Belief and Moore's Paradox

Declan Smithies
Draft for SPAWN, Syracuse University
August 14-16, 2013

How do we know what we believe? Any credible answer to this question must account for an asymmetry between first-person and third-person perspectives – that is, between the way in which we know what we ourselves believe and the way in which we know what others believe. In fact, this asymmetry has two aspects. First, our way of knowing what we believe is different from our ways of knowing what others believe. And second, our way of knowing what we believe is better – that is, more reliable and less vulnerable to ignorance and error – than our ways of knowing what others believe. In Alex Byrne’s (2005) useful terminology, our epistemic access to our own beliefs is both peculiar and privileged.

Gilbert Ryle famously denied the existence of an asymmetry between first-person and third-person perspectives:

Knowledge of what there is to be known about other people is restored to approximate parity with self-knowledge. The sorts of things that I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same. (1949: 149)

On the Rylean view, we know what we ourselves believe in much the same way that we know what others believe – that is, by inference from observation of behavior. But although the Rylean denies that we have peculiar access to our beliefs, he can accept a weak version of the claim that we have privileged access to our beliefs because we have more observational data from which to draw inferences about our own beliefs than the beliefs of others.

In denying that our epistemic access to our own belief is peculiar, however, the Rylean underestimates the extent to which it is privileged. Paul Boghossian summarizes the objection as follows:

The difficulty is not merely that, contrary to appearance and the canons of epistemic practice, he [i.e. the Rylean] has to construe the knowledge as inferential. The difficulty is that he has to construe it as involving inferences from premises about behavior that you could not possibly possess. (1989: 67)

Consider, for example, my belief that the population of the United Kingdom is more than 60 million people. I know that I believe this, but I don’t know it by observing my behavior. Indeed, I don’t even know how this belief would manifest itself in my behavior. Perhaps it would incline me to accept a bet that pays decent odds on the
population being more than 60 million, but my knowledge that I would accept such a bet depends upon my knowledge of what I believe, rather than vice versa.

Let us assume, then, that the Rylean is mistaken and that we do have some distinctively first-personal way of knowing what we believe. Still, the question remains, what exactly is the nature of this way in which we know what we believe? One might say that we know what we believe by introspection, but the term ‘introspection’ here is merely a placeholder. The task remains for a theory of introspection to give a more substantial account of what this introspective way of knowing what we believe consists in and how it is different from and better than our ways of knowing what others believe.

The main goal of this paper is to argue that what I call the simple theory of introspection (Smithies 2012a) can be extended to account for our introspective knowledge of what we believe as well as our introspective knowledge of conscious experience. The plan for the paper is as follows. Section one summarizes the simple theory and explains how the central motivations for the simple theory support the extension from conscious experience to belief. Sections two and three develop a new argument for extending the simple theory on the grounds that it best explains the irrationality of believing Moorean conjunctions. Section four examines the role that conscious experience plays in explaining our introspective knowledge of belief. Finally, section five explores some questions about the nature of epistemic idealization in defending the simple theory against objections.

1. The Simple Theory of Introspection

The simple theory of introspection is an epistemological theory, rather than a psychological theory. It is an account of the nature of introspective justification, which makes introspective knowledge possible. It is not an account of the psychological mechanisms or processes that we use in forming introspectively justified beliefs and acquiring introspective knowledge. As far as the simple theory is concerned, it is an open empirical question which psychological mechanisms or processes we use in forming our introspective beliefs, so long as they are formed in a way that is more or less directly sensitive to the mental states they concern.

According to the simple theory of introspection, being in a mental state of the right kind puts one in a position to know by introspection that one is in a mental state of that kind. Assuming that introspective skepticism is false, the simple theory implies that there are some mental states for which the following biconditional holds:

One is in a position to know by introspection that one is in M iff one is in M.

The left-to-right direction is fairly trivial, since knowledge is factive: quite generally, one is in a position to know that \( p \) only if it is a fact that \( p \). The opposite direction is more controversial, however, and sets the simple theory apart from its rivals, such as inner sense and reliability theories, which impose more demanding requirements

2
to the effect that one is in a position to know by introspection that one is in a mental state only if one represents that one is in that mental state or one has an introspective mechanism that reliably tracks that mental state. On the simple theory, in contrast, one can be in a position to know by introspection that one is in a certain mental state just by virtue of being in that mental state.

This paper is primarily concerned with introspective justification, rather than introspective knowledge. In fact, the simple theory of introspective knowledge can be viewed as a consequence of a simple theory of introspective justification that is more fundamental in the order of explanation. According to the simple theory, the source of introspective justification is identical with its subject matter. That is, one has introspective justification to believe that one is in a mental state of a certain kind just by virtue of being in a mental state of that kind. Assuming once more that introspective skepticisms is false, the simple theory implies that there are certain kinds of mental states for which the following biconditional holds:

One has introspective justification to believe that one is in M iff one is in M.

Given the further assumptions that introspective justification is indefeasible and immune from Gettier cases, any case in which one has introspective justification to believe that one is in a certain mental state is also a case in which one is in a position to know by introspection that one is in that mental state.

Not all mental states are sources of introspective justification. Psychophysical experiments have shown that unconscious visual representations of masked stimuli can prime subsequent