use possesses the kind of intelligibility that constitutes some of its utterances as assertions, then it will necessarily be *entitled* to self-ascribe beliefs in accordance with Moran's Transparency Condition (even if it does not grasp this entitlement); and that (2) a creature must grasp this entitlement if it is to make comprehending use of the first person. It is compatible with these points that a creature might learn to make full-fledged, comprehending assertions without learning to make assertions in the first person at all. (Actually I doubt this is possible, but showing its impossibility would require further argument.) But even if a creature can learn to make assertions without learning to make self-ascriptive assertions, we can still say this: such a creature will already have all the skills and information it needs to make immediate, authoritative self-ascriptions of belief *except* mastery of the use of the expression "I believe" in accordance with Moran's Transparency Rule. For its speech will already have the kind of intelligibility that makes its assertoric utterances normally (barring, e.g., special efforts on its part to dissimulate) expressions$_A$ of its beliefs.

Moreover, although I have been focusing on what would need to be added to Finkelstein's account of self-knowledge to remedy its deficiencies, the points I have emphasized must surely be acknowledged by any satisfactory account of self-knowledge. For any account of self-knowledge must be intelligible as an account of *knowledge* – which it will only be if the utterances that express it are intelligible as genuine assertions – and in particular of *self*-knowledge – which it will only be if the relevant utterances are intelligible as involving a comprehending use of the first person.[28] Hence, if I am right that immediate, authoritative knowledge of one's own beliefs is a necessary concomitant of the capacities for assertion and for the comprehending use of the first person, and requires no special account where these capacities are present, it follows that the story of how we know our own beliefs will form a fundamental

and independent department in any account of self-knowledge, whatever explanation it goes on to give of our capacity to know when we are hungry or in pain. The expressivist story about the basis of these latter species of knowledge seems to me attractive, and its attractions are only increased if it is cut free from the demand that it account for our knowledge of our deliberated attitudes. I do not insist on the correctness of the expressivist story about our knowledge of our own sensations, however. What is important for my purposes is just that the kind of knowledge we have of our own *beliefs* should be recognized as distinct and in an important sense fundamental. It is fundamental, not because we find in it a model that can be generalized to account for all kinds of self-*knowledge*, but because this kind of self-knowledge is a precondition of self-*consciousness* – of the comprehending use of the first person. If this is right, then what Moran has given us is not an account of the authority of avowals in general, but a characterization of the framework into which any account of the authority of other kinds of avowals must fit.

I have been focusing in this section on what is at stake in our capacity to make transparent self-ascriptions of belief. Let me conclude with a few remarks about other deliberated attitudes. There are, I think, two major kinds to consider: cognitive attitudes, which involve the holding-true of some proposition, and conative attitudes, which involve regarding some action as to-be-performed or some object as to-be-attained. (I leave it undecided whether there are attitudes that have both aspects.)

For cognitive attitudes, Moran's kind of self-knowledge is relevant precisely because these attitudes involve belief, and are open to deliberation in virtue of the fact that the relevant beliefs are open to deliberation. Thus, if it is true that propositional anger that *p* must involve the belief that the relevant fact constitutes an insult to me, or that propositional fear that *p* must involve the belief that the relevant fact constitutes a threat, then such attitudes will be knowable transparently to the extent that our capacity to deliberate about the relevant beliefs renders our anger and fear themselves open to deliberation. These attitudes will also, of course, involve passions that have a measure of independence from belief, and this

---

[28] Once again, I am taking it for granted that we have *knowledge* of our own minds. If someone is inclined to deny this, however, my point can be defended on the weaker assumption that our claims about our own mental states are genuine

begins to explain why such attitudes can prove recalcitrant in the face of reflection on reasons. To this extent, there is room for an appeal to some analogue of the expressivist story about pain in explaining our knowledge of such attitudes: learning to say when I am angry or afraid will no doubt involve learning to use words for what was formerly expressed by other kinds of behavior. But for such a thing as propositional anger or fear to be possible at all, these primitive stems of anger and fear must intertwine with our capacity for articulate belief, so that the resulting emotions at least normally bear a relation to our application of the concepts *insult* and *threat*.

In the case of conative attitudes such as desire and intention, the story is different but analogous. Grounds for conative attitudes are not reasons for thinking something true but reasons for thinking something desirable, reasons why there is "something to be said" for a course of action or an object of attainment. My capacity to know my own conative attitudes through deliberation is, like my capacity to know my own beliefs through deliberation, connected with my capacity to answer a certain sort of "Why?"-question; but the relevant "Why?" is different: it is, I think, the "certain sense of the question 'Why?'" that G. E. M. Anscombe investigates in her *Intention* (1963). Nevertheless, like our capacity to answer the truth-oriented "Why?"-question, our capacity to answer this practical "Why?"-question is arguably basic to our very status as rational thinkers and agents. Demonstrating this basicness would involve showing two things:

(P1) that the capacity to say why it would be desirable to do something is a necessary condition of the capacity to make non-observational statements about what one is doing; and

(P2) that the capacity to use the first person in non-observational statements about what one is doing is a necessary condition for understanding the first person at all.

These points are the analogues of the two points I have made about the truth-oriented "Why?"-question:

(T1) that the capacity to say why something is true is a necessary condition of the capacity to express$_A$ what one thinks (i.e., to make assertions); and

(T2) that the capacity to use the first-person in self-ascriptions of thought is a necessary

---

assertions.

condition for understanding the first person at all.

I will not argue for the practical analogues of these points. (P1) is defended in Anscombe's *Intention*, and there is something like a defense of (P2) toward the end of her essay "The First Person" (1975).[29] If (P2) were accepted, it would follow that no creature can possess any explicit *self*-knowledge unless it has the capacity for non-observational knowledge of what it is doing, just as no creature can possess any explicit self-knowledge unless it has the capacity for transparent knowledge of its own beliefs. If (P1) were accepted, we would have an explanation of the possibility of such knowledge that connected it with the ability to deliberate. Moreover, since it seems that I cannot understand an utterance of mine as an assertion unless I understand making that utterance as my own voluntary action, (P1) and (P2) would together entail that any creature capable of making comprehending assertions at all – of expressing$_A$ its beliefs in speech – must not only be capable of answering the truth-oriented "Why?"-question, but the practical "Why?"-question as well.[30]

These topics would obviously merit a much more extensive discussion in a full-scale account of self-knowledge. My purpose here, however, has just been to sketch how the points I have made about our knowledge of our own beliefs might be extended to the larger set of attitudes on which Moran takes his account to bear, and to indicate how this account might be *fundamental* to the understanding of our knowledge of our own deliberated attitudes without being a *total* account of such knowledge.

---

[29] For a fuller defense of (P2), and one that rejects Anscombe's notorious claim that "I" is not a referring expression, see John McDowell, "Referring to Oneself" (1998a).

[30] That there is a connection between being able to make comprehending assertions and understanding oneself as an agent is argued in McDowell 1998a, §VI. If this is right, it suggests an argument for the thesis that a creature capable of making comprehending assertions must possess a first-person concept. The argument would be that (1) the ability to make comprehending assertions requires an understanding of assertions as one's own voluntary actions, and (2) understanding something as one's own voluntary action requires the ability to self-ascribe actions. But (3) an utterance is only intelligible as a *self*-ascription of action if it involves a comprehending use of the first person, so our thesis follows.

6. CONCLUSION: TWO KINDS OF SELF-KNOWLEDGE

I have been arguing that an account of our knowledge of our own deliberated attitudes must form a fundamental and distinct part of any satisfactory account of self-knowledge. I take it as obvious that a satisfactory account will also have to tell some other story about our knowledge of our own sensations and appetites: this is one point on which Moran and his critics agree. If this is right, then we must reject (UA) and admit that our ability to speak authoritatively about our own minds draws on (at least) two different kinds of self-knowledge. I want to conclude by emphasizing two ways in which the resulting outlook resembles the Kantian view of self-knowledge that I mentioned at the outset, and by saying something about a topic that I marked early on but have said little about recently: the sense in which our knowledge of our deliberated attitudes involves a kind of agency.

One similarity between Kant's view and the view advocated here is that both draw a sharp distinction between the way we know our own judgments about what is the case and what to do, on the one hand, and the way we know our own sensations and appetites, on the other. A way of putting the thesis of the last section is to say that our immediate, authoritative knowledge of our own judgments is a necessary byproduct of our ability to reason – to make up our mind – about what is the case and what to do. Whether or not the expressivist account of our knowledge of our own sensations and appetites is on the right track, however, it seems clear that our knowledge of our own sensations and appetites is not in this sense maker's knowledge. Sensations and appetites are states that come to pass with us, not states we arrive at through deliberation. And this sounds strikingly like what Kant says: that whereas our apperceptive knowledge of our own judgments is a knowledge of "what we are doing," our knowledge of our sensations and appetites through inner sense is a knowledge of what we "undergo."[31]

---

[31] The characterizations I am quoting are from Kant's *Anthropology*, §24 (Ak. 7:161). But the idea that apperception gives us knowledge of ourselves *qua* active (or "spontaneous"), while inner sense gives us knowledge of ourselves *qua* passive (or "receptive"), is common to all of Kant's discussions of the topic. Nor is the idea new to Kant. Aristotle remarks in *De Anima* that "[s]ensation depends, as we have said, on a process of movement or affection from without" (1984, Vol. I, 416b33-4), and goes on to argue that thought contrasts with sensation in this respect.

This Kantian contrast between an active and a passive form of self-knowledge has been a source of puzzlement to commentators. Our discussion of Moran and Finkelstein, however, has equipped us to see a point in such a distinction. For on the one hand, we have seen that it is attractive to understand our knowledge of what we believe as reflecting our capacity for a kind of agency – the capacity to make up our minds on the basis of grounds for belief.[32] The point of this invocation of agency is not that every actual self-ascription of belief reflects a present *exercise* of the latter capacity. The connection is rather at the level of the capacities themselves: only a creature *capable* of making utterances in a way that is responsive to the sense of the question "Why?" that asks for grounds for holding-true can express$_A$ its beliefs in speech at all, and once this capacity is in place, the capacity to make immediate, authoritative *self-ascriptions* of belief requires only a harnessing of the former capacity to the use of the expression "I believe." The point of saying that our capacity to self-ascribe beliefs reflects a capacity for agency thus comes to this: the kind of thing a belief is is to be understood in terms of its relation to reasons of a certain kind (reasons bearing on the truth of the proposition believed), and the kind of thing a self-ascription of belief is is to be understood in terms of its connection with the ability to take account of such reasons in explicit deliberation. By contrast, our knowledge of our own sensations plainly is not connected in *this* way with reasons and deliberation. Sensations are states on which the notion of deliberation, and the associated notion of reasons for and against, do not get a grip. They are, as Kant says, states we passively undergo.

