PHILOSOPHICAL TOPICS
VOL. 33, NO. 1, SPRING 2005

Introspection

Alex Byrne
Massachusetts Institute of Technology

> The word introspection need hardly be defined—it means, of course, the looking into our own minds and reporting what we there discover.
>
> —James, *The Principles of Psychology*

> 'Introspection' is a term of art and one for which little use is found in the self-descriptions of untheoretical people.
>
> —Ryle, *The Concept of Mind*

1. INTRODUCTION

I know various contingent truths about my environment by perception. For example, by looking, I know that there is a computer before me; by hearing, I know that someone is talking in the corridor; by tasting, I know that the coffee has no sugar. I know these things because I have some built-in mechanisms specialized for detecting the state of my environment. One of these mechanisms, for instance, is presently transducing electromagnetic radiation (in a narrow band of wavelengths) coming from the computer and the desk on which it sits. How that mechanism works is a complicated story—to put it mildly—and of course much remains unknown. But we can at least produce more-or-less plausible sketches of how the mechanism can start from retinal irradiation, and go on to deliver knowledge of my surroundings. Moreover, in the sort of world we inhabit, specialized detection

mechanisms that are causally affected by the things they detect have no serious competition—seeing the computer by seeing an idea of the computer in the divine mind, for example, is not a feasible alternative.

In addition to these contingent truths about my environment, I also know various contingent truths about my psychology. For example, I know that I *see* a computer, that I *believe* that there is someone in the corridor, that I *prefer* coffee without sugar. How do I do know these things? Well, unless it's magic, I must have some sort of mechanism (perhaps more than one) for detecting my own mental states—something rather like my visual, auditory, and gustatory systems, although directed to my mental life. That is, I have knowledge of my mental life by a special kind of *perception*, or,

> a little more cautiously, . . . something that resembles perception. But unlike *sense*-perception, it is not directed towards our current environment and/or our current bodily state. It is perception of the mental. Such "inner" perception is traditionally called introspection, or introspective awareness. (Armstrong 1981, 60; see also Armstrong 1968, ch. 15; Lycan 1987, ch. 6; 1996, ch. 2; Nichols and Stich 2003, 160–64)

This *inner-sense theory* sounds like enlightened common sense; as Shoemaker remarks, it "can seem a truism" (1994, 223). However, it is not infrequently taken to be a crass mistake.[1]

The main point of this paper is that the proponents and opponents of the inner-sense theory should split the difference. There *is* a mechanism for detecting one's mental states but—as will be explained later—in an important respect it does *not* "resemble perception."

The positive account will come at the end. The next section notes two features of self-knowledge that any theory should explain.

2. PRIVILEGED AND PECULIAR ACCESS

Self-knowledge is often contrasted with knowledge of the mental states of others in the following two ways. First, knowledge of one's mental states is *privileged* in comparison to knowledge of others' minds. Roughly: beliefs about one's mental states acquired through the usual route are more likely to amount to knowledge than beliefs about others' mental states (and, more generally, beliefs about one's environment). At any rate, knowledge of one's own mental state is more likely when the state is neither factive nor object-entailing. One may well falsely believe that the cat is indoors; hence one may well falsely believe that one *knows* that the cat is indoors or *sees* that the cat is indoors. Similarly, one may well falsely believe that one sees *the cat*. But it is harder to err in believing that one *believes* that the cat is indoors, or that it *looks to one* that the cat is indoors.

To say that we have privileged access is not to say that beliefs about one's present mental states always amount to knowledge. Such beliefs need not even be true. One can falsely believe that one is angry, that one wants a beer, that one believes that one is happy, for example. More controversially, one can even falsely believe that it looks to one that something is red, or that one has a headache. Nonetheless, although error may always be a possibility, in a typical situation it is easier to be right about one's (non-factive, non-object-entailing) mental states (that one believes that the cat is indoors, say) than about the mental states of another (that Fred believes that the cat is indoors), or the corresponding tract of one's environment (that the cat is indoors).[2]

Second, knowledge of one's mental states is *peculiar* in comparison to one's knowledge of others' minds. One has a special method or way of knowing that one believes that the cat is indoors, that one sees the cat, that one intends to put the cat out, and so on, which one cannot use to discover that someone *else* is in the same mental state.

Our access to others' minds is importantly similar to our access to the nonpsychological aspects of our environment: one can come to know that the cat is indoors by seeing that it is, and one can likewise come to know that the cat wants to be fed and that Fred wants the sushi deluxe. Our peculiar access to our own minds is not like this: one can come to know that one wants the sushi deluxe without observing oneself at all.

Privileged and peculiar access can come apart.[3] Behaviorists typically hold that one has access to one's own mind in the same way that one has access to others' minds—by observing behavior. Yet a behaviorist might well agree that one has privileged access to one's own mind, simply because one is typically much better positioned than others to observe one's behavior.[4]

Thus Ryle:

> The superiority of the speaker's knowledge of what he is doing over that of the listener does not indicate that he has Privileged Access to facts of a type inevitably inaccessible to the listener, but only that he is in a very good position to know what the listener is in a very poor position to know. The turns taken by a man's conversation do not startle or perplex his wife as much as they had surprised and puzzled his fiancée, nor do close colleagues have to explain themselves to each other as much as they have to explain themselves to their new pupils. (1949, 171)[5]

Conversely, imagine a proponent of inner sense who holds that one's "inner eye" is very unreliable by comparison with one's outer eyes. The psychologist Karl Lashley likened introspection to astigmatic vision, claiming that "[t]he subjective view is a partial and distorted analysis" (1923, 338).[6] On this account, we have peculiar but *under*privileged access.

The inner-sense theory does offer a nice explanation of *peculiar* access: for obvious architectural reasons, the (presumably neural) mechanism of inner sense

is only sensitive to the subject's own mental states. In exactly the same style, our faculty of proprioception explains the "peculiar access" we have to the position of our own limbs.

Self-knowledge is a large topic. To keep things manageable, the focus—following the philosophers discussed in the next three sections—will be on knowledge of one's *beliefs*. The final section briefly widens the view.

3. SELF-KNOWLEDGE AS SELF-CONSTITUTION

This section reinforces the initial suspicion that the inner-sense theory must be right through an examination of a notable recent alternative, presented in Moran's subtle and original *Authority and Estrangement*.[7] A main theme of that book is that the problem of self-knowledge is misleadingly conceived as one of "epistemic access (whether quasi-perceptual or not) to a special realm" (Moran 2001, 32). In that respect, self-knowledge is unlike mathematical knowledge, knowledge of others' minds, knowledge of the past, and so on. The problem is as much one of moral psychology as it is of epistemology: we must think of "[t]he special features of first-person access...in terms of the special responsibilities the person has in virtue of the mental life in question being *his own*" (32).

Moran's account gives a central role to the "transparent" nature of belief, as expressed in the following well-known passage from Evans:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?," I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" (Evans 1982, 225)[8]

Moran first formulates Evans's observation as a "claim of transparency":

> With respect to the attitude of belief, the claim of transparency tells us that the first-person question "Do I believe P?" is "transparent" to, answered in the same way as, the outward-directed question as to the truth of P itself. (Moran 2001, 66)

However, as Moran notes, sometimes the question "Do I believe P?" is *not* transparent in this way, for instance in "various familiar therapeutic contexts" (85). So the correct formulation of Evans's observation is that one can *typically* answer the question "Do I believe P?" simply by considering whether P is true. In Moran's terminology, an answer to the question "Do I believe P?" typically obeys the "Transparency Condition":

> A statement of one's belief about X is said to obey the Transparency Condition when the statement is made by consideration of the facts

> about X itself, and not by either an "inward glance" or by observation of one's own behavior. (101)

The qualifications about no "inward glance," or observation of one's behavior, can be glossed as follows. Concluding that P is true is sufficient for one justifiably to claim that one believes P: no *additional* evidence—a fortiori evidence about oneself—is required.

According to Moran, transparency shows that arriving at self-knowledge (specifically, knowledge of one's beliefs) is not accurately viewed as a process of *self-discovery*, but rather as a process of *self-constitution*. Coming to know whether one believes P is not a matter of taking a "theoretical" or disinterested stance toward oneself, of the sort one adopts toward another person when his beliefs are the subject matter of inquiry. Rather, it is a matter of "making up one's mind" as to the truth of P. Further, transparency explains "the special features of first-person knowledge (roughly . . . the immediacy, authority, and special relation to rationality of ordinary self-knowledge)" (Moran 2003, 410).