I do not claim that this contrast between an active and a passive form of self-knowledge is perfectly clear; my point is just that the impulse to draw such a contrast is not unintelligible – that it has a basis in ordinary things we want to say about states of these different kinds. An elucidation of this contrast between activity and passivity might begin by considering the different kinds of explanations that

---

[32] Moran himself suggests that this thought has a Kantian provenance (see Moran 2001, Chapter 4, §7), although he does not discuss Kant's views in any detail.

we give of these different kinds of states.[33] For instance, it seems an important fact about explanations of the form

(E) *S* believes that *p* because she believes that *q*

that the subject of such an explanation is expected to be able to *produce* the relevant ground: if she is asked why she thinks that *p*, she should be able to say, "Well, *q*." Indeed, she must not only be able to produce the relevant ground; she must take it to *be* a ground – must take it to speak in favor of the truth of the claim that *p*. If either of these conditions is not met, we will withdraw (E), either denying that *S* believes that *p*, or at least denying that she does so on the ground that *q*.[34] By contrast, explanations of sensations are not subject to such conditions. This, I think, begins to bring out the sense in which what I believe is *up to me*, whereas what I sense is not: it belongs to the kind of state belief is that, for a subject capable of reflecting on reasons, what the subject believes depends on what *she takes there to be grounds* for believing, whereas it belongs to the kind of state sensing is that what a subject senses is not responsive to any such assessment.

Having made these observations, we are in a position to respond to the Shah-Velleman point mentioned in §2: the point that often when we ask ourselves whether we believe that *p*, we find our minds

---

[33] For a development of the idea that explanations of belief are associated with a distinctive kind of "because," see Sebastian Rödl's *First-Person Order* (forthcoming). My remarks on this point, and the perspective of this chapter more generally, are deeply indebted to Rödl's work on the first person. See also Rödl 1998, of which the forthcoming book is a (much revised) translation.

[34] Perhaps the *sentence* "*S* believes that *p* because she believes that *q*" could also be used to report a causal relationship for which the constraint stated in the text does not hold. The kind of use I have in mind would be one reporting what is commonly called a "deviant causal chain." Suppose, for instance, that a person came to believe that *q*, and suppose that, as a result of an odd flaw in her *psyche*, this belief caused her to have a nervous breakdown, one result of which was her acquiring the unshakeable conviction that *p*. Perhaps we could then explain her belief that *p* using the sentence in (E), and in this case there would be no requirement that she be able to produce the explanatory belief or that she regard *q* as a good ground for believing that *p*. But this, I want to suggest, would not be the same *explanation* we have when (E) is used in the normal way. We can see this from the fact that a person intending to use (E) in the normal way would withdraw the claim if it turned out that *S* could not identify the belief that *q* as her ground. This reflects the fact that there is a kind of "Why?"-question to which only a ground known to the subject, and endorsed by the subject as a ground, can be a true answer. (This remark may be too concessive in any case. In the case of a deviant causal chain like the one described, the proper explanation would presumably be "*S* believes that *p* because she *believed* that *q*." This contrasts with (E), in which both of the connected propositions are in the present tense. The normal form of explanation of a belief by another grounding belief does not involve reference to a temporal sequence. This point is connected, I think, with Kant's reasons for claiming that apperception stands outside the temporal form of our sensibility. See A533/B561 and *Anthropology*, I, I, §7 [Ak. 7:142].)

already made up. This is certainly true, but if it is supposed to show a deep difference between the way we know beliefs that we arrive at through present deliberation and the way we know beliefs that we simply call to mind, it misses the point of describing our knowledge of our own beliefs as a kind of active knowledge. The point, to repeat, is that, given the connection for me between the question whether I believe that *p* and the question whether *p*, and given the connection between my capacity to say whether *p* and my capacity to answer the truth-oriented "Why?"-question, it follows that my capacity to answer questions about what I believe is necessarily tied to my *capacity* to deliberate. We could put it this way: when I speak about what I believe, I am speaking about *how my mind is made up* – even if I am not *making* it up at the moment, even if I *never* went through a process of conscious deliberation about the belief in question. For only because I am able to make utterances in a way that reflects an appreciation of the truth-oriented "Why?"-question – an appreciation manifested on those occasions when I actually make up my mind by deliberating, but present even where no actual deliberation has taken place in virtue of the truth of counterfactuals like the ones emphasized in the last paragraph – am I capable of speaking about my beliefs at all.[35] To make this point is not to deny that there can be beliefs that prove recalcitrant to reflection, perhaps even beliefs that we only come to recognize in ourselves through self-observation and self-analysis. It is only to state the rule against whose background such exceptions are intelligible.

On the Shah-Velleman picture, to answer the question whether I *already* believe that *p*, I must put the question whether *p* to myself "as a stimulus applied … for the empirical purpose of eliciting a response." But surely this is a mad picture of how I normally know what I believe, even in cases where the belief is already extant and I do not now consider grounds for and against. Shah and Velleman's description makes it sound as though I test my belief as to whether *p* as I might call into a well to see if

---

[35] Compare Anscombe's remark on the significance of Aristotle's account of the practical reasoning:

> [I]f Aristotle's account were supposed to describe actual mental processes, it would in general be quite absurd. The interest of the account is that it describes an order that is there whenever actions are done with intentions. (1963, §42, p. 80)

there is an echo. If this were the case – if I just found the assertoric "*p*!" coming back when the interrogative "*p*?" was sent in – then it is hard to see how this reply could figure in my present reflection as anything but the testimony of an alien voice, whose rational significance for my thinking now was an open question. But surely it is crucial that things are not normally like this: normally, the fact that I believe something, and know myself to believe it, has immediate rational significance for my present reflection, whether or not I am now forming the relevant belief in response to a deliberation. Whatever way I have of retaining beliefs must retain their rational significance for me; otherwise it isn't *beliefs* that are being retained. And the explanation of how beliefs can retain their rational significance is that an extant belief that *p* is the same *kind* of thing as a newly formed belief that *p*: it is my answer to the question whether *p*, formed and maintained in a way that is normally responsive to my reflection on grounds.[36]

This is all I will say about the first point of contact with Kant: the emphasis on the distinction between our knowledge of our own judgments and our knowledge of our own sensations, and the thought that this distinction maps onto a contrast between an active and a passive kind of self-knowledge. I want now to turn to a second point of contact: the thought that there is a relation of dependency between these two kinds of self-knowledge. Kant tells us that it must be possible for the apperceptive "I think" to accompany all of my representations if my representations are to be thinkable at all (B132). This implies, and it is clear from other passages that Kant holds, that the representations I receive through "inner sense," in particular, would not be thinkable by me if I did not have the capacity for apperception.[37] Let me restate this, without argument, in a way that makes its significance clearer: the

---

My point about the connection between belief and deliberation, similarly, is that what we see displayed in explicit deliberation about what to believe is an order that must be there whenever beliefs are held for reasons.

[36] I am indebted for the language of "immediate rational significance," and for the general shape of the point made here, to Burge 1996 and 1998.

[37] See B140, and compare the much-discussed letter to Marcus Herz of May 26, 1789 in which Kant claims that if I were a creature lacking the capacity for apperception, my representations would take place "without my knowing the slightest thing thereby, not even what my own condition was" (Ak. 11:52), as well as the discussion in Kant's *Reflexionen* of the question "Is it an experience that we think?" (R5661, Ak. 18:318-9).

claim is that I would not be able to think about the kinds of states that are the objects of inner sense – sensations, appetites, and other kinds of mental "affection" – if I did not also have the distinctively active sort of awareness I have of my own thoughts and judgments. Now, put this way, Kant's claim is again strikingly similar to a claim for which I have argued. For I have argued that our ability to self-ascribe sensations and appetites depends on our ability to make immediate, authoritative self-ascriptions of belief, since our very capacity to make assertions at all depends on our capacity to make utterances in a way that reflects an appreciation of the truth-oriented "Why?"-question, and our ability to make comprehending assertions in the first person, in particular, depends on our having mastered the use of the first person in transparent self-ascriptions of belief.

My argument for this dependency turned on the thought that a creature has only mastered the use of the first person, which must figure in any genuine *self*-ascription, if it has mastered its use in transparent self-ascriptions of belief. And this, again, is a Kantian thought. Kant puts it in his way by saying that "the *I* is only the consciousness of my thinking" (B413). What he means is that this element of thought and intelligent speech has its significance in virtue of its connection with a certain kind of knowledge, the knowledge we have of our own thoughts in thinking them, in making up our minds.

A way of putting my point in this chapter is to say that, unless we recognize this kind of self-knowledge as fundamental, and distinct from our knowledge of what we sense, we will not be able to see clearly what makes the object of self-knowledge a *self*. For to be a self is to be a thinker and an agent, and to be a thinker and an agent is to be capable of a kind of activity that stands in contrast to the passivity of sensation. Nor is this merely a point that must be acknowledged by *theorists* of self-knowledge. Even to be capable of having such mundane thoughts as "I'm in pain," we must have a conception of the active subject to which sensations belong. For, as I have argued, to understand the use of the first person in ascriptions of sensation requires understanding its use in ascriptions of belief, ascriptions that can be made on the basis of a deliberation. It requires, in other words, that our representation of the subject whose sensations are in question imply that this subject also possesses a capacity for spontaneity. And

this, in fact, is just what Kant says about "the representation *I*": it is a representation of "the spontaneity of a thinking subject" (B278). Our conclusion is that only a creature possessed of such a representation can know itself.

VI. FAILURES OF SELF-KNOWLEDGE

> What is the form of this ignorance, an ignorance of something I cannot just not know? Is it to be thought of as keeping a secret? But in what form can I keep a secret from myself, keep silent? To keep silent around myself I have to silence myself; I keep myself in the dark by darkening myself.
>
> Stanley Cavell, *The Claim of Reason* (1979), p. 388

1. INTRODUCTION

Descartes famously held that "there can be nothing in the mind, in so far as it is a thinking thing, of which it is not aware."[1] There is general agreement among contemporary philosophers, however, that this Cartesian identification of the mental with the conscious must be rejected. Several factors have contributed to this consensus: psychoanalysis has had an influence, as have advances in the cognitive sciences, and shifts in the philosophy of mind itself. The result is that contemporary philosophers of mind are, in general, quite at ease with the idea of mental representations that fall outside the scope of consciousness. Indeed, a significant number of contemporary philosophers regard it as unacceptable for an account of mind to tie mentality to consciousness in any very intimate way. Thus we hear that

> [t]he reasons for which people believe things are rarely conscious. People often believe things for good reasons, which given them knowledge, without being able to say what those reasons are.[2]

And that

---

[1] René Descartes, *Philosophical Writings* (1984), Vol. II, p. 171.

[2] Gilbert Harman, *Thought* (1973), p. 28.

> we have no more reason to identify being a mental state with being a conscious state than we have to identify physical objects with physical objects that somebody sees… Consciousness seems central to mentality only because it is so basic to how we know our own mental states.[3]

To a philosopher who finds these sorts of views congenial, my account of the representational powers of a rational creature, which ties the very idea of a representation to the capacity for self-consciousness, may look like a reversion to Cartesianism. Now, if to be a Cartesian is just to hold that there is some important connection between the capacity for thought and the capacity for self-conscious awareness of thought, then I am happy to have my view called "Cartesian." It is important, however, to be clear about what such Cartesianism does and does not entail. Does it require that we deny the existence of unconscious thoughts, or of reasons that operate "beneath the radar" of consciousness? In this final chapter, I consider what sense my Kantian approach to representations can make of the idea of an unconscious representation — a representation that the subject is *not* in a position to "accompany with 'I think'." My aim is to show that the account of mind that I have been developing need not deny what is obvious, namely that we sometimes come to understand a person better by positing failures of self-knowledge on her part. Indeed, I want to suggest that my account is not only *compatible* with such explanations, but can in fact help to clarify their significance, and the limits on their intelligibility. For, although some philosophers write as if the idea of, e.g., an unconscious thought were no stranger than the idea of a conscious one, I think anyone who reflects seriously on the idea of an unconscious thought will find it quite weird. One advantage I shall claim for my account is that it explains this weirdness, while not requiring us to deny that the idea of an unconscious thought can have its uses.

My underlying concern in this chapter, then, is with the very idea of an unconscious thought. My *topic*, however, is more specific. To see what is weird about the idea of an unconscious thought, it is helpful to reflect on the widespread sense that there is something paradoxical about the very idea of self-

---

[3] David M. Rosenthal, "Two Concepts of Consciousness" (1986), p. 330.

deception. For when we describe a subject as self-deceived, we seem to posit in her an underlying recognition of a fact of which she is not consciously aware. And of course, not being consciously aware of the fact in question, she cannot reflect self-consciously on her recognition of this fact, either. So we seem to have here an explanation that posits a personal-level representational state that is inaccessible to self-consciousness. If we can make sense of such explanations, we will have the beginnings of an account of unconscious thought.

*

Consider, then, the sort of claim we make about a person when we say such things as the following: "He's self-deceived," "He's lying to himself," "He's in denial," "He can't bring himself to face the facts." These various descriptions have two significant features in common: first, they all posit a failure on the subject's part to recognize facts which he could normally be expected to recognize; and secondly, they all imply that this failure is itself motivated. What the subject fails to recognize may, in the first instance, be either a fact about himself or a fact about the world beyond himself; but whatever else he fails to perceive, one thing he must not recognize is the role of his own motives in shaping what he thinks. We may thus say that, in applying these sorts of descriptions, we explain a person's thoughts or actions by positing a *failure of self-knowledge* on his part.

The fact that people are capable of such failures has long been a source of wonder and puzzlement. In St. Augustine's *Confessions*, for instance, we find Augustine marveling at his capacity to hide the truth of his situation from himself:

> I had placed myself behind my own back because I did not wish to observe myself… I had known [my iniquity], but deceived myself, refused to admit it, and pushed it out of my mind.[4]

The sense of paradox reflected in Augustine's talk of placing himself behind his own back is a

---

[4] *Confessions*, VIII. vi [16].

characteristic reaction to the discovery that one has been deceiving oneself. A similar sense is captured in a famous scene from Henry James' *The Ambassadors*, a scene in which the protagonist, Strether, is finally forced to recognize the nature of the relationship between two people he has come to admire. James describes the revelation as follows:

> He almost blushed, in the dark, for the way he had dressed the possibility in vagueness, as a little girl might have dressed her doll… 'What on earth—that's what I want to know now—had you then supposed?' He recognized at last that he had really been trying all along to suppose nothing.[5]

What makes it apt to describe these as cases of self-deception is precisely this suggestion of an ongoing effort to ignore some unwelcome fact. The subject, it seems, has not merely believed something wishfully or foolishly; he has actively *tried* to maintain his belief in the face of contrary indications. Moreover, the apparent purposefulness of this activity suggests that the subject at some level worries that what he believes is not so. For what could guide him in steering his attention away from certain facts but an obscure recognition of their unwelcome implications? And what could motivate his ongoing effort to seek out rationalizations for his preferred belief but a persistent sense of doubt?

It can hardly be denied that such descriptions are sometimes apt: there are occasions on which they seem to express the best sense we can make of a person. When we pause to reflect on these descriptions, however, it is hard to see what they can mean. For how can a person hide doubts from *himself*? We can sympathize with Augustine's wonder at how he placed himself behind his own back, and Strether's surprise at discovering that he has been trying to suppose nothing, because we can imagine ourselves having these reactions. To discover that one has been deceiving oneself will involve finding oneself hard to understand: "How," one will want to ask, "could I not have known?" But what exactly is hard to understand here, and what sort of understanding do we want?

Many philosophical discussions of self-deception assume that what is hard to understand about

self-deception can be captured in a paradox, and that what we want is an account of how this paradox can be resolved.[6] On this view, the basic problem of self-deception is simply to understand how the very same person can be both deceiver and deceived, given that a deceiver must know what she is up to, whereas a person who is deceived presumably must not know that a deception is taking place. The main philosophical labor is then devoted to considering strategies for resolving this paradox, and two kinds of strategies tend to dominate the discussion: (1) strategies that seek to *resolve* the paradox by positing some kind of division in the self-deceiver's mind, so that her belief can be located in one part of her mind and her doubt or deceptive intention in another; and (2) strategies that seek to *dissolve* the paradox by denying that real cases of self-deception need involve the compresence of belief and doubt, or of an intention to deceive and an obliviousness to that very intention.

My aim in this chapter is not to join this debate but to question the outlook that produces it. Although interesting work has been done in each of these two traditions, I think there is an explanatory task that both have neglected, a task which is obscured by their shared focus on resolving paradoxes. For surely what is puzzling about self-deception is not just that the self-deceived subject is the bearer of contrary psychological predicates. What is puzzling is that a person can be out of touch with her own thoughts in this striking way. This is what makes the question "How could I not have known?" an apt expression of our puzzlement: all paradoxes aside, we find it remarkable that a person could fail to know her own doubts, and her own efforts to conceal them. If this is right, however, then it seems that a satisfactory account of self-deception must begin by reflecting on why a person *should* be expected to

---

[5] *The Ambassadors*, Book Eleventh, Chapter III. This passage was brought to my attention by Patrick Gardiner, "Error, Faith, and Self-Deception" (1976), a helpful paper to which I am indebted at several points in what follows.

[6] This emphasis on paradoxes is reflected, for instance, in Alfred Mele's review article "Recent Work on Self-Deception" (Mele 1987b), whose whole structure is dictated by the assumption that the task of an account of self-deception is to resolve paradoxes. Whatever the merits of this assumption, it does reflect how most writers on self-deception saw their task at the time when Mele wrote. Nor have things changed a great deal since: although there have been a few discussions of self-deception that place it in the context of larger questions about self-knowledge (notably in Moran 2001 and Scott-Kakures 2002, both of which will receive some mention below), most writers still treat self-deception as a

know her own doubts. Why should a failure of self-knowledge seem problematic at all? To try to resolve paradoxes about self-deception without considering this question is to address the symptoms of our puzzlement without considering its source.

I will argue that our puzzlement over self-deception has its source in facts about self-knowledge. To understand what makes self-deception problematic, I will claim, we must first understand why a self-conscious subject must normally know what she believes, and what her grounds are for believing it. Once we understand this, we will be in a position to see why self-deception *should* seem puzzling – why there should be limits to our ability to make sense of a person by seeing her in such a light. It will emerge that the aura of paradox surrounding self-deception is a reflection of a deep fact about the connection between belief and self-consciousness, a fact that is fundamental to the possibility of rational thought.

I begin (§2) by describing two challenges that an account of self-deception must meet, challenges that accounts which focus exclusively on resolving paradoxes tend to ignore. Having articulated these challenges, I then turn (§§3-4) to a consideration of two landmark papers on self-deception, one by Donald Davidson and another by Mark Johnston. Davidson and Johnston's papers are among the founding documents in the debate over how to address the paradox of self-deception: Davidson is one of the most famous proponents of the divided mind strategy, while Johnston is one of its most prominent critics. I argue, however, that underlying their disagreement about how to address the paradox of self-deception is a more basic dispute about the character of ordinary rational belief-formation. Thinking through this disagreement will put us in a position to see a connection between the puzzling character of self-deception and the importance of self-consciousness (§5). And having seen this connection, we will be able to see why self-deception is a pathology special to rational creatures, and what truth there is in the suggestion that a self-deceiver must have a divided mind (§6).

---

source of paradoxes that can be considered more or less in isolation.

2. TWO CHALLENGES FOR AN ACCOUNT OF SELF-DECEPTION

What should an account of self-deception explain? As I have already mentioned, many writers on self-deception assume that the primary task of a philosophical account of self-deception should be to resolve a paradox. In my view, however, the phenomenon of self-deception presents explanatory challenges which are not well-captured in paradoxes, and which accounts that focus on resolving paradoxes consequently tend to overlook. In this section, I articulate two such challenges, and argue that some prominent treatments of self-deception do not address them.

*

If the problem of self-deception is just that it generates a paradox about how the same person can be both deceiver and deceived, then it seems that we ought to be equally puzzled about how a person can remind himself of his dentist-appointment or teach himself French.[7] For it is possible to generate formally similar paradoxes about these activities. To remind someone of something, one must presumably already remember it; but then how can one remind oneself? To teach someone some body of knowledge, one must presumably already possess that knowledge; but then how can one teach oneself? Yet no one is long puzzled by *these* paradoxes. We can easily see how a person could remind herself of an appointment or teach herself French: she must simply take steps to ensure that, without anyone else's intervention, the relevant end is achieved; and we can think of various steps she could take (e.g., tying a string around her finger, studying a textbook).

In the case of self-deception, by contrast, the "how possible?" question genuinely troubles us. It is not that we are unable to describe states that could be called "being self-deceived" and steps a person

could take to put herself in such a state. But to just the extent that the state is unproblematic, it seems not to be what we had wanted to describe. If this is right, however, then posing paradoxes about how one person can be both deceiver and deceived does not get to the root of our puzzlement, for it leaves us with no account of why we are more puzzled by self-deception than we are by the reflexive variants of other transitive activities. This, then, is a first challenge for an account of self-deception: before attempting to show how self-deception is possible, it should explain what makes it seem specially *impossible*. And merely to say that we are inclined to understand deceiving oneself on the model of deceiving another is not to fulfill this task, for we need to understand *why* we are inclined to conceive of self-deception in this way, when we do not fall into this trap with just any reflexive counterpart of a transitive construction.[8]

---

[7] These examples are due to T. S. Champlin, *Reflexive Paradoxes* (1988).

[8] Many treatments of self-deception give this explanatory task remarkably short shrift. For instance, several recent discussions seem committed to the view that puzzlement over self-deception results only from outdated ideas about how minds work. Thus we are told that, although "the traditional view" of self-deception supposes that motivated biasing of belief can only be explained by positing an agency in the subject that carries out a project of deception, psychological research shows that

> [c]old or unmotivated biased cognition is not explained on the model of intentional action, and motivation can prime mechanisms for the cold biasing of data in us without our being aware, or believing, that evidence favors a certain proposition. (Mele 1997, p. 95)

Or else it is emphasized that

> [q]ua irrational beliefs, self-deceptive states are not unique. They make us feel uncomfortable because states of self-deception threaten the conviction that beliefs and evaluative attitudes 'don't mix': that they are generated through entirely different kinds of processes. Ideally, beliefs ought to be formed exclusively in response to epistemically relevant factors. The psychology of the formation of beliefs, however, suggests that they are systematically sensitive to a wide range of factors, many of which are epistemically irrelevant. (Lazar 1999, p. 286)

The suggestion, in short, is that an up-to-date understanding of how nonrational biasing factors affect belief-formation will see self-deception as just one instance of tendencies that are pervasive and unproblematic. And the implication seems to be that what has made self-deception *seem* puzzling to earlier writers is just some naïve, rationalistic assumption about how beliefs must originate – the assumption that all biasing requires a biaser, or that beliefs and evaluative attitudes "don't mix," or something of the sort.