Moran's argument from transparency to the self-constitution thesis makes use of a distinction between "theoretical" and "practical or deliberative" questions:

> a *theoretical* question about oneself. . .is one that is answered by a *discovery of the fact of which one was ignorant*, whereas a *practical or deliberative* question is answered by a *decision or commitment of some sort* and it is not a response to ignorance of some antecedent fact about oneself. (Moran 2001, 58, my emphasis)

And:

> a *'deliberative'* question about one's state of mind . . . [is] a question that is answered by *making up one's mind*, one way or the other, coming to some resolution. (Moran 2003, 404, my emphasis)

For example, distinguish two sorts of situations in which one might ask the question "What will I wear?" (see Moran 2001, 56; cf. Anscombe 1963, § 2). First, one is preparing to get dressed for the annual philosophy department party. Second, one has just been sentenced to five years for embezzling the philosophy department funds, and has yet to be issued with standard prison clothing. In the first case, the question calls for a *decision*: one considers the sartorial pros and cons, and selects the purple tie. In the second case, the question is answered by a *discovery*: the judge announces that prisoners in Massachusetts wear orange jumpsuits.

As this example shows, the distinction is not strictly speaking one between *questions*—ignoring temporal complications, it is the *same* question both times—but rather between ways of answering questions. And, indeed, Moran later writes of answering a question in "deliberative or theoretical spirit," taking a "deliberative or theoretical stance" to a question, and so forth.[9] When one answers a question "Am I F?" in a *deliberative* spirit, one engages in practical or theoretical reasoning whose outcome (a belief in the theoretical case, an action/intention in the practical

case) determines either that one is F, or that one is not F. That is, the outcome of one's reasoning determines the answer. When one answers a question "Am I F?" in a *theoretical* spirit, one engages in theoretical reasoning whose outcome is simply to uncover the answer, not to determine it.

The distinction applies to questions like "Do I believe P?" One might address this question in a theoretical spirit, treating it "as a more or less purely psychological question about a certain person, as one may enquire into the beliefs of someone else" (Moran 2001, 67). Alternatively, one might address this question in a deliberative spirit, as a matter of making up one's mind about P. Take, for example, the question "Do I believe Alice is a threat to my career?" as asked by her colleague Bert. After looking back over his behavior toward Alice—anonymously rejecting one of Alice's papers that criticizes Bert's pet theory, etc., etc.—Bert might conclude that he has this belief. Alternatively, Bert might address the question in a deliberative spirit, and investigate whether Alice really is a threat to Bert's career. Perhaps the result of the investigation is that Alice is harmless, and Bert thereby concludes that he believes that Alice is not a threat. We can imagine Bert addressing the question in both a deliberative and theoretical spirit, raising the uncomfortable possibility of discovering that he has inconsistent beliefs.[10]

Here is how Moran links the "deliberative/theoretical" distinction with transparency:

> With respect to belief, the claim of transparency is that from within the first-person perspective, I treat the question of my belief about P as equivalent to the question of the truth of P. What I think we can see now is that the basis for this equivalence hinges on the role of deliberative considerations about one's attitudes. For what the "logical" claim of transparency requires is the deferral of the theoretical question "What do I believe?" to the deliberative question "What am I to believe?". And in the case of the attitude of belief, answering a deliberative question is a matter of determining what is true.
>
> When we unpack the idea in this way, we see that the vehicle of transparency in each case lies in the requirement that I address myself to the question of my state of mind in a *deliberative* spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the beliefs of someone who happens also to be me. (63)

Suppose I ask myself "Do I believe P?" and that I answer "I believe P" by determining that P is true. Then, according to Moran, I have answered this question by "a decision or commitment of some sort," and not "by a discovery of the fact of which I was ignorant." Transparency shows, in other words, that knowledge that one believes P, when arrived at by considering whether P is true, is a matter of "making up one's mind" that P is true.

However, Moran's conclusion is overdrawn. It is true that often one answers the question "Do I believe P?" in a deliberative spirit. It is natural to imagine this happening with Evans's question "Do you think there is going to be a third world

war?" One has not previously considered the likelihood of a third world war; one studies the relevant geopolitical facts, and makes up one's mind. But, precisely because it suggests this sort of context, in this respect Evans's example is misleading. Consider the question "Do I believe that I live in Cambridge, Massachusetts?" or "Do I believe that Moran is the author of *Authority and Estrangement*?" These questions can be answered transparently, by considering the relevant facts of location and authorship, but I do not need to make up my mind.[11] On the contrary, it is already made up. I have believed for some time that I live in Cambridge, and that Moran is the author of *Authority and Estrangement*. I can know that I believe I live in Cambridge, for example, by remembering the nonpsychological fact that I live in Cambridge.

So transparency does *not* show that knowledge of one's beliefs is in general a matter of making up one's mind. And there are further reasons to be suspicious of any tight connection between transparency and the thesis that self-knowledge involves self-constitution. Moran concentrates almost exclusively on the transparency of belief, but perception provides other examples (as noted in Evans 1982, 224–25; see also Dretske 1995, 2003). One can know that one sees the cat by an "outward look" at the cat. One determines that the cat is there, and concludes that one sees the cat. However, *seeing the cat* is not in *any* sense a matter of making up one's mind, or "coming to some resolution"—one can see the cat without having any beliefs about it. When one comes to know that one sees the cat by looking at the cat, one has simply discovered "some antecedent fact about oneself."

Although the significance of Moran's special cases should not be overlooked, self-knowledge of mental states in general (or even of beliefs in particular), with its distinctive features of privileged and peculiar access, cannot possibly be explained in terms of "self-constitution." Moran's emphasis on *transparency* is quite another matter, however, and much will be made of that later.[12]

4. AGAINST INNER SENSE I: PRELIMINARY SKIRMISHINGS

The inner-sense theory—at least when left at a high level of abstraction—is of considerable initial appeal. Given its unpopularity, one might expect the standard arguments against it to be decisive refutations, or near enough. This section and the next try to dampen any expectation down.

4.1 INNER SENSE IS UNLIKE PARADIGMATIC PERCEPTION

The metaphor of an "inner eye" should not be taken very seriously—and not just because of the darkness inside the brain. For example, there are visual experiences but no *introspective* experiences:

> . . . on nobody's view is the awareness of one's headache mediated by an appearance *of* the headache. And in the case of attitudes like belief, there

> is simply nothing quasi-experiential in the offing to begin with. (Moran 2001, 14)

Moran also mentions the apparent absence of any organ of introspection (13). Shoemaker (1994) presents a comprehensive tabulation of the disanalogies between inner sense and paradigms of "outer sense." According to the *object perception model*, as Shoemaker calls it, introspection is like ordinary (visual) perception in the following four respects: it typically involves (a) "awareness of facts ... by means of awareness of objects"; (b) "'identification information' about the object of perception"; (c) "perception of...intrinsic, nonrelational properties"; and (d) potential for the objects of perception to be "objects of attention" (1994, 205–6). Since (a)–(d) do *not* characterize our typical ways of attaining self-knowledge, Shoemaker concludes that the object perception model is thoroughly misguided.