What is striking about such accounts is how little they say about why these assumptions should ever have seemed compelling. No doubt recent studies of factors that bias cognition have improved our understanding of the role that nonrational forces play in our thinking, but surely the existence of such forces was not unknown to earlier philosophers. And in any case, puzzlement over self-deception is not merely a thing of the past, or a phenomenon restricted to philosophers: *anyone* who tries to imagine making the kind of discovery described by Augustine or James

*

A second challenge for an account of self-deception is to explain what a mind must be like if it is even to be capable of this sort of pathology. This, again, is a topic about which many prominent accounts have little to say.[9] Indeed, a number of influential accounts define self-deception in such a way that it is not even clear why a self-deceiver should have to be a rational creature at all. In a widely-discussed paper, for instance, John V. Canfield and Don F. Gustavson argue that to deceive oneself is to acquire a wishful belief in "belief-averse circumstances."[10] This account, while crude, is arguably the progenitor of the dominant contemporary approach, which seeks to explain self-deception as the result of blind but motivated psychological tendencies that bias a subject's gathering or weighing of evidence.[11] Now, on reflection, there seems to be no principled reason why a dog could not come to believe something as a result of this sort of process. I once spent a few days looking after a Labrador Retriever called "Scout" while her owner was out of town. During those days, Scout barked plaintively at any

---

will, I think, be puzzled about how such a thing could have occurred. If this is right, however, then it seems insufficient to trace this puzzlement to some merely optional assumption about how minds work. The challenge is not only to say what assumptions about belief-formation lie behind this puzzlement, but to explain why these assumptions have a grip on us, and why this grip persists in the face of all the evidence, both scientific and anecdotal, of how intelligent people can be led astray.

[9] An important exception is Dion Scott-Kakures' perceptive "At 'Permanent Risk': Reasoning and Self-Knowledge in Self-Deception" (2002). Scott-Kakures and I agree in thinking that there is a close connection between the capacity for self-deception and the capacity to reflect on one's own beliefs, and that we can see this connection by reflecting on why we do not suppose that mere brutes can deceive themselves. However, although we agree on these points, our conceptions of what it is to reflect on one's own beliefs are very different. I cannot offer a detailed comparison of our views here, but it begins to bring out the difference between us to note that the "transparency" of first-person present-tense belief-ascriptions, which I discuss below in §5, plays no role in Scott-Kakures's account. I shall argue that it is precisely this feature of the concept of belief that makes self-deception puzzling, and that represents the key constraint on an account of this phenomenon.

[10] See their "Self-Deception" (1962).

[11] Contemporary examples of this approach include Barnes 1997, Lazar 1999, and Mele 2001, as well as Johnston 1988, which I discuss extensively below. The accounts put forward by these authors are less obviously subject to the problem about allowing for nonrational self-deceivers than is the account discussed in the text, since these more recent accounts tend to posit tendencies toward biased "hypothesis-testing" and "evidence-gathering," which sound like activities that only a rational creature could engage in. But (1) we are typically not told how to distinguish real hypothesis-testing and evidence-gathering from its nonrational analogue; and in any case (2) no account is offered of why this difference should matter, why the kind of irrationality that rational creatures can exhibit is importantly different from the kind exhibited by nonrational creatures. In the absence of such an account, characterizing self-deception by appeal to biased "hypothesis-testing" or distorted "evidence-gathering" is *ad hoc*.

noise in the stairwell, apparently supposing that the person on the stairs was her returning master. In less stressful circumstances, however, Scout is able to tell the tread of her master from that of other persons. Did her desire to believe that her master had returned lead her to misjudge signs which in other circumstances would not have misled her? Nothing turns on whether this was actually the explanation of her uncharacteristic barking; all that matters is that it seems coherent to suppose that this *might* have been the explanation. If self-deception is simply motivated belief in belief-averse circumstances, then this would seem to count as a case of her being self-deceived.[12]

Yet surely intuition speaks against calling this a case of self-deception. Surely, indeed, intuition speaks against calling *anything* in the life of a dog a case of self-deception: the possibility of self-deception seems to arise as a concomitant of a kind of rationality of which such creatures are simply not capable. We can of course speak of dogs as self-deceived if we like, and the question whether most English speakers would accept such a usage is obviously not one that can be settled *a priori*. But the real question is not what most English speakers would accept, but how useful it is to compare such canine "self-deception" with human self-deception. In what follows, I shall argue that we must draw a distinction here if we are to give a satisfactory explanation of our puzzlement over self-deception. In the meantime, I rest my case on the plain fact that nobody is really inclined to speak of brutes as self-deceived, except perhaps to make a philosophical point. This suggests that the applicability of ordinary talk about self-deception requires a certain background: a self-deceiver must have a mind of a certain complexity. A satisfactory account of self-deception should explain what the relevant complexity is and why it is necessary as a background to self-deception.

---

[12] This is a true story, but see Scott-Kakures 2002 for a similar example.

3. DAVIDSON ON DECEPTION AND DIVISION

One account of self-deception that does suggest answers to these challenges is the account given by Donald Davidson in his "Deception and Division" (1985a). Davidson's discussion is distinguished by an attempt to say what makes self-deception not just superficially paradoxical but interestingly problematic, and his answer hinges on the thought that there are general difficulties in understanding how irrationality is even possible. It is in response to these difficulties that Davidson makes his well-known suggestion that we should understand the mind of the self-deceiver as partitioned into distinct sub-agencies. For my purposes here, however, the more interesting aspect of Davidson's view is not his positive account of self-deception but his characterization of the difficulties that such an account must resolve.

On Davidson's view, the philosophical interest of self-deception is closely related to the interest of other forms of irrationality such as weakness of the will.[13] In general, he suggests, what is difficult to understand about irrationality is how a judgment that, all things considered, a certain proposition is better supported by the evidence than another incompatible proposition, or that a certain course of action is more desirable than some alternative course, can coexist with a belief in the less probable proposition or a decision to pursue the less desirable course. He often puts the problem as a difficulty about reconciling the possibility of irrationality with the general character of propositional attitude explanation: any given propositional *content* is the content that it is at least partly in virtue of the complex pattern of logical relationships in which it stands, and any given propositional *attitude* is rightly ascribed to a subject only insofar as it stands in rational relationships to other attitudes of the subject which reflect the logical relationships that characterize its content.

Davidson takes this constraint to be a direct consequence of the special kind of explanation in

---

[13] Davidson's most extensive discussion of self-deception occurs in his "Deception and Division" (1985a), but my account of his view also draws on his 1980b, 1982, and 1985b.

which attitudes like belief have their home: in the fundamental case, the causes we cite in explaining a person's beliefs are not causes that merely happen to rationalize what they cause; they explain what they cause precisely by rationalizing it. It belongs to the essence of propositional attitudes, in short, to figure in a special kind of rational-causal explanation. To find a system of attitudes *irrational*, however, is precisely to find that the attitudes do not live up to the basic rational principles in virtue of which the logical relations between propositions constitute a framework for the explanation of thought and action. The difficulty then is to understand how such a combination of attitudes could coexist, given the fundamental connection between the ascription of propositional attitudes and the project of finding rationality in a subject's thought and action. And the problem is especially acute in the case of self-deception, where a recognition that, all things considered, it is more likely that not-*p* seems not merely to coexist with a belief that *p*, but in some sense to sustain it.

Davidson's explanation of how such a situation is possible famously involves two elements: first, the idea of a mental cause that does not rationalize what it causes; and secondly, the idea of a partition in a subject's mind that prevents the subject from becoming jointly conscious of both members of a pair of conflicting attitudes. Since a person who is irrational believes something that is not consistent with what he takes the totality of evidence to support, or intentionally does something that he does not judge, all things considered, to be the desirable thing to do, irrationality must involve some mental state (often a wish or desire) causing a belief or an action which the subject himself does not take to be rationalized by that state. And to make sense of such nonrational attitude-causation, Davidson suggests, we need to posit a partition in the subject's mind: the partition marks the boundary across which the ordinary pattern of rational causation – the pattern by reference to which we are able, in the first instance, to make sense of the propositional attitudes of a single subject at all – breaks down. To posit such a boundary is not to explain *why* unreason has prevailed in a given case, but it is at least to make the idea of irrationality coherent: it gives us a way of accommodating directly conflicting attitudes within the mind

of a single subject, while holding onto the idea that in general we make sense of a person's mental states at all by reference to what Davidson famously calls a "constitutive ideal of rationality" (1980c, p. 223).

4. JOHNSTON ON INTENTIONALISM AND RATIONAL CAUSATION

In the literature on self-deception, Davidson is often grouped together with other so-called "intentionalists" – other philosophers who argue that genuine self-deception must involve the intention to deceive oneself.[14] This grouping has some justification: Davidson does claim the self-deceiver must "intend the 'deception'" (1985a, p. 144) and must act "with the intention of producing" his preferred belief (1985a, p. 145). And more generally, he suggests that we should understand the interaction of parts in an irrational person's mind on the model of the interaction between persons (see Davidson 1982, pp. 300-301). These claims have been widely criticized. For, it is argued, intentionalism rests on a failure to see that there can be processes which bias cognition, and which are motivated, but which do not involve any agency that acts with the intention of leading cognition astray.[15]

I will not attempt to defend Davidson against this criticism: I think the criticism is persuasive, as far as it goes. The question I want to consider is how far it does go: what implications does the criticism have for Davidson's more basic claim about the connection between propositional attitude psychology and a special form of rational-causal explanation? For this purpose, it will be useful to look at one of the seminal critiques of intentionalism: Mark Johnston's "Self-Deception and the Nature of Mind" (1988). Johnston's paper is one of the few discussions that criticizes not only Davidson's claim that self-deception must involve the intent to deceive, but also his more basic conviction that attitudes like belief

---

[14] Another important advocate of this view is David Pears, who argues for it in his *Motivated Irrationality* (1984).

[15] I have already cited three proponents of this view: Mele 1987a, 1997 and 2001, Barnes 1997 and Lazar 1999.

have their primary home in a special form of explanation. Reflecting on Johnston's critique will help us to distinguish the question of the truth of intentionalism from the question of the primacy of rational causation, and thus will set the stage for a consideration of the latter topic in its own right.

According to Johnston, what forces Davidson to posit a partition in the mind of the self-deceiver is only the unwarranted assumption that

(A) A self-deceiver's *motivated* self-misleading must be understood as *intentional* self-misleading.