However, as Shoemaker also notes, these points of disanalogy do not dispose of the idea that we detect our mental states by means of some kind of causal mechanism. This, in conjunction with the claim that our mental states obtain "independently of the perceiving of them" (206), Shoemaker calls the *broad perceptual model*.[13] He takes Armstrong to be defending just the broad perceptual model, rather than pressing any especially close comparison with senses like vision. Why insist on calling it the "perceptual model," though? Shoemaker points out that some clear cases of perception depart from the stereotype of vision. Smell, for instance, does not afford "identification information": "Smelling a skunk does not put one in a position to make demonstrative reference to a particular skunk" (223). There is no evident organ of proprioception (207). And at least some have claimed there are no proprioceptive experiences either: according to Anscombe, there are no "separately describable sensations ... when we know the position of our limbs" (1963, §8; see also §28 and Moran 2001, 19). Moran gives another example where a "quasi-perceptual presentation" seems to be lacking, namely our sense of time (2001, 19–20). Although Moran himself takes this to be a case of knowledge *without* perception, it is quite unclear why we should not take it instead as an example of the diversity of perceptual mechanisms, and as illustrating the perils of concentrating too much on the visual case.[14]

What we learn from the present objection is that if there is such a thing as inner sense, it is quite unlike paradigmatic outer senses—in particular, it is quite unlike vision. That is an interesting and important observation, but is not damaging to the inner-sense theory.[15]

4.2 INNER SENSE LEADS TO ALIENATED SELF-KNOWLEDGE

A considerably more interesting objection, due to Moran, is this. The inner-sense theory offers "a picture of self-knowledge as a kind of mind-reading applied to oneself, a faculty that happens to be aimed in one direction rather than another." However, "our ordinary self-knowledge [is] different from this sort of self-telepathy" (2001, 91). In particular:

> in ordinary circumstances a claim concerning one's attitudes counts as a claim about their *objects*, about the world one's attitudes are directed on . . . the expression of one's belief carries a commitment to its truth. (92)

That is, if in ordinary circumstances I say "I believe that it's raining," I am not disinterestedly reporting on the mental states of someone who happens to be me. I am also *committed* to the meteorological hypothesis that it's raining. To put the point in terms familiar from Moore's paradox, I am not prepared to follow up my psychological report with "... but it isn't raining." As Moran notes, in special circumstances I *can* do that—perhaps one sunny day my therapist convinces me that my obsessive umbrella-carrying is best explained by the hypothesis that I believe that it's raining. At the therapist's, I may say "I believe that it's raining" without "avow[ing] the embedded proposition ... itself" (85). But this is quite atypical: Moran's worry seems to be that *all* knowledge of one's beliefs would be of this "noncommittal" sort, if the inner-sense theory were correct.

If this is the worry, it is readily defused. Let us say that a belief is *alienated* just in case the belief is to a significant extent inferentially isolated—in particular, it is not expressible by the subject in unembedded speech. Beliefs uncovered at the therapist's provide standard examples. On the face of it, the (admittedly not entirely clear) alienated/unalienated distinction can be drawn without supposing that the subject has self-knowledge, and thus without smuggling in any question-begging assumptions. Suppose one has an unalienated belief that it's raining—one asserts that it's raining when queried, and the belief functions in the usual way to guide present action and future planning. Suppose now that the inner-sense theory is true, and that one's faculty of "self-telepathy" delivers the verdict that one believes that it's raining. *Pace* Moran, one's claim that one has this belief *will* carry "a commitment to its truth." Because the belief detected is unalienated, one will *not* say "Fancy that, I believe that it's raining! I wonder if that belief of mine is true?"[16]

4.3 INNER SENSE CAN'T DETECT EXTRINSIC PROPERTIES

In "Content and Self-Knowledge" (1989), Boghossian argues that the "apparently inevitable" thesis of externalism about mental content leads to the absurd conclusion that "we could not know our own minds" (149), thus presenting a paradox, which he leaves unresolved. Part of Boghossian's case involves ruling out inner sense (or "inner observation") as a source of self-knowledge. And "inner observation" is construed quite loosely, along the lines of Shoemaker's "broad perceptual model":

> It makes no difference to the argument of this paper if you think of inner observation as amounting to traditional introspection, or if you think of it as amounting to the operation of some Armstrong-style "brain scanner." What *is* crucial to inner observation models of self-knowledge is the claim that beliefs about one's own thoughts are justified by the deliverances of some internal monitoring capacity, much like beliefs about the external environment are justified by the deliverances of an external monitoring capacity (perception). (149, fn. 1)

According to externalism, the property of *believing that* p (for many fillings for '*p*') is *extrinsic* or, in Boghossian's terminology, *relational*. For instance, recalling Putnam's Twin Earth thought experiment, two individuals may be intrinsically or internally just alike, with only one having the property of believing that water is wet. According to Boghossian, externalism and the inner-sense theory are incompatible because:

> you cannot tell by mere inspection of an object that it has a given relational or extrinsic property. This principle is backed up by appeal to the following two claims, both of which strike me as incontestable. That you cannot know that an object has a given relational property merely by knowing about its *intrinsic* properties. And that mere inspection of an object gives you at most knowledge of its intrinsic properties. (162)

To this it might be objected that one *can* tell by "mere inspection" that a dime one is holding has the extrinsic property of being worth ten cents. Boghossian replies that this is not *mere* inspection, because:

> the process by which we know the coin's value is not really [mere] inspection, it's inference: you have to deduce that the coin is worth ten cents from your knowledge of its intrinsic properties plus your knowledge of how those intrinsic properties are correlated with possession of monetary value. And our knowledge of thought is not like that. (163–64)

Boghossian's principle, then, is that if one perceives (only) an object o, and has no relevant background information, one cannot thereby come to know that o is F, where Fness is an *extrinsic* property of o.

Bohgossian mostly speaks of "thoughts," but the paradox is supposed to cover "standing states" like belief (157), so we may just consider this case. Suppose a person S believes that *p*. What is the object o, and property Fness, to which we can apply Boghossian's principle? There are three candidates. Given that the object o is *perceived*, they all presuppose—contrary to Boghossian's apparent advertisement—something like Shoemaker's "object perception model."

First, it might be suggested that o is S's "token belief that *p*," with Fness being the (extrinsic) property of having the content that *p*. But that can't be right. Familiarly, believing is a state of a person, not an event or process. When someone believes that *p*, there is no psychological particular in her head or elsewhere that *is* her belief—"token beliefs" are a philosopher's invention.

To introduce the second candidate, note that S's being in the state of believing that *p* might involve the presence of a certain mental representation in the subject's head—say, a token sentence *s* in the language of thought. Could this string of Mentalese symbols *s* be the perceived object o, with Fness being, as before, the property of having the content that *p*? No—granting these speculations about mental representations for the sake of the argument, *s* is never an object of perception in the first place.

Finally, the third candidate for o is S herself, with Fness being the property (state) of believing that *p*. Unlike sentences in the head, sometimes oneself *is* among

the objects of perception, as when one looks in a mirror. However, the resulting version of Boghossian's principle is not going to deliver any conflict between the inner-sense theory and externalism because, in a situation in which inner sense is allegedly operative, one can come to know that one believes that *p without* perceiving oneself.

It might be retorted that if inner sense cannot be turned on oneself, one can hardly employ it to come to know that one believes that *p*. If that is right, then we have a swift refutation of the inner-sense theory that does not rely on externalism. But it isn't right: one can come to know through ordinary vision, without seeing *oneself*, that one is facing a table, for example. And that point, incidentally, shows that Boghossian's principle is, at least taken without qualifications, incorrect: *being in front of oneself* is an *extrinsic* property of the table.

As Boghossian's discussion nicely brings out, the "object perception model" of introspection fails for want of appropriate objects. But that is *not* to impugn the inner-sense theory—Shoemaker's "broad perceptual model" is left untouched.

5. AGAINST INNER SENSE II: SHOEMAKER

In a number of papers, Shoemaker has developed an argument against the inner-sense theory that simultaneously serves as an argument for his own view. According to Shoemaker, "there is a conceptual, constitutive connection between the existence of certain sorts of mental entities and their introspective accessibility" (1994, 225). This is "a version of the view that certain mental facts are 'self-intimating' or 'self-presenting', but a much weaker version" (225) than the strong view associated with Descartes.

One way in which it is weaker is that Shoemaker does not think that believing that *p* entails believing that one believes that *p*. Taken out of context, he can be read that way ("it is of the essence of many kinds of mental states and phenomena to reveal themselves to introspection" [242]),[17] but his view is that the entailment only goes through if other conditions are added. As he puts it, "believing that one believes that *P* can be just believing that *P* plus having a certain level of rationality, intelligence, and so on" (244).

Shoemaker's argument against the inner-sense theory is that it predicts the possibility of a condition analogous to ordinary blindness, deafness, ageusia (loss of taste), and so on, which Shoemaker calls *self-blindness*:

> To be self-blind with respect to a certain kind of mental fact or phenomenon, a creature must have the ability to conceive of those facts and phenomena (just as the person who is literally blind will be able to conceive of those states of affairs she is unable to learn about visually) . . . And it is only introspective access to those phenomena that the creature is supposed to lack; it is not precluded that she should learn of them in the way others might learn of them, i.e., by observing her own behavior, or by discovering facts about her own neurophysiological states. (226)

The blind are as rational, intelligent, and conceptually competent as the rest of us—they merely lack a particular mechanism specialized for detecting states of their environment. If the inner-sense theory is right, then presumably the mechanism specialized for detecting one's mental states could be absent or inoperative, while sparing the subject's rationality, intelligence, and conceptual competence. But, according to Shoemaker, such self-blindness is not possible.[18]

Following Shoemaker, let us say that a *rational agent* is a "person with normal intelligence, rationality, and conceptual capacity" (236). To say that self-blindness is impossible is to say that, necessarily, any rational agent has the sort of privileged and peculiar access to her mental states that we typically enjoy.