If we assume that self-deception is an intentional project, it will of course be hard to see how a unified subject can deceive himself. For to do something intentionally seems to require knowing what one is doing and why one is doing it, and precisely this knowledge seems to stand in the way of one's being deceived. Hence, if we make assumption (A), it will seem necessary to posit a partition in the self-deceiver's mind, placing the agent who intends the deception on one side of the partition and his victim on the other. But, Johnston suggests, we can acknowledge that a self-deceiver motivatedly avoids the truth without supposing that he intentionally leads himself astray. For self-deception can be understood as a *subintentional* mental process: one that serves an interest of the subject, and whose occurrence is no accident, but which is not *undertaken for the sake of* that interest, or indeed for any reason. If I hope that *p* and am anxious that not-*p*, it may be in my interest to believe that *p*, and ignore evidence that not-*p*; for this would relieve my anxiety. And why, Johnston asks, should there not be a "mental tropism" which disposes a person in such circumstances to form the agreeable belief? To the extent that such tropisms are adaptive, it is perfectly intelligible that they might arise as a result of conditioning, or indeed be part of our evolutionary inheritance; and there is no reason why their operation should require that the subject be aware of what is going on. But if this is right, then – Johnston suggests – self-deception can simply be understood as a condition in which

> anxiety that one's desire that *p* will not be satisfied is reduced by one's acquisition of the belief that *p* (wishful thinking) and one's ceasing to acknowledge one's recognition of the evidence that not-*p* (repression) (Johnston 1988, p. 86)

where all of this occurs tropistically.

I think Johnston is clearly right that the notion of a mental tropism is intelligible, and that its intelligibility shows assumption (A) to be false. It should also be clear, however, that the appeal to mental tropisms does not directly address the more basic difficulty that Davidson finds in self-deception. For this difficulty does not depend on the assumption that self-deception is an intentional process: it turns on a thought about the connection between propositional attitude explanations and rationality that does not specially concern intentions. The difficulty is just to understand how it can be correct to ascribe propositional attitudes to a subject when those attitudes do not operate in the subject's mind according to rational principles. If this is a problem, it remains one even if we follow Johnston in conceiving of self-deception as a subintentional process.

Once we have the idea of a subintentional mental tropism in view, however, it becomes more difficult to see why nonrational mental causation *should* be problematic. Why shouldn't a propositional attitude cause other propositional attitudes that it does not rationalize? Why shouldn't we simply see self-deception as a counterexample to a view of mind which requires that propositional attitude ascriptions be governed by a constitutive ideal of rationality? This is precisely what Johnston concludes: that the intelligibility of tropistic explanation shows Davidson's conception of the mental to be unmotivated. There is not, he suggests, some special form of "rational causation" that is the primary *modus operandi* of mental states; there are simply various tropisms, some which tend toward rational outcomes and others which do not. Those tropisms that qualify as rational do so "not because of some *sui generis* manner—rational causation—in which the one attitude causes another but because the one attitude is in fact a reason for the other" (1988, p. 87). And a rational creature is simply one with "the capacity to inhibit conscious changes in attitude when [it] recognizes that those changes are not well grounded" (*loc. cit.*).

In short, although Davidson's account of irrationality promised to explain what makes self-

deception interestingly problematic rather than just superficially paradoxical, Johnston's criticisms make the interest that Davidson finds in the phenomenon look dubious. They make us lose our grip on why there should be anything specially problematic in the idea of nonrational mental causation, and, by the same token, they make it hard to see why one should hold that propositional attitude ascriptions are associated with a distinctive kind of explanation. Johnston need not deny that it is part of our concept of mental states like belief that they are *usually* caused by states that rationalize them; but he will deny that there is a distinctive *kind* of explanation at issue here. That is, he will deny that the kind of explanation of a belief we find in

(E1) *S* believes that *q* because she believes that *p* and believes that if *p* then *q*

differs in essentials from the kind we find in

(E2) *S* believes that *q* because she wishes that *q* and is anxious that not-*q*.

My own view is that Davidson is right to see a fundamental difference between the kinds of explanation found in (E1) and (E2). In the next section, I offer some reasons for holding that the sort of explanation offered in (E1) differs in kind from, and possesses an important kind of primacy relative to, the sort offered in (E2). Seeing this difference and this primacy will help us to see the truth in Davidson's idea that the primary form of explanation of belief is one that requires us to see the believing subject as rational. It will also help us to recover our sense of what is problematic about self-deception, and why it should require something like a split in the self.

5. BELIEF, REASONS FOR BELIEF, AND CONSCIOUSNESS

What is special about explanations like (E1) will become clearer if we consider a topic that neither Davidson nor Johnston says much about: the connection between belief and reflective *consciousness* of belief. It is a striking fact about the analytical literature on self-deception that it for the most part ignores

this topic. For surely part of what seems odd about self-deception is that we want to say that a self-deceiver does not recognize her own doubts. A self-deceiver must not only be deceived *by* herself; she must also be deceived *in* herself: she must fail to know her own mind. The difficulty is to see how this is possible.

But why, it will be asked, should it not be possible? Why should there be anything puzzling in the idea of an unconscious doubt? To see what might be puzzling about this idea, it is helpful to reflect on what conscious beliefs are like. Discussions of self-deception typically take it for granted that we cannot normally believe "counterevidentially" – that we cannot believe a proposition the evidence against which we consciously recognize to be overwhelming. But why is this so? What is the connection between belief and conscious recognition of evidence?

We can begin to see the connection here by recalling a point that has been a recurring theme in the preceding chapters: that the kind of creature that can think about its own beliefs, and thus can contemplate the possibility of changing them, will necessarily be a creature that can ask itself *why* it believes what it does, and can appreciate the special relevance of a certain kind of answer to this question. In our discussion of Moran in the last chapter, we noted that first-person present-tense belief-ascriptions are normally subject to the following principle:

(T) I can determine whether I now believe that *p* by considering whether *to* believe that *p* – a question I can answer by considering grounds for thinking that *p* is *true*.

(T) is often called the "Transparency Principle" because it appears to make the question whether I believe that *p* "transparent" to the question whether *p*: it says that I can settle the former question by settling the latter. Indeed, there seems to be an even stronger point in the offing. It is not just that I normally *can* settle the question of what I believe in this way; it seems that a person who *did not* in general settle the question of what he believed in this way – who habitually answered questions about what he believed with responses like "Judging from my behavior, it seems that I believe that *p*" or "I find myself with a yen to believe that *q*, although I can't think of any reason for believing this" – would suffer from a

radical kind of self-alienation, which would call into question the significance of these very judgments about himself. For the question would arise whether he genuinely believed these claims about his own beliefs, or whether these too should count as just more data for observation, more expressions of blind inclinations to aver. If a person can make up his mind on the basis of evidence at all, then it seems that he should be able to express his conclusion by saying "I believe that …", where what fills the ellipsis is the conclusion he has reached. But if the subject can make up his mind about whether he seems to believe that *p*, then why can't he make up his mind about whether *p*? And if he can make up his mind about whether *p*, then why does he need to rely on observation of his own behavior, or unaccountable yens, to say what he believes?

The principle of belief-ascription that we have called (T) thus appears to be intimately connected with the kind of state that belief is: a belief just is a creature's assessment of what is the case, its "take" on the truth, and a subject who did not recognize the entitlement expressed in (T) would not have grasped this. But if to understand what a belief is requires understanding the entitlement expressed in (T), then it is a mistake to conceive of belief as a state of a person's mental mechanism which, as it happens, that person can be trained reliably to detect. For if, in the normal case, I determine what I believe by considering what is true, then our normal "consciousness" of what we presently believe is not, in the first instance, a consciousness of how things are *with us* at all. In the normal case, I ascertain what I believe by thinking about what is true. Moreover, these observations about the connection between what I believe and what I take myself to have grounds for believing imply that *explanations* of belief like (E1) – explanations that account for a belief by citing other beliefs, memories or experiences that speak in favor of its truth – have a special status. For the causes of the belief that *q* cited in (E1) are precisely grounds for thinking that *q* is *true*, and explain the subject's belief precisely in virtue of being grounds of this sort. The "because" in (E1) corresponds to a "Why?"-question that could be put to the subject – "Why do you believe that *q*?" – and the beliefs mentioned on the right-hand side of the "because"

purport to give her grounds, grounds she could produce if asked.[16] The connection asserted in (E1) is thus not one that could hold without the subject's being aware of it: if, upon being asked why she believed that *q*, she could not produce the fact that *p* and the fact that if *p* then *q* as grounds, we would conclude that she did not believe it on this basis.[17]

If these claims about the connection between belief and grounds for belief are correct, then we must reject Johnston's picture of rational belief-formation as a matter of "accepting" tropistic belief-transitions when they are warranted and "inhibiting" them when they are not. For on the one hand, no "transition" from believing that *p* and that if *p* then *q* to believing that *q* would count as a case of my believing that *q because* I believe that *p* and that if *p* then *q*, in the relevant sense of "because," unless the inference was one I would accept as warranted. And on the other hand, if I judge that a transition is *not* warranted, then it seems that I do not have to *do* something to inhibit it: to judge that it is not warranted just is not to believe the conclusion, or at least not to believe it on this basis.[18]

These facts can only look mysterious from Johnston's perspective, for Johnston seems to assume that the following states are, in general, distinct:

(i) believing that *p*,

---

[16] This is not to suggest that our subject's ground for believing that *q* is merely *that she believes* that *p* and that if *p* then *q*. On the contrary: her grounds will normally be the (supposed) facts themselves, not the mere fact that she holds certain beliefs. But my point is that, given the connection between what a subject believes and what she takes to be true, there is not normally any opposition *from her standpoint* between believing something because *p* and believing something because she believes that *p*. For the question whether she believes that *p* just is for her the question whether *p* is true.

[17] Perhaps the *sentence* that occurs in (E1) could also be used to report a causal relationship for which the constraint stated in the text does not hold. The kind of use I have in mind would be one reporting what is commonly called a "deviant causal chain." Suppose, for instance, that a person came to believe that *p* and that if *p* then *q*; and suppose that, as a result of an odd flaw in his *psyche*, these two beliefs caused him to have a nervous breakdown, one result of which was his acquiring the unshakeable conviction that *q*. We could then explain his belief that *q* using the sentence that occurs in (E1), and in this case there would be no requirement that he regard the causally-explanatory beliefs as grounds for his belief that *q*. But this, I want to suggest, would not be the same *explanation* we have when (E1) is used in the normal way. We can see this from the fact that a person intending to use (E1) in the normal way would withdraw the claim if it turned out that the subject in question could not identify the beliefs that *p* and that if *p* then *q* as her grounds. This reflects the fact that there is a kind of "Why?"-question to which only a ground known to the subject, and endorsed by the subject as a ground, can be a true answer.

[18] Nor, of course, is it clear what such "accepting" and "inhibiting" could amount to. If my judging that a certain claim is unwarranted does not *constitute* my changing my mind, then it is unclear what steps I *could* take to change it.

(ii) believing that one believes that *p*,

(iii) believing that, all things considered, the evidence supports the proposition that *p*.

But if this were correct, then it would be quite baffling what entitled us to assume, as we ordinarily do without a second thought, that I can determine whether I believe that *p* by assessing evidence as to whether *p*, and that a belief about whether I believe that *p* formed on this basis will constitute knowledge of my own first-order belief as to whether *p*. It would be baffling, that is, how my being in a state of kind (iii) could justify my entering a state of kind (ii), and why the resulting state of kind (ii) should count as knowledge of whether I am in a state of kind (i). The only account of these connections available to Johnston, it seems, would be one that adverted to reliable tropisitic connections between states of these various kinds. But such an account would falsify the epistemology of this sort of self-knowledge: not only do we not normally regard a subject's entitlement to judge about her own belief that *p* on the basis of evidence that *p* as depending on the substantive assumption that there is a reliable correlation between her so judging and her believing that *p*, but, as noted earlier, we would regard a subject who *did* look for evidence of such a correlation as not in authority to speak about her own beliefs.