Concentrating on beliefs, Shoemaker's basic strategy is this:

> What I shall be arguing, in the first instance, is that if someone is equal in intelligence, rationality, and conceptual capacity to a normal person, she will, in consequence of that, behave in ways that provide the best possible evidence that she is aware of her own beliefs . . . to the same extent as a normal person would be, and so is not self-blind. (236)

Suppose that rational agent George is self-blind. One might think that George's condition could be easily diagnosed, because he will sometimes say, "It's raining but I don't believe it is," or the like. No, according to Shoemaker: George's rational agency "will be enough to make [him] appreciate the logical impropriety of affirming something while denying that one believes it" (237). George, then, will not betray his self-blindness in *this* way. Might he betray it in some other way? For instance, wouldn't George be flummoxed if asked, "Do you believe that it's raining?" That takes us to step C, below, of Shoemaker's attempt to reduce to absurdity the hypothesis that George is self-blind (1988, 34–45):

A. Self-blind speaker George will recognize the paradoxical character of '*p* but I don't believe that *p*'.[19]
B. Since George is a rational agent, this recognition will lead him to avoid Moore-paradoxical sentences.
C. Further, George will recognize that he should give the same answer to 'Do you believe that *p*?' and '*p*?'
D. Continuing this line of argument: plausibly, there is "nothing in his behavior, verbal or otherwise, that would give away the fact that he lacks self-acquaintance" (i.e. the ordinary kind of self-knowledge of one's beliefs) (36).
E. If George really is self-blind, "how can we be sure ... that self-blindness is not the normal condition of mankind?" (36).
F. "[I]t seems better to take the considerations [above] as a reductio ad absurdum of the view that self-blindness [with respect to beliefs] is a possibility" (36).

Shoemaker then briefly argues that this sort of argument can be extended (with qualifications) to other states (45–48; see also 1994, 237).

As Shoemaker's intricate discussion amply illustrates, the argument to step D raises some very difficult and complicated issues, and one might well take it to founder somewhere along the way. For instance, consider the following objection:

> [T]here are conceivable circumstances in which the total evidence available to a man supports the proposition that it is raining while the total third-person evidence supports the proposition that he does not believe that it is raining... If George is self-blind, then in the envisaged circumstance he is going to be very puzzled. He knows that Moore-paradoxical sentences are to be avoided. Yet it will seem to him that such an utterance is warranted by the evidence... And now there will be something—namely his expression of puzzlement—that distinguishes him from the normal person. (1988, 42)

Shoemaker replies that this case "is not really conceivable":

> There is a contradiction involved in the idea that the total evidence available to someone might unambiguously support the proposition that it is raining and that the total third-person evidence might unambiguously support the proposition that the person does not believe that it is raining. For the total third-person evidence concerning what someone believes about the weather should include what evidence he has about the weather—and if it includes the fact that his total evidence about the weather points unambiguously toward the conclusion that it is raining, then it cannot point unambiguously toward the conclusion that he doesn't believe that it is raining. (43)

However, Shoemaker's reply is incorrect. Suppose I am self-blind, and my evidence is this: (a) the cat has come indoors soaking wet; (b) the weather forecast is for rain; (c) I am going out without my umbrella, carrying important papers that will spoil if it's raining. This evidence "points unambiguously toward the conclusion that it is raining"; it also points unambiguously toward the conclusion that I *don't believe* that it is raining—if I knew someone *else* behaved in this way, I would reasonably conclude that she does not believe that it's raining. (Assume that, somehow, I have determined that I dislike getting wet and ruining important papers.)

According to Shoemaker, this reasoning goes wrong because the evidence cited is not my *total* evidence: "the total third-person evidence concerning what someone believes about the weather should include what evidence he has about the weather." Thus, I have *another* item of evidence to be weighed in with the rest, namely that my evidence about the weather is that the cat came in soaking wet, etc. And if so, as Shoemaker says, this undercuts the conclusion that I do, after all, lack the belief that it's raining—if I knew that someone *else* had evidence that pointed unambiguously toward the conclusion that it's raining, then even if she walks out without an umbrella, that would not show that she doesn't believe that it's raining. Rather, it would suggest other hypotheses—perhaps that she believes that her umbrella is lost.

But what is it to "have evidence" about the weather? Presumably, it is (at least) to *believe* facts that confirm or disconfirm meteorological hypotheses. What's more, Shoemaker himself must think this, otherwise the objection he is trying to rebut would be a nonstarter. If I don't *believe* that the cat came in soaking wet, etc., it will not "seem to me" that the Moore-paradoxical sentence 'It's raining but I don't believe that it's raining' is true, and so there will be no "expression of puzzlement" that distinguishes me from the normal person.

Thus, the fact that *my evidence* includes the fact that the cat came in soaking wet entails that I *believe* that the cat came in soaking wet. Hence, to insist that my total evidence should include what evidence I have about the weather is tantamount to assuming that I have knowledge (or true beliefs) about what I believe. But—since I am supposed to be self-blind—this is exactly what cannot be assumed.

Even if we waive these difficulties in reaching step D, the rest of the argument is hardly plain sailing. Suppose that step D is secured: George, our allegedly self-blind man, behaves in every way like a man who has the ordinary sort of self-knowledge. Why are we supposed to agree that George really does have self-knowledge? Why hasn't Shoemaker just outlined a strategy for *faking* self-knowledge? (Shoemaker himself, of course, is no behaviorist.)

Further, even we grant every step of the argument, and agree that George *does* have self-knowledge, that doesn't obviously show anything about *us*. In particular, it doesn't show that we have no faculty of inner sense. Admittedly, if George has self-knowledge, then we—at least, those of us who are "rational agents"—could come by self-knowledge without deploying inner sense. But—given the sophistication of George's reasoning—why doesn't this simply show that rational agents have a *backup* to their faculty of inner sense? An analogy: imagine that, by exploiting various subtle nonvisual cues (auditory, olfactory, etc.), a suitably clever person could have the normal sort of knowledge of her environment but without opening her eyes. When facing a strawberry, for instance, she immediately identifies it as red (suppose it gives off a distinctive odor). This would not indicate that the visual system is a myth.[20]

6. SHOEMAKER'S INSIGHTS, AND PRIVILEGED ACCESS AGAIN

Although Shoemaker hasn't shown that self-blindness is an impossibility, or that the inner-sense theory is mistaken, his discussion contains some crucial insights.

Let us say (vaguely but serviceably) that a theory of self-knowledge is *economical* just in case it implies that self-knowledge is produced solely by epistemic capacities needed for other domains of enquiry. Ryleanism is economical: the capacities for *self*-knowledge are precisely the capacities for knowledge of the minds of *others*. Shoemaker's theory is also economical: here the relevant capacities are "normal intelligence, rationality, and conceptual capacity." The inner-sense theory, on the other hand, is *extravagant*: the organs of outer perception, our general capacity for rationality, and so forth, do not account for all our self-knowledge—for that, an additional mechanism, an "inner eye," is needed.

Shoemaker's particular economical theory of self-knowledge suggests, somewhat ironically, that the "broad perceptual model" is ill-named. Recall that, according to it, we detect our own mental states by means of a causal mechanism, and the states obtain "independently of their being perceived" (Shoemaker 1994, 204).

However, this does not distinguish perception from the psychological process

of *reasoning*, which, like perception, can extend knowledge. Initially Holmes knows that Mr. Orange has been shot, that Mr. Pink has an alibi, that Mr. White's fingerprints are on the gun, and so on; reasoning from this evidence results, suppose, in Holmes knowing that Mr. White committed the murder. Holmes has detected Mr. White's guilt by means of a causal mechanism, and the fact detected obtains independently of the detecting of it. Of course, perception has entered somewhere along the line—Holmes dusted the gun for prints himself, say. But Holmes has not *perceived* that Mr. White committed the murder, as he might see various arches, loops, and whorls on the handle of the gun. The so-called "broad perceptual model" is consistent with self-knowledge not involving the perception of anything *mental*.

One might think that this loophole doesn't amount to much. A Rylean might insist that we never strictly speaking *perceive* that someone is in a certain mental state, but instead infer that she is from her behavior (see sec. 2 above). Such a version of Ryleanism would fit the "broad perceptual model" without involving the perception of anything mental, but what is the interest of that?[21]

But there is another way in which self-knowledge might involve reasoning without the perception of anything mental. Recall the quotation from Evans about belief and transparency (sec. 3). One apparently finds out that one believes that it's raining by determining that it's raining: knowledge that one has this belief, insofar as it rests on perceptual evidence at all, rests on perceptual evidence about the weather, not on perceptual evidence of one's behavior or anything mental. That is, one *reasons* from the evidence that it's raining, to the conclusion that one believes that it's raining. If this procedure can yield self-knowledge, and if it involves the (causal) *detection* of the belief that it's raining, then this would be an instance of the "broad perceptual model" without being either Ryleanism or the inner-sense theory.

And although we haven't yet found a compelling reason to reject the inner-sense theory, there is a reason for pursuing an alternative. As noted in section 2, the inner-sense theory neatly explains *peculiar* access. But it does not explain *privileged* access. In fact, it leaves it something of a mystery. Why is inner sense less prone to error than the outer senses? (Recall Lashley's astigmatic inner eye.) And why is there not (or, at any rate, not obviously) an actual psychological condition that approximates Shoemaker's "self-blindness"?

7. TRANSPARENT RULES

The account to follow appeals to the notion of *following a rule*, specifically an *epistemic* rule. This apparatus of epistemic rules needs to be explained first.

7.1. EPISTEMIC RULES

Holmes's reasoning to the conclusion that Mr. White killed Mr. Orange is complex, and his methods resist easy summary. Presumably Holmes's reasoning is somehow rule-governed, but it is not clear how to identify the rules. On the other hand, some

reasoning is considerably simpler. For example, Mrs. Hudson might hear the doorbell ring, and conclude that there is someone at the door. By hearing that the doorbell is ringing, Mrs. Hudson knows that the doorbell is ringing; by reasoning, she knows that there is someone at the door.

It is natural to say that Mrs. Hudson acquires knowledge of her visitors by following a simple recipe or rule. If we say that an *epistemic rule* is a conditional of the following form:

R If conditions C obtain, believe that p[22]

then the epistemic rule that Mrs. Hudson follows is:

> DOORBELL If the doorbell rings, believe that there is someone at the door[23]

What does it mean to say that Mrs. Hudson *follows* this rule on a particular occasion? For present purposes this semi-stipulative answer will suffice: Mrs. Hudson believes that there is someone at the door *because* she recognizes that the doorbell is ringing. The 'because' is intended to mark the kind of reason-giving causal connection that is often discussed under the rubric of 'the basing relation'. Mrs. Hudson might recognize that the doorbell is ringing, and believe that there is someone at the door for some *other* reason; in this case, she does not form her belief because she recognizes that the doorbell is ringing.

So S follows the rule R ('If conditions C obtain, believe that p') on a particular occasion iff on that occasion:

(i) S believes that p because she recognizes that conditions C obtain

which implies:

(ii) S recognizes (hence knows) that conditions C obtain

(iii) conditions C obtain

(iv) S believes that p

Following DOORBELL tends to produce knowledge about one's visitors (or so we may suppose), and hence it is a *good* rule. Following *bad* rules tends to produce false and unjustified beliefs, for example:

> NEWS If the *Weekly World News* reports that p, believe that p

NEWS is also an example of a *schematic* rule. One *follows* a schematic rule just in case one follows a rule that is an instance of the schematic rule; a schematic rule is *good* to the extent that its instances are.

If the antecedent conditions C of an epistemic rule R are not specified in terms of the rule follower's mental states, R is *neutral*. A schematic rule is neutral just in case some of its instances are. Thus, the claim that S can follow a neutral rule does not presuppose that S has the capacity for self-knowledge. DOORBELL and NEWS are neutral rules; 'If you intend to go swimming, believe that you will get wet' is not.[24]

Self-knowledge is our topic, not skepticism: knowledge of one's environment (including others' actions and mental states) and reasoning (specifically, rule-following of the kind just sketched) can be taken for granted. So, in the present con-

text, it is not in dispute that we follow neutral rules, including neutral rules with mentalistic fillings for '*p*', like 'If S has a rash, believe that S feels itchy'; neither is it in dispute that some neutral rules are good rules.

Moran's "claim of transparency" (sec. 3) can be recast using the apparatus of epistemic rules as follows. Knowledge of one's beliefs may be obtained by following the neutral schematic rule:

BEL If *p*, believe that you believe that *p*

Since the antecedent of BEL expresses the content of the mental state that the rule-follower ends up believing she is in, BEL can be called a *transparent* rule.

7.2 THE PUZZLE OF TRANSPARENCY

But how can following BEL lead to self-knowledge? In his contribution to a symposium on *Authority and Estrangement*, Moran acutely observes that there is a puzzle here:

> the claim of Transparency *is* something of a paradox: how can a question referring to a matter of empirical psychological fact about a particular person be legitimately answered without appeal to the evidence about that person, but rather by appeal to a quite independent body of evidence? (2003, 413)

This *puzzle of transparency* can be expressed in the terminology of epistemic rules as follows. Apparently, knowledge of what one believes is often the result of following the neutral schematic rule BEL, yet surely this is a *bad* rule: that *p* is the case does not even make it *likely* that one believes that it is the case.[25]

However, recall the "rule of necessitation" in modal logic. According this rule of inference, if a sentence '*p*' is a line of a proof, one may write down the necessitation of '*p*', '$\Box p$', as a subsequent line. Artificially forcing this into a format similar to that of "epistemic rules," the rule of necessitation becomes:

NEC If '*p*' is a line, you may write '$\Box p$' as a subsequent line

NEC, it seems, does not preserve truth, and so—in an extended sense—is a "bad" rule. It doesn't follow from the fact that the cat is indoors that *necessarily* the cat is indoors. The cat's being indoors doesn't even make it *likely* that this state of affairs could not have been otherwise.

But, of course, the rule of necessitation is not a bad rule. In fact, it's a necessarily truth-preserving rule. The reason is that—assuming that the only initial premises of a proof are axioms—whenever one is in a position to follow the rule by writing down '$\Box p$', '*p*' *is* a necessary truth. The axioms of a system of modal logic are themselves necessary truths, and whatever follows from them by the other rules are also necessary truths. So whenever one is in circumstances in which the rule applies—whenever, that is, one is confronted with a proof whose initial premises are axioms—every line of the proof is a necessary truth. If the allowable substituends

for '*p*' include sentences about the location of cats, then the rule of necessitation is a bad rule. But if (as intended) it is kept within the confines of modal logic, the rule is perfectly good.

Having noticed this, it is a short step to noticing that something analogous holds for BEL. One is only in a position to follow BEL by believing that one believes that *p* when one has recognized that *p*. And recognizing that *p* is (inter alia) coming to *believe* that *p*.[26] BEL is *self-verifying* in this sense: if it is followed, the resulting second-order belief is true. Compare a third-person version of BEL:

BEL-3 If *p*, believe that Fred believes that *p*

BEL-3 is of course not self-verifying: the result of following it may be (indeed, is very likely to be) a false belief about Fred's beliefs.

Given that we follow rules like DOORBELL, it should not be in dispute that we *can* follow BEL. Given the plausibility of Evans's observation about the procedure we actually follow, it should not be in dispute that we *do* follow BEL. The puzzle of transparency is solved by noting that BEL is self-verifying; since the goodness of rules like DOORBELL can be assumed, it should not be in dispute that following BEL will often produce *knowledge* of what one believes. BEL offers an obvious explanation of *peculiar* access: as just noted, BEL-3 is a very bad rule indeed. But, if BEL is to be the whole story, privileged access must also be explained. At a minimum, we need to show that BEL is significantly *better*—more knowledge-conducive—than rules whose consequents concern others' mental states.

7.3 PRIVILEGED ACCESS EXPLAINED

Since following a rule like DOORBELL will deliver beliefs that are about as likely to amount to knowledge as our beliefs about others' mental states, for simplicity DOORBELL can go proxy for a good rule whose consequent concerns others' mental states. In what ways is BEL better than DOORBELL?