*

Having seen these connections between believing, knowing what one believes, and taking to be true, we can see more clearly why the ascription of beliefs to a subject must be governed by a "constitutive ideal of rationality," and why understanding certain kinds of flagrant violation of this ideal requires positing a partition in the subject's mind. Davidson is right to insist that to ascribe beliefs to a subject at all, we must normally be able to see a connection between what she believes and her own assessment of reasons; but this requirement of interpretive "charity" is merely the outward aspect of a fundamental fact about what it is to believe something and how this is connected with taking something to be evidence. The fundamental fact is this:

(B) What I believe is what I take to be true, and to judge that there are adequate grounds for a certain conclusion just is to take that conclusion to be true.

Given this fact about the nature of belief, there is simply no room in an undivided mind for directly conflicting attitudes like belief that *p* and recognition that the evidence warrants a different conclusion. This is not to deny that an undivided mind can tolerate all kinds of bad reasoning, unwarranted assumption, failure to recognize the implications of one's own beliefs, and so on. We can make sense of these kinds of irrationality precisely because we can take the subject not to recognize her own mistake. But when it is not simply a matter of failing to grasp the connection between propositions, or failing to see that some belief is open to question – when it is a matter of actually *recognizing* the unfavorable implications of some fact for what one believes – then this way out is not available.[19]

It is this basic connection between believing and taking to be true that makes the idea of self-deception so puzzling. For if we are to recognize cases of genuine, ongoing self-deception, we must find

---

[19] Some philosophers maintain that there can be "epistemically akratic" subjects who believe that *p* while recognizing that the balance of evidence points toward the conclusion that *p* is not true. (See, e.g., Heil 1984, Mele 1987a, Chapter 8, Scanlon 1998, pp. 35-36.) Would a case of epistemic akrasia constitute a counterexample to my claim that a subject who consciously takes a certain conclusion not to be warranted will not believe it? It is important to think carefully about what the examples of epistemic akrasia are supposed to be.

Often the example given is precisely of a person who is self-deceived, or else of a person who a psychoanalyst would describe as unconsciously believing what he consciously judges to be ill-grounded. Cases of epistemic akrasia that involve *unconscious* beliefs, however, do not tell against my claim as a remark about the structure we must find in the mind of a person who *consciously* believes something.

A more problematic kind of case would be one in which a person consciously recognizes that the evidence warrants one conclusion, but nevertheless consciously believes something else. Can we make sense of such a case? We can certainly conceive of a subject who recognizes that the balance of evidence points toward a certain conclusion but who does not believe that conclusion. This is intelligible so long as the subject does not take the available evidence to be conclusive. Thus a mother might know that the balance of available evidence points toward the conclusion that her missing child is dead, might acknowledge that another person in similar circumstances ought to accept this conclusion, and might nevertheless continue to believe that in *her* case the conclusion is wrong, perhaps sustaining the conviction that her case is the exception by contemplating scenarios in which the child turns up alive, etc. But it should be evident that such a case either does *not* involve the subject's believing that the evidence warrants another conclusion, or else involves something less than genuine belief that the missing son is alive. For on the one hand, if the subject believes that, despite the tendency of the available evidence to support one conclusion, another conclusion is actually the true one, then this is to say that she judges the available evidence *not* to warrant the conclusion that her child is dead. This may be wishful thinking on her part, but as described it is not counterexample to the claim that a subject who consciously takes a certain conclusion to be unwarranted cannot believe it. And on the other hand, if what we mean by saying that she believes her son is alive is that she has a recurring impulse to think this, no matter how well she understands that it is not so, then surely we should not say strictly speaking that she *believes* this. A fuller description would clarify that although her condition has some features of belief, her view of the matter lacks the kind of unity that would make an unqualified ascription of belief correct. We can of course describe her condition as a case of "akratic belief" if we like, but if we do we are using the modifier "akratic" in something like the way the modifier "plastic" is used in the construction "plastic fruit": a plastic fruit is not actually a fruit, although it resembles one.

subjects who meet the following description: they believe some proposition for reasons that they themselves know ("deep down") to be unsound or inadequate. But if what a person believes just is what he takes there to be sound reason for believing, then the suggestion that a subject believes a proposition for reasons he knows to be unsound or inadequate faces a dilemma. For what is the self-deceiver's relationship to his reasons for belief supposed to be? Does he really believe them, and believe that they demonstrate the proposition of which he aims to convince himself? In that case, although it may be true that he *is deceived*, it is hard to see how he can be accused of presently *deceiving* himself; for although his conviction may be wishful or foolish, nevertheless it seems that he really is convinced, and it is not clear what space is left for suspicions to the contrary. But on the other hand, if he does not believe the things he "tells himself," or does not believe that they show what he wishes they showed, then it is hard to see what his inward recital of grounds can ever achieve, how it can ever amount to his actually being *deceived*. In short, it is hard to see how there can be *room* for a state of self-deception or an activity of deceptively convincing oneself of something.[20]

What forces us to see the self-deceiver's mind as divided, then, is not the dispensable assumption that a self-deceiver must intend her own deception. It is the fact that, if one's mind has the normal kind of unity, then one's belief as to whether *p* just is the sum of one's thinking about whether *p* is true. If this is right, then the much-discussed question whether a self-deceiver must intend her own deception is a red herring. We can give up the claim that there must be some agency in the self-deceiver with the intent to deceive, and accept Johnston's idea that the forces that operate to lead the self-deceiver astray are tropistic. This will still leave us with a problem about what to make of the idea that the self-deceiver *at some level knows that her belief is not well-founded*. If we reject this sort of description as incoherent, we are

---

[20] A similar difficulty can be raised about the kind of self-misleading involved in "turning one's attention away from something." For does the subject appreciate that the thing he turns his attention away from speaks against his preferred belief, or does he not? If the former, how can he be deceived? If the latter, in what sense can he be said to be deceiving himself?

rejecting the kind of talk that provoked our interest in the first place: we can of course fix on something *else* to call "self-deception," but this will be to change the subject. But if we do not reject this sort of description, we must find some interpretation of the phrase "at some level" which leaves room for a subject to hold a belief "at one level" while regarding the reasons for that belief as inadequate "at another." And, as we have seen, it is hard to see how any interpretation that takes the self-deceiver's mind to have the normal kind of unity could have this effect.

6. MAKING SENSE OF MENTAL DIVISION

To say that a self-deceiver's mind cannot have the normal kind of unity, however, is not yet to explain what her mental division amounts to. We do not want to say that a self-deceiver literally suffers from multiple personality, or that she is like a person possessed by a demon. We want to say, rather, that *the very same subject* "at one level" believes what "at another level" she doubts. But what are we to make of this talk of different levels in a single subject? Let me conclude with a few remarks about this topic.

*

It will be helpful first to return to our question about why dogs can't deceive themselves. One striking thing about Johnston's tropistic account of self-deception is that it does not suggest any straightforward answer to this question. Presumably we could discover that dogs suffer from intelligible but nonrational mental tropisms which lead them to believe what would satisfy their desires and ignore indications to the contrary. We could of course call this "canine self-deception," but to suggest that it is the same phenomenon we find in the human case would, I think, be to overlook what is really interesting about the latter. For in the canine case there is nothing corresponding to obscure doubt, and consequently there is no pressure to see the mind of a canine "self-deceiver" as divided. If this is right, then one way to approach the question about what to make of the idea of a divided mind is to ask what

it is about a human mind that allows for this picture.

Davidson's account of irrationality suggests at least a schematic answer to this question. The answer is closest to the surface in his well-known paper on weakness of the will, where he argues that, if we are to make sense of the possibility of such weakness, we must picture a person as having not just two competing desires, but a *will* that chooses between them (see Davidson 1982b, pp. 35-6). What having a will comes to, on Davidson's account, is being able to make "all-things-considered judgments," judgments that reflect one's assessment of what course of action is best when all the available information is taken into account. A person's being weak-willed then consists in his intentionally acting in a manner inconsistent with his own all-things-considered judgment. Now, this suggests an explanation of why we do not take dogs to be capable of *akrasia*: namely, that we do not take them to be capable of making all-things-considered judgments at all. And although Davidson does not develop the point explicitly in the case of self-deception, it is easy to imagine a similar account of why we do not suppose that dogs can deceive themselves. The thought would be that a human being is capable of a kind of mental act which constitutes his taking a stand as to whether *p* is true, given the totality of evidence available to him. In this case, it would be pleonastic to call this his "all-things-considered judgment"; it is part of the purport of *any* judgment that it is true, and hence correct all-things-considered. Nevertheless, there is plainly a difference between a creature that can judge and one that merely acquires beliefs from perception, retains them in memory, and perhaps possesses various uncritical habits of "generalization" and "inference." A creature that can judge is one that can consider the question "Why should I think that *p*?", where that is a question of the grounds for taking *p* to be true. Such a creature can raise *for itself* the question of the soundness of its beliefs, and this requires that it be able to suspend belief, to consider propositions in the modalities of the possible and the hypothetical.[21]

---

[21] I borrow the final phrase from Brian O'Shaughnessy's helpful discussion of the difference between rational and

Inasmuch as we do not take dogs to be capable of this sort of reflection at all, we do not credit them with the capacity to believe something inconsistent with their own judgment either. But this, on the Davidsonian view, is a necessary condition for being self-deceived.

We want to say that, in the self-deceiver's case, the cart is pulling the horse: the self-deceiver does not believe what she does because she recognizes this and that as grounds for belief; she recognizes this and that as grounds because this is what she wishes to believe. Her mental condition has the *structure* of belief – there is the right kind of pattern of grounds and conclusions – but the explanation travels in the wrong direction: she notes these facts because they are convenient, and seeks grounds for questioning those ones because they point in an unwelcome direction. It is precisely here that Johnston's idea of a non-accidental but non-intentional mental tropism is helpful: the self-deceiver's "noting" and "seeking" are presumably to be understood in such terms. But such tropistic belief-formation only counts as *irrational* in the context of a capacity for self-conscious, *rational* belief-formation – for judgment – which is not to be understood tropistically. Lacking this capacity, a creature lacks the kind of mental complexity relative to which we can identify phenomena like "telling oneself" and "trying to convince oneself." Such a creature may believe something *for* a certain reason, and we may want to say that the relevant reason only weighed with the creature as a result of the biasing force of some powerful desire, but the critical capacity is absent that would make this lapse blameworthy.

*

What is attractive in Davidson's view is his recognition that self-deception is a pathology of judgment, something that is possible only in a mind with the kind of complexity that makes judgment possible. What is not attractive is his assumption that the only way to make self-deception intelligible is to suppose that the self-deceiver's mind contains two agencies, one of which intentionally misleads the other. Johnston has a battery of persuasive objections to this idea, but I will not repeat them. I take it

---

nonrational cognition in his recent book *Consciousness and the World* (2002, pp. 111-112).

that this literal conception of the self-deceiver's mental division is obviously too wooden, making self-deception look as if it must involve something like multiple personality. What holds Davidson's conception in place, though, is only the conviction that this is how we *must* understand self-deception if we are to acknowledge that it involves both belief and doubt. I want to suggest that we can concede this point, and concede that it forces us to conceive of the self-deceiver's mind as in a sense divided, without accepting Davidson's literal conception of what such division amounts to.

We are led to speak of the self-deceiver's mind as divided because we want to say that she at some level doubts the very thing she gets herself to believe. But why do we want to say this? Not because she expresses doubt in the ordinary way – namely, by being ready to acknowledge it and take it into account. Our reason for saying that the self-deceiver doubts her own belief, presumably, is that she goes to such lengths to prop it up and to avoid lines of inquiry that would bring it into question. In other words, when we say that the self-deceiver "at some level" doubts what she herself believes, this ascription of doubt is *not* governed by the standards that in the first instance govern our ascription of attitudes like belief and doubt. For our normal standards of belief-ascription recognize the entitlement expressed in (T): they take the question what a subject believes to be conceptually linked to her own assessment of grounds. If we sever this connection, as we do when we speak of a self-deceiver as "at some level" doubting her own belief, we are using the term "doubting" in a new way – a way that may come naturally, but whose structure needs investigation.

Davidson's mistake is to suppose that, when we say that the self-deceiver harbors a doubt about the adequacy of her grounds, we are positing another belief on a par with the belief those grounds support: the belief that a certain proposition is not well-founded. This is what forces the literal understanding of the self-deceiver's mental division: if we suppose that the self-deceiver holds two full-fledged beliefs, each governed by the standards associated with principle (T), then it seems we must suppose that there are two sub-minds, each capable for its own part of holding beliefs on the basis of an

assessment of grounds. Once we clarify our reasons for saying that the self-deceiver harbors persistent doubt, however, we can see that the self-deceiver's conscious belief and her unconscious doubt are *not* on a par: whereas the conscious belief is ascribed according to ordinary criteria, the underlying doubt is ascribed on a different basis, one that *presupposes* the (qualified) aptness of the first ascription. To call a person self-deceived is to say, in the first place, that, judged according to normal standards, her belief is that *p*. But it is at the same time to suggest that these normal standards apply imperfectly in the case at hand: that this individual is not displaying enough unity in her speech and behavior to support our ordinary way of ascribing beliefs to a rational subject.

Fundamentally, what the self-deceiver is guilty of is a kind of abdication of epistemic responsibility. She finds reasons to believe what she wants to believe, but she systematically and motivatedly fails to exercise her capacity to reflect on the soundness of those reasons. But – as I have emphasized throughout this dissertation – a rational creature's attitude toward the proposition that *p* is one of belief only insofar as she holds the attitude in a manner that is open to reflection on reasons, open to the question "Why do I think that *p* is true?" And this openness is required not only in regard to the belief justified, but in regard to the grounds produced as justifiers. To the extent that we find a person guilty of a systematic failure to exercise her capacity for self-criticism, therefore, we have reason to say that a whole interconnected network of her apparent beliefs are actually something less than beliefs. To say that she at some level doubts these beliefs is to picture this situation as involving a divided mind. But the aptness of this *picture* does not require that there be distinct subsystems in the person with different beliefs; its aptness is that the subject in one sense believes what in another sense she does not quite believe. And the paradoxicality of this situation consists in the fact that, if these two senses come entirely apart, then there is nothing left to call "belief."

BIBLIOGRAPHY*

Ameriks, Karl. 1994. "Understanding Apperception Today." In *Kant and Contemporary Epistemology,* ed. Paolo Parrini. Dordrecht: Kluwer Academic Publishers.
_____. 2000. *Kant's Theory of Mind.* New Edition. Oxford: Oxford University Press.

Anscombe, G. E. M. 1963. *Intention.* Second Edition. Oxford: Basil Blackwell.
_____. 1975. "The First Person." In *Mind and Language*, ed. Samuel Guttenplan. Oxford: Oxford University Press. Reprinted in Cassam 1994.

Aristotle. 1984. *The Complete Works of Aristotle.* 2 Vols. Ed. Jonathan Barnes. Princeton: Princeton University Press.

Armstrong, D. M. 1968. *A Materialist Theory of the Mind.* New York: Humanities Press.

Augustine. 1991. *Confessions.* Trans. Henry Chadwick. Oxford: Oxford University Press.

Bar-On, Dorit and Douglas C. Long. 2001. "Avowals and First-Person Privilege." *Philosophy and Phenomenological Research*, 62, pp. 311-335.

Barnes, Annette. 1997. *Seeing through Self-Deception.* Cambridge: Cambridge University Press.

Bennett, Jonathan. 1964. *Rationality.* New York: Humanities Press.
_____. 1966. *Kant's Analytic.* Cambridge: Cambridge University Press.
_____. 1974. *Kant's Dialectic.* Cambridge: Cambridge University Press.

Bilgrami, Akeel. 1998. "Self-Knowledge and Resentment." In Wright, Smith and MacDonald 1998.

Brandom, Robert B. 1994. *Making It Explicit.* Cambridge: Harvard University Press.
_____. 1995. "Knowledge and the Social Articulation of the Space of Reasons." *Philosophy and Phenomenological Research*, 55, pp. 895-908.
_____. 2000. "Insights and Blindspots of Reliabilism." In his *Articulating Reasons.* Cambridge: Harvard University Press.

Brewer, Bill. 1999. *Perception and Reason.* Oxford: Clarendon Press.

Brook, Andrew. 1994. *Kant and the Mind.* Cambridge: Cambridge University Press.

Burge, Tyler. 1996. "Our Entitlement to Self-Knowledge." *Proceedings of the Aristotelian Society*, 96.
_____. 1998. "Reason and the First Person." In Wright, Smith and MacDonald 1998.
_____. 2003. "Perceptual Entitlement." *Philosophy and Phenomenological Research*, 67, pp. 503-548.

* Note: Where a reprint is cited, page citations in the dissertation are to the reprint.

Canfield, John V. and Don F. Gustavson. 1962. "Self-Deception." *Analysis*, 23, pp. 32-36.

Cassam, Quassim, ed. 1994. *Self-Knowledge.* Oxford: Oxford University Press.

Cassam, Quassim. 1997. *Self and World.* Oxford: Oxford University Press.

Cavell, Stanley. 1979. *The Claim of Reason.* Oxford: Oxford University Press.

Champlin, T. S. 1988. *Reflexive Paradoxes.* London: Routledge.

Chater, Nick and Cecilia Heyes. 1994. "Animal Concepts: Content and Discontent." *Mind & Language*, 9, pp. 209-246.

Davidson, Donald. 1980a. *Essays on Actions and Events.* Oxford: Oxford University Press.
_____. 1980b. "How Is Weakness of the Will Possible?" In Davidson 1980a.
_____. 1980c. "Mental Events." In Davidson 1980a.
_____. 1982a. "Paradoxes of Irrationality." In *Philosophical Essays on Freud*, ed. James Hopkins and Richard Wollheim. Cambridge: Cambridge University Press.
_____. 1982b. "Rational Animals." *Dialectica*, 36, pp. 317-327.
_____. 1984a. "First Person Authority." *Dialectica*, 38, pp. 101-11.
_____. 1984b. "Thought and Talk." In his *Inquiries into Truth and Interpretation.* Oxford: Oxford University Press.
_____. 1985a. "Deception and Division." In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore & Brian McLaughlin. Oxford: Basil Blackwell.
_____. 1985b. "Incoherence and Irrationality." *Dialectica*, 39, pp. 345-54.

Descartes, René. 1984. *The Philosophical Writings of Descartes.* 2 Vols. Trans. John Cottingham, Robert Stootoff and Dugald Murdoch. Cambridge: Cambridge University Press.

Dretske, Fred. 1981. *Knowledge and the Flow of Information.* Cambridge: MIT Press.
_____. 1986. "Misrepresentation." In *Belief: Form, Content, and Function.* Ed. Radu J. Bogdan. Oxford: Clarendon Press.

Edgley, Roy. 1969. *Reason in Theory and Practice.* London: Hutchison.

Eilan, Naomi and Johannes Roessler, eds. 2003. *Agency and Self-Awareness.* Oxford: Oxford University Press.

Engstrom, Stephen. Unpublished. "Kant on Objective Validity, Truth, and Judgment."

Evans, Gareth. 1982. *The Varieties of Reference.* Oxford: Oxford University Press.
_____. 1985a. *Collected Papers.* Oxford: Oxford University Press.
_____. 1985b. "Semantic Theory and Tacit Knowledge." In Evans 1985a.
_____. 1985c. "Things without the Mind—A Commentary upon Chapter Two of Strawson's *Individuals*." In Evans 1985a.

Falvey, Kevin. 2000. "The Basis of First-Person Authority." *Philosophical Topics*, 28, pp. 69-99.

Finkelstein, David H. 2003. *Expression and the Inner.* Cambridge: Harvard University Press.

Frege, Gottlob. 1997. "Thought." In *The Frege Reader*, ed. Michael Beaney. Oxford: Blackwell Publishers.

Fricker, Elizabeth. 1998. "Self-Knowledge: Special Access versus Artefact of Grammar —A Dichotomy Rejected." In Wright, Smith and MacDonald 1998.

Gardiner, Patrick. 1976. "Error, Faith, and Self-Deception." In *The Philosophy of Mind*, ed. Jonathan Glover. Oxford: Oxford University Press.

Geach, Peter. 1957. *Mental Acts*. New York: Humanities Press.
_____. 1960. "Ascriptivism." *Philosophical Review*, 69, pp. 221-225.
_____. 1965. "Assertion." *Philosophical Review*, 74, pp. 449-465.

Goldman, Alvin I. 1976. "Discrimination and Perceptual Knowledge." *Journal of Philosophy*, vol. 73, no. 20, pp. 771-791.

Guyer, Paul. 1980. "Kant on Apperception and *a priori* Synthesis." *American Philosophical Quarterly*, 17, pp. 205-212.

Hampshire, Stuart. 1965. *Freedom of the Individual*. Princeton: Princeton University Press.

Harman, Gilbert. 1973. *Thought*. Princeton: Princeton University Press.
_____. 1986. *Change in View*. Cambridge: MIT Press.

Heck, Richard G. 2000. "Nonconceptual Content and the 'Space of Reasons'." *The Philosophical Review*, vol. 109, no. 4, pp. 483-523.

Hegel, G. W. F. 1975. *Hegel's Logic* (Part I of the *Encyclopedia of the Philosophical Sciences*). Trans. William Wallace. Oxford: Oxford University Press.

Heil, John. 1983. "Doxastic Agency." *Philosophical Studies*, vol. 43, pp. 355-364.
_____. 1984. "Doxastic Incontinence." *Mind*, 93, pp. 56-70.

Henrich, Dieter. 1971. "Self-Consciousness: A Critical Introduction to a Theory." *Man and World*, vol. 4, pp. 3-28.
_____. 1982. "Fichte's Original Insight." *Contemporary German Philosophy*, vol. 1, pp. 15-53.

Hume, David. 1978. *A Treatise of Human Nature*. Second Edition. Ed. L.A. Selby-Bigge and P.H. Nidditch. Oxford: Clarendon Press.