One immediate advantage of BEL over DOORBELL is that the former but not the latter is self-verifying. Suppose one follows DOORBELL, and so knows that the doorbell is ringing and believes that there is someone at the door. One's belief that there is someone at the door is probably true, but it may be false. Suppose one also follows BEL: in particular, one recognizes that the doorbell is ringing and thereby believes that one believes that the doorbell is ringing. Because BEL is self-verifying, the truth of one's second-order belief is guaranteed.

Suppose there is someone at the door, and so the belief produced by following DOORBELL is true—how likely is it to be knowledge? Following Sosa (1999) and Williamson (2001, ch. 5), say that one's belief that *p* is *safe* just in case one's belief could not easily have been false.[27] Safety is a plausible necessary condition for knowledge; absent countervailing considerations (such as having excellent but misleading evidence that not-*p*), safety can be used as a rough-and-ready diagnostic tool for the presence of knowledge where the proposition in question is contingent. Could one easily have been wrong about the presence of a visitor? The ways in

which one could have falsely believed that there is someone at the door can be classified into three types:

> *Type I*: not-*p*, and one falsely believes that conditions C obtain, thereby believing that *p*. Perhaps the sound made by a passing ice cream truck might have been mistaken for the ringing of the doorbell, leading to the false belief that there is someone at the door.
>
> *Type II*: not-*p*, and one truly believes that conditions C obtain, thereby believing that *p*. Perhaps a wiring defect might have caused the doorbell to ring, leading to the false belief that there is someone at the door.
>
> *Type III*: not-*p*, and one believes that *p*, but not because one knows or believes that conditions C obtain. Perhaps too much coffee might have lead one to believe that there is someone at the door, even if the stoop had been deserted.

By hypothesis, there is someone at the door. Also by hypothesis, one *follows* DOORBELL, which entails one *knows* that the doorbell is ringing. Hence one could not easily have been wrong about that, and so Type I errors are remote possibilities. And, given certain assumptions that will obtain in many realistic cases (the doorbell has no wiring defects, the coffee is not that psychoactive, etc.), Type II and Type III errors are also remote possibilities and could not easily have happened. However, in other realistic cases these errors *are* nearby possibilities, and hence one's true belief that there is someone at the door will not be knowledge.

Consider now the belief that one believes that there is someone at the door; could one easily have been wrong? It is not possible to make a Type I error: one cannot falsely believe that the doorbell is ringing without believing that the doorbell is ringing. Type II errors are likewise ruled out: one cannot truly believe that the doorbell is ringing without believing that the doorbell is ringing.

If one follows BEL, only Type III errors are a threat to one's knowledge: perhaps too much coffee would have lead one to believe that one believes that the doorbell is ringing, even if one had not believed that the doorbell is ringing. With the modest assumption that Type III errors are equally likely when following BEL as when following DOORBELL, the true beliefs produced by following BEL are more likely to amount to knowledge than the true beliefs produced by following DOORBELL.[28]

Sometimes one will not succeed in following DOORBELL because one believes but does not know that the doorbell is ringing (maybe a passing ice cream truck induces a false belief). Say that S *tries* to follow rule R iff S believes that *p* because S *believes* that conditions C obtain. That S follows R entails that she tries to follow R, but not conversely. If one tries to follow DOORBELL but does not succeed, then one will not *know* that there is someone at the door; if one's belief about a visitor is true, that is just an accident. The visitor could have easily been delayed, with the truck passing as it actually did, in which case one would have falsely believed that there is someone at the door. That is, a Type I error is a nearby possibility.

Sometimes one will not succeed in following BEL either: one will merely try to follow it, and believe but not know that the doorbell is ringing. But one's second

order belief that one believes that the doorbell is ringing will be *true*. As before, Type I and II errors are not possible. Hence this situation will be commonplace: trying to follow BEL, one investigates whether *p*, *mistakenly* concludes that *p*, and thereby comes to *know* that one believes that *p*. (In these cases, one will know that one believes that *p* on the basis of no evidence at all.)

BEL, then, has considerable epistemic virtues, but it is important to not overstate them. Consider the following quotation from Evans (continuing the quotation given in section 3 above):

> I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p* . . . If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states: even the most determined sceptic cannot find here a gap in which to insert his knife. (Evans 1982, 225)

This passage can be read as suggesting that (a) the ability to investigate whether *p* brings with it the ability to find out whether one believes that *p* (assuming one has the concept of belief), and that (b) following BEL cannot fail to produce knowledge of one's beliefs.

Whether or not this is Evans's view, on the present account it is incorrect.[29] There is no guarantee that one can follow (or try to follow) BEL, no matter how talented one is at gaining knowledge of cats, doorbells, other minds, and so on. Neither is there any guarantee that the beliefs produced by following (or trying to follow) BEL will amount to knowledge.

However, as a contingent matter, trying to follow BEL will usually produce knowledge of what one believes. Venturing out on a limb—of course the matter requires more discussion—we may tentatively conclude that privileged access is thereby explained.

7.4 SHOEMAKER'S INSIGHTS VALIDATED

The hypothesis that privileged and peculiar access is explained by our following (or trying to follow) BEL *is* a version of Shoemaker's "broad perceptual model." Suppose someone follows an instance of BEL, and thereby believes that she believes that it's raining. She looks out of the window, say, and sees that it's raining. A causal transition between mental states occurs, as it does when one believes that there is someone at the door because one has recognized that the doorbell is ringing. The subject believes (knows) that it's raining, which causes her to believe that she believes that it's raining. Thus there is an appropriate causal mechanism. Further, the state detected is independent of its detection. The subject might not have followed BEL, in which case the first-order belief would have been present without the second-order belief. What's more (we may fairly suppose), someone might believe that it's raining, possess the concept of belief, and yet not even have the capacity to follow BEL.

Yet the account is not a version of the inner-sense theory. It is *economical*, like behaviorism. Taking the capacity to follow good neutral rules for granted, knowledge of what one believes comes along more-or-less for free. Since this capacity belongs to the department of reasoning, not perceiving, Shoemaker's idea that the source of self-knowledge can be traced to "rationality" is vindicated, albeit not via his preferred route.

8. BEYOND BELIEF?

It is often pointed out the phenomenon of transparency is quite limited.[30] It covers belief and perception, but what about, to take the obvious example, knowledge of one's *wants* or *desires*? I come to know that I *believe* that I have a beer by looking outward and discovering a beer in my hand. One does not typically come to know that one *wants* a beer by the same procedure. Typically, when I want a beer, that is because I *don't* have one. And I often have privileged and peculiar access, when in such a beerless condition, to the fact that I want a beer. So some other account is required to explain one's knowledge of one's desires, to say nothing of wishes, hopes, fears, expectations, thoughts, intentions, imaginings, and the rest.[31]

So far, none of this is an *objection* to the account of the previous section, just some observations that highlight the limitations of the approach. But there is an impending problem—the *puzzle of opacity*. Suppose that the epistemology of nontransparent states is *extravagant*, and hence that self-knowledge of wants, hopes, intentions, and so forth cannot be explained in terms of our ability to follow neutral rules. In short: a faculty of introspection is needed. Then the puzzle is this: why isn't inner sense ever operative in the case of transparent mental states? Why is the epistemology of these states always (apparently) of the transparent and economical sort?

Perhaps the puzzle of opacity can be solved, consistently with an extravagant epistemology of nontransparent mental states. But a two-tiered account of self-knowledge—economical in the case of transparent mental states, extravagant in the case of the rest—is not an entirely comfortable position. This at least motivates an exploration of whether extravagance is really needed.

Let us briefly consider desire. There are desires and desires: likes, wants, preferences, cravings, lusts, wishes, and so on. A first-person epistemology for all members of this heterogeneous category will not be attempted here. Instead, as a tractable example, take the preference for one of a range of options. I have neither beer nor wine, and am offered one or the other. Here is the beer, the culmination of centuries of Belgian brewing tradition. There is the wine, the product of my host's home winemaking hobby. I prefer the beer to the wine. How do I know that?

As just noted, clearly transparency does not apply here—I do not (usually) find out that I prefer the beer by finding out that I am holding a glass of the stuff. However,

often my eyes are still "directed outward—upon the world." I can investigate my preferences by attending to the *beer* and the *wine*, and their relative merits (and perhaps to the host as well, in particular her tendency to take offense). I conclude that the beer wins over the wine, and thereby conclude that I prefer the beer. The relevant (neutral) rule is roughly this:

> DES If ψing is a better option than χing, believe that you prefer to ψ than to χ[32]

DES is a neutral rule, so the capacity to follow it does not presuppose the capacity for self-knowledge. But is DES a *good* rule? One's preferences tend to line up with one's beliefs about the merits of the options: if one believes that ψing is a better option than χing, one typically prefers to ψ than to χ. On the face of it, DES is a good rule; moreover, it can be used to explain privileged access to one's desires in a way that closely (although not perfectly) parallels the explanation given in section 7.3 above.