Hurley, Susan L. 1998. *Consciousness in Action*. Cambridge: Harvard University Press.
_____. 2001. "Overintellectualizing the Mind." *Philosophy and Phenomenological Research*, 63, pp. 423-431.

Johnston, Mark. 1988. "Self-Deception and the Nature of Mind." In *Perspectives on Self-Deception*, ed. Brian P. McLaughlin & Amélie Oksenberg Rorty. Berkeley: University of California Press.

Kahneman, D., P. Slovic and A. Tversky, eds. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kant, Immanuel. 1900. *Gesammelte Schriften*. 29 vols. Ed. Royal Prussian Academy of Sciences. Berlin: Georg Reimer.
_____. 1928. *Critique of Judgement*. Trans. James Creed Meredith. Oxford: Oxford University Press.
_____. 1929. *Critique of Pure Reason*. Trans. Norman Kemp Smith. New York: St. Martin's Press.

_____. 1950. *Prolegomena to Any Future Metaphysics*. Trans. Lewis White Beck. New York: Macmillan Publishing Company.
_____. 1974a. *Anthropology from a Pragmatic Point of View*. Trans. Mary J. Gregor. The Hague: Martinus Nijhoff.
_____. 1974b. *Logic*. Trans. Robert S. Hartman and Wolfgang Schwarz. Indianapolis: Bobbs-Merrill.
_____. 1993. *Grounding for the Metaphysics of Morals*. Third Edition. Trans. James W. Ellington. Indianapolis: Hackett Publishing Company.
_____. 1996a. *Critique of Practical Reason*. Trans. Mary J. Gregor. In Immanuel Kant, *Practical Philosophy*. Cambridge: Cambridge University Press.
_____. 1996b. *Critique of Pure Reason*. Trans. Werner J. Pluhar. Indianapolis: Hackett Publishing Company.
_____. 1997a. *Critique of Pure Reason*. Trans. Paul Guyer & Allen W. Wood. Cambridge: Cambridge University Press.
_____. 1997b. *Lectures on Metaphysics*. Ed. & Trans. Karl Ameriks & Steve Naragon. Cambridge: Cambridge University Press.
_____. 1999. *Immanuel Kant: Correspondence*. Ed. & Trans. Arnulf Zweig. Cambridge: Cambridge University Press.

Kaplan, David. 1989. "Demonstratives." In *Themes from Kaplan*, ed. Joseph Almog. Oxford: Oxford University Press.

Kemp Smith, Norman. 1918. *A Commentary to Kant's Critique of Pure Reason*. London: Macmillan.

Keller, Pierre. 1998. *Kant and the Demands of Self-Consciousness*. Cambridge: Cambridge University Press.

Kitcher, Patricia. 1982. "Kant on Self-Identity." *Philosophical Review*, vol. 91. pp. 41-72.
_____. 1990. *Kant's Transcendental Psychology*. Oxford: Oxford University Press.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.
_____. 2002. "Self-Constitution: Action, Identity and Integrity." Given as the John Locke Lectures at Oxford University. Available at: *http://www.people.fas.harvard.edu/~korsgaar/#Publications*.

Lazar, Ariela. 1999. "Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'." *Mind*, 108, pp. 265-290.

Lehrer, Keith. 1979. "The Gettier Problem and the Analysis of Knowledge." In *Justification and Knowledge*. Ed. George S. Pappas. Dordrecht: D. Reidel.

Leibniz, Gottfried Wilhelm. 1991. *Monadology*. Ed. & Trans. Nicholas Rescher. Pittsburgh: University of Pittsburgh Press.

Lewis, David. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, 50, 3, pp. 249-258.

Locke, John. 1975. *An Essay concerning Human Understanding*. Ed. P. H. Niddich. Oxford: Oxford University Press.

Lycan, William G. 1998. "Consciousness as Internal Monitoring." In *The Nature of Consciousness,* ed. Ned Block, Owen Flanagan and Güven Güzeldere. Cambridge: MIT Press.

Malcolm, Norman. 1977. "Thoughtless Brutes." In his *Thought and Knowledge*. Ithaca, New York: Cornell University Press.

McDowell, John. 1985. "Functionalism and Anomalous Monism." In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Ed. Ernest LePore and Brian McLaughlin. Oxford: Blackwell.
_____. 1994a. "The Content of Perceptual Experience." *The Philosophical Quarterly,* 44, pp. 190-205.
_____. 1994b. *Mind and World*. Cambridge: Harvard University Press.
_____. 1995. "Knowledge and the Internal." *Philosophy and Phenomenological Research*, 55, pp. 877-93.
_____. 1998a. "Referring to Oneself." In *The Philosophy of P.F. Strawson*, ed. Lewis E. Hahn. Chicago: Open Court Publishing Company.
_____. 1998b. "The Woodbridge Lectures 1997: Having the World in View." *Journal of Philosophy*, 95, pp. 431-491.

McGinn, Colin. 1997. *The Character of Mind*. Revised Edition. Oxford: Oxford University Press.

Mele, Alfred R. 1987a. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford: Oxford University Press.
_____. 1987b. "Recent Work on Self-Deception." *American Philosophical Quarterly*, 24, 1, pp. 1-17.
_____. 1997. "Real Self-Deception." *Behavioral and Brain Sciences*, 20, pp. 91-102.
_____. 2001. *Self-Deception Unmasked*. Princeton: Princeton University Press.

Moran, Richard. 1994. "Interpretation Theory and the First Person." *Philosophical Quarterly*, 44, pp. 154-73.
_____. 2001. *Authority and Estrangement*. Princeton: Princeton University Press.
_____. 2003. "Responses to O'Brien and Shoemaker." *European Journal of Philosophy*, 11, pp. 402-419.

Moravcsik, Julius. 1994. "Essences, Powers, and Generic Propositions." In *Unity, Identity, and Explanation in Aristotle's Metaphysics*, ed. T. Scaltsas, D. Charles, and M. L. Gill. Oxford: Oxford University Press.

Naragon, Steve. 1990. "Kant on Descartes and the Brutes." *Kant-Studien*, 81, pp. 1-23.

Nichols, Shaun and Stephen P. Stich. 2003. *Mindreading*. Oxford: Clarendon Press.

Nisbett, R. E. and L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, New Jersey: Prentice-Hall.

Noë, Alva. 2002. "Is Perspectival Self-Consciousness Nonconceptual?" *Philosophical Quarterly*, 52, 207, pp. 185-194.

O'Brien, Lucy. 2003. "Moran on Agency and Self-Knowledge." *European Journal of Philosophy*, 11, pp. 375-390.

O'Shaughnessy, Brian. 2002. *Consciousness and the World*. Oxford: Oxford University Press.

Pears, David. 1984. *Motivated Irrationality*. Oxford: Clarendon Press.

Powell, C. Thomas. 1990. *Kant's Theory of Self-Consciousness*. Oxford: Clarendon Press.

Quine, W. V. 1960. *Word and Object*. Cambridge: MIT Press.

Rödl, Sebastian. 1998. *Selbstbezug und Normativität*. Paderborn: Ferdinand Schöningh.
_____. Forthcoming. *First-Person Order*. Cambridge: Harvard University Press.

Rosenthal, David M. 1986. "Two Concepts of Consciousness." *Philosophical Studies,* 49, pp. 329-59.

Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Barnes & Noble Books.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge: Harvard University Press.

Scott-Kakures, Dion. 2002. "At 'Permanent Risk': Reasoning and Self-Knowledge in Self-Deception." *Philosophy and Phenomenological Research*, 65, pp. 576-603.

Sellars, Wilfrid. 1963. "Empiricism and the Philosophy of Mind." In *Science, Perception and Reality*. Atascadero, California: Ridgeview Publishing Company.
_____. 1967. *Science and Metaphysics*. Atascadero, CA: Ridgeview Publishing Company.
_____. 1974. "Some Remarks on Kant's Theory of Experience." In his *Essays in Philosophy and Its History*. Boston: D. Reidel Publishing Company.

Shah, Nishi and J. David Velleman. 2004. "Doxastic Deliberation." Unpublished manuscript available at *http://www-personal.umich.edu/~velleman/Work/Doxastic.pdf*.

Shoemaker, Sydney. 1963. *Self-Knowledge and Self-Identity*. Ithaca: Cornell University Press.
_____. 1988. "On Knowing One's Own Mind." *Philosophical Perspectives*, 2, pp. 183-209.
_____. 1990. "First-Person Access." *Philosophical Perspectives*, 4, pp. 187-214.
_____. 1991. "Rationality and Self-Consciousness." In *The Opened Curtain: A U.S.-Soviet Philosophy Summit*, ed. Keith Lehrer and Ernest Sosa. San Francisco: Westview Press.
_____. 2003. "Moran on Self-Knowledge." *European Journal of Philosophy*, 11, pp. 391-401.

Stich, Stephen P. 1985. "Could Man Be an Irrational Animal?" *Synthese*, 64, pp. 115-135.

Strawson, P. F. 1959. *Individuals*. London: Methuen & Company.
_____. 1966. *The Bounds of Sense*. London: Methuen & Company.

Thompson, Michael. 1995. "The Representation of Life." In *Virtues and Reasons: Philippa Foot and Moral Theory*, ed. Rosalind Hursthouse, Gavin Lawrence, and Warren Quinn. Oxford: Oxford University Press.

Tomasello, Michael and Josep Call. 1997. *Primate Cognition*. Oxford: Oxford University Press.

Tugendhat, Ernst. 1982. *Traditional and Analytical Philosophy* (English translation of *Vorlesungen zur Einführung in die sprachanalytische Philosophie*). Trans. P. A. Gorner. Cambridge: Cambridge University Press.
_____. 1986. *Self-Consciousness and Self-Determination*. Trans. Paul Stern. Cambridge: MIT Press.

Weiskrantz, L. 1986. *Blind Sight*. Oxford: Oxford University Press.

Williams, Bernard. 1973. "Deciding to Believe." In his *Problems of the Self*. Cambridge: Cambridge University Press.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Trans. G. E. M. Anscombe. New York: Macmillan Publishing Company.

_____. 1967. *Zettel*. Ed. G. E. M. Anscombe and G. H. von Wright. Trans. G. E. M. Anscombe. Oxford: Basil Blackwell.
_____. 1980. *Remarks on the Philosophy of Psychology*. 2 Vols. Ed. G. E. M. Anscombe and G. H. von Wright. Trans. G. E. M. Anscombe. Oxford: Basil Blackwell.

Wright, Crispin. 1987. "On Making Up One's Mind: Wittgenstein on Intention." In *Logic, Philosophy of Science and Epistemology: Proceedings of the 11th International Wittgenstein Symposium*, ed. Paul Weingartner & Gerhard Schurz. Vienna: Hölder-Pichler-Tempsky.
_____. 1991. "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention." In *Meaning Scepticism*, ed. Klaus Puhl. New York: De Gruyter.
_____. 1998. "Self-Knowledge: The Wittgenstein Legacy." In Wright, Smith and MacDonald 1998.

Wright, Crispin, Barry C. Smith and Cynthia MacDonald, eds. 1998. *Knowing Our Own Minds*. Oxford: Clarendon Press.