Unlike BEL, DES is not a transparent rule. And there is another important difference. No doubt *Socrates* would be happy with the claim that one prefers to ψ than to χ whenever one believes that ψing is a better option than χing, but—as is familiar from the literature on weakness of will—he appears to have been wrong about that. Unlike BEL, DES is not self-verifying. Slavishly following (or trying to follow) DES will sometimes lead to a *false* belief about one's mental state.

Consider an example. Suppose I believe that the wine is a better option than the beer, because my host will take offense if I choose the beer. Nonetheless, I selfishly prefer the beer to the wine. If I follow DES, I will falsely believe that my preferences are the other way round. The problem is not that such a mistake can never happen—it can. (I believe that it would be considerably better to read *Mind and World* this evening than to watch *The Real World*; thinking myself a person who is sensitive to intellectual virtues, I believe that I prefer the former to the latter. However, I find myself turning on the television, leaving the book unopened.)[33] The problem, rather, is that sometimes one *knows* that one's preferences are at odds with one's better judgments. In particular, despite believing that the wine is better than the beer (or, alternatively, not having an opinion on the matter), I may well have privileged and peculiar access to the fact that, all things considered, I prefer the beer to the wine.

The issue is whether this sort of knowledge will required an extravagant faculty of introspection. And here the buck is passed from preference to intention, because in the situation just described, it is plausible that I know I prefer the beer to the wine because I know that I *intend* to have the beer.

Further investigation will have to be deferred.[34] To summarize the conclusion so far: at least with respect to belief, the inner-sense theory is partly right. There is an inner mechanism for detecting one's beliefs. But the inner-sense theory is also partly wrong: the mechanism comes with our capacity for reasoning about the external world—there is no inner eye.

NOTES

Distant ancestors of parts of this paper were included in talks given at Alberta, Calgary, Stanford, USC, the University of Texas at Austin, Union College, Western Washington University, Vermont, and a Metaphysics and Epistemology conference in Dubrovnik, Croatia. The numerous comments I received on those occasions greatly improved this paper. I am especially grateful to the participants in Ned Block and Thomas Nagel's seminar on language and mind at NYU, and in my graduate seminars on other minds and self-knowledge at MIT. For comments on the penultimate draft, thanks to David Hilbert, Richard Holton, Ed Minar, and Susanna Siegel.

1. Cf. Wright 1998, 24: "The privileged observation explanation [of "first-third person asymmetries in ordinary psychological discourse"] is unquestionably a neat one. What it *does* need philosophy to teach is its utter hopelessness."
2. As much psychological research has shown, we are often mistaken about our reasons for belief or action (Wilson 2002); however, none of this undermines the (relatively modest) sort of privileged access claimed in the text. For some discussion of what this research does and doesn't show, see Wilson 2002, 104–15; see also Nichols and Stich 2003, 161.
3. Indeed, they actually come apart for *knowing that the cat is indoors, seeing the cat*, and the like; we have peculiar but not (an impressive kind of) privileged access to these states.
4. The claim of "Privileged Access," in Ryle's sense, is this: "(1) ... a mind cannot help being constantly aware of all the supposed occupants of its private stage, and (2) ... it can also deliberately scrutinize by a species of non-sensuous perception at least some of its own states and operations" (1949, 148). In the contemporary literature, 'privileged access' is often used approximately for what (in the text) is described as privileged *and* peculiar access (see, e.g., Alston 1971; Moran 2001, 9–10).
5. For the near-Rylean position in psychology, see Bem 1972. Although one (perhaps incautious) statement of Ryle's official view is that "in principle, as distinct from practice, John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe" (1949, 149), he has no account of the third-personal method by which (to take Ryle's own examples) "I can catch myself daydreaming," or "catch myself engaged in a piece of silent soliloquy" (160).
6. This quotation, together with Lashley's comparison with astigmatism, appears in Lyons 1986, 29. Ryle's characterization of "inner perception" denies that there are "any counterparts to deafness, astigmatism" (1949, 157).
7. Another equally notable recent alternative is Bar-On 2005, to be passed over for reasons of space.
8. See also Dretske 1994, 1995, and Gordon 1996. A similar view can be found in Husserl: see Thomasson 2003 for an interesting discussion.
9. See Moran 2001, 63, 64, 65, 67.
10. At one point Moran contrasts the two ways of answering the question "Do I believe P?" as follows:

    > In characterizing the two sorts of questions one may direct towards one's state of mind, the term 'deliberative' is best seen at this point in contrast to 'theoretical', the primary point being to mark the difference between that enquiry which terminates in a true description of my state, and one which terminates in the formulation or endorsement of an attitude. (2001, 63)

    However, this is misleading (and is not Moran's considered view). In successfully answering the question "Do I believe P?", whether in a deliberative or theoretical spirit, one comes to have a true belief about one's beliefs, and so in both cases the enquiry "terminates in a true description of [one's] state."
11. For what is essentially the same point, see Peacocke 1998, 215–16.

12. As will become apparent, the account given in section 7 classifies Moran's special cases together with examples where one's mind is already made up, as both involving (in Moran's phrase) "epistemic access ... to a special realm."
13. It is also part of the broad perceptual model that mental states could obtain even if no creature had the capacity for introspection (1994, 206, 224–25). For simplicity this will be left tacit.
14. For a brief review of psychological work on time perception, see Allan 2000.
15. Moran puts another objection to the inner-sense theory as follows:

    the simple belief that Wagner died happy is constituted by a host of inferential commitments concerning related matters (about Wagner, death, happiness, and much else) and the truth of various counterfactuals. How, then, one may ask, could all of *this* be presented to my immediate inner perception when I am aware of what I believe about Wagner? I don't even know what "all" of this *is* ... (Moran 2001, 14)

    It is not obvious that Moran takes this objection to be particularly troubling (see 17 fn. 13); in any event, it is fallacious. Supposing that a constitutive account of elephants adverts to a host of evolutionary relationships of which I am ignorant, it does not follow that I can't "immediately" perceive that there is an elephant on the lawn.
16. A related worry is that if the inner-sense theory is correct, one could *sometimes* use one's inner eye to detect an alienated belief; without the benefit of therapy, or reflecting on one's past pattern of behavior, one could sometimes discover that one has the (alienated) belief that it's raining. If (as seems plausible) this never in fact happens, the inner-sense theorist owes us an explanation of why not. However, it is unclear why she can't provide one, given the large functional difference between alienated and unalienated beliefs.
17. Shoemaker is of course using 'introspection' broadly here, to denote the special method (perceptual or not) we have of finding out about our own mental states.
18. It might be argued that someone without an inner sense would lack the concept of belief, as it might be argued that the blind lack color concepts (cf. Peacocke 1992, 151–62, and Shoemaker 1994, 236, fn. 3). But this highly controversial claim can be set aside here.
19. This sort of Moore-sentence should be distinguished from the one that figured in section 4.2, namely 'p but I believe that not-p'. The latter sort of sentence *is* assertable in (atypical) circumstances.
20. Shoemaker's reply to this objection is (extrapolating back from his 1994) that (a) the argument against the inner-sense theory does not assume that George's reasoning goes on in us ("obviously it doesn't" [1994, 239]) and that (b) Mother Nature would not have taken the trouble to install "an *additional* mechanism ... whose impact on behavior is completely redundant" (240). But if we do not *in fact* run through George's reasoning, how can the "availability of the reasoning" (240) *explain* our behavior? After all, going by Shoemaker's own description in his 1988, George's behavior is not explained by the mere availability of the reasoning.
21. The fact that Ryleanism in general is an instance of the broad perceptual model (and, in some variants, the object perception model) does not affect the discussion in Shoemaker 1994, which clearly presupposes that Ryleanism is false.
22. Since judging is the act that results in the state of belief, perhaps the consequent is better put as 'judge that p'. This is simply a stylistic or presentational issue, however. The linguistic formulation of the rule only plays a heuristic role—*all* the work is done by the account of *following* a rule (see immediately below).
23. No doubt the epistemological story is considerably more complicated; DOORBELL should be treated as a harmless simplification.
24. 'You' refers to the rule-follower; tenses are to be interpreted so that the time the rule is followed counts as the present.
25. The *locus classicus* for the puzzle of transparency (as it arises for perception) is Dretske 2003; see also Martin 1998, 117–18.
26. Of course, there are many differences between the rule of necessitation and BEL. For one thing, logical rules of inference are not rules of reasoning. With this cautionary remark in mind, there is another point of analogy. It is a mistake to think that the rule of necessitation is equivalent to the (invalid) axiom schema '$p \supset \Box p$' plus modus ponens; likewise, it is a mistake to think of fol-

lowing BEL as equivalent to (falsely) assuming that for all *P*, if *P* is true then one believes *P*, which would make one's reasoning from the premise that it's raining to the conclusion that one believes that it's raining demonstrative.

In his earlier writings on transparency, Dretske likened the phenomenon to examples of "displaced perception" such as the following: I see that the bathroom scale on which I am standing reads "170" and infer that I weigh 170 pounds (Dretske 1994, 263; 1995, 41). This reasoning stands or falls with the prior reasonableness of the assumption (or "connecting belief" [1995, 42]) that if the bathroom scale on which I am standing reads "170" then I weigh 170 pounds, and so is importantly *dis*analogous to BEL. Against Dretske, Aydede (2003) complains, in effect, that the connecting beliefs in the mental case are often false (see also Lycan 2003, 16–17, and 26–27 n. 1). (For Dretske's account of the difference between the mental case and examples like the scale, see Dretske 1995, 60–61.)

27. The formulation in the text is (approximately) Williamson's. For simplicity, situations that could easily have obtained in which one does not falsely believe that *p* but rather falsely believes something else will be ignored. See Williamson 2001, 101–2.
28. Some with "internalist" sympathies might insist that considerations of safety and the like are not enough: if following BEL leads to knowledge, the knowledge-conducive properties of BEL have to be in some way "accessible" to the rule-follower. This issue is just an instance of the general debate between externalism and internalism (see, e.g., Goldman 2001). On the face of it, the present proposal does not make an internalist account of self-knowledge *especially* problematic (compared to, say, an internalist account of perceptual knowledge); accordingly the externalism/internalism debate need not be examined here.
29. In the case of perception, which he contrasts with belief, Evans does deny that transparency "produce[s] infallible knowledge" (1982, 228).
30. See Goldman 2000, 182–83; Nichols and Stich 2003, 194; Finkelstein 2003, postscript; Bar-On 2004, 114–18.
31. The BEL-style approach cannot be applied to perception without modification. For example, the rule 'If *p*, believe that you see ("visually") that *p*' is not good. One often knows that the cat is indoors without seeing that it is.
32. Something like this view is implicit in Moran 2001 (see Finkelstein 2003, 161). A similar suggestion is also made in Gertler 2003, section 2.3. DES and some of the subsequent points in the text can be extracted from Shoemaker 1988, 47–48.
33. For a probing discussion of other sorts of examples, see Arpaly 2003.
34. See Byrne, in preparation.

REFERENCES

Allan, L. G. 2000. "Time Perception." In *Encyclopedia of Psychology*, ed. A. Kazdin. Washington, DC: American Psychological Association.

Alston, W. 1971. "Varieties of Privileged Access." *American Philosophical Quarterly* 8: 223–41.

Anscombe, G. E. M. 1963. *Intention*. 2d ed. Oxford: Blackwell.

Armstrong, D. M. 1968. *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Armstrong, D. M. 1981. *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press.

Arpaly, N. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Aydede, M. 2003. "Is Introspection Inferential?" In *Privileged Access: Philosophical Accounts of Self-Knowledge*, ed. B. Gertler. Aldershot: Ashgate Publishing.

Bar-On, D. 2005. *Speaking My Mind*. Oxford: Oxford University Press.

Bem, D. J. 1972. "Self-Perception Theory." In *Advances in Experimental Social Psychology*, ed. L. Berkowitz. vol. 6. New York: Academic Press.

Boghossian, P. 1989. "Content and Self-Knowledge." *Philosophical Topics* 17: 5–26. Page reference to the reprinting in Ludlow and Martin 1998.

Byrne, A. In preparation. *Transparency and Self-Knowledge*. Oxford: Oxford University Press.

Dretske, F. 1994. "Introspection." *Proceedings of the Aristotelian Society* 94: 263–78.

Dretske, F. 1995. *Naturalizing the Mind.* Cambridge, MA: MIT Press.
Dretske, F. 2003. "How Do You Know You Are Not a Zombie?" In *Privileged Access: Philosophical Accounts of Self-Knowledge*, ed. B. Gertler. Aldershot: Ashgate Publishing.
Evans, G. 1982. *The Varieties of Reference.* Oxford: Oxford University Press.
Finkelstein, D. H. 2003. *Expression and the Inner.* Cambridge, MA: Harvard University Press.
Gertler, B. 2003. "Self-Knowledge." In *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta (Spring 2003), URL = *http://plato.stanford.edu/archives/spr2003/entries/self-knowledge/*
Goldman, A. 1993. "The Psychology of Folk Psychology." *Behavioral and Brain Sciences* 16: 15–28.
Goldman, A. 2000. "The Mentalizing Folk." In *Metarepresentations*, ed. D. Sperber. Oxford: Oxford University.
Goldman, A. 2001. "Internalism Exposed." In *Knowledge, Truth, and Duty*, ed. M. Steup. Oxford: Oxford University Press.
Gordon, R. 1996. "'Radical' Simulationism." In *Theories of Theories of Mind*, ed. P. Carruthers and P. Smith. Cambridge: Cambridge University Press.
Lashley, K. 1923. "The Behavioristic Interpretation of Consciousness II." *Psychological Review* 30: 329–53.
Ludlow, P., and N. Martin, eds. 1998. *Externalism and Self-Knowledge.* Stanford: CSLI.
Lycan, W. G. 1987. *Consciousness.* Cambridge, MA: MIT Press.
Lycan, W. G. 1996. *Consciousness and Experience.* Cambridge, MA: MIT Press.
Lycan, W. G. 2003. "Dretske's Ways of Introspecting." In *Privileged Access: Philosophical Accounts of Self-knowledge*, ed. B. Gertler. Aldershot: Ashgate Publishing.
Lyons, W. 1986. *The Disappearance of Introspection.* Cambridge, MA: MIT Press.
Martin, M. 1998. "An Eye Directed Outward." In *Knowing our Own Minds*, ed. C. Wright, B. Smith, and C. Macdonald. Oxford: Oxford University Press.
Moran, R. 2001. *Authority and Estrangement.* Princeton: Princeton University Press.
Moran, R. 2003. "Responses to O'Brien and Shoemaker." *European Journal of Philosophy* 11: 402–19.
Nichols, S., and S. Stich. 2003. *Mindreading.* Oxford: Oxford University Press.
Peacocke, C. 1992. *A Study of Concepts.* Cambridge, MA: MIT Press.
Peacocke, C. 1998. "Conscious Attitudes, Attention, and Self-Knowledge." In *Knowing Our Own Minds*, ed. C. Wright, B. Smith, and C. Macdonald. Oxford: Oxford University Press.
Ryle, G. 1949. *The Concept of Mind.* London: Hutchinson. Page reference to the 1980 Penguin reprint.
Shoemaker, S. 1988. "On Knowing One's Own Mind." *Philosophical Perspectives* 2: 183–209. Page reference to the reprinting in Shoemaker 1996.
Shoemaker, S. 1994. "Self-Knowledge and 'Inner-Sense'." *Philosophy and Phenomenological Research* 54: 249–314. Page reference to the reprinting in Shoemaker 1996.
Shoemaker, S. 1996. *The First-Person Perspective and Other Essays.* Cambridge: Cambridge University Press.
Sosa, E. 1999. "How to Defeat Opposition to Moore." *Philosophical Perspectives* 13: 141–53.
Thomasson, A. 2003. "Introspection and Phenomenological Method." *Phenomenology and the Cognitive Sciences* 2: 239–54.
Williamson, T. 2001. *Knowledge and Its Limits.* Oxford: Oxford University Press.
Wilson, T. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Cambridge, MA: Harvard University Press.
Wright, C. 1998. "Self-Knowledge: The Wittgensteinian Legacy." *Knowing Our Own Minds*, ed. C. Wright, B. Smith, and C. Macdonald. Oxford: Oxford University Press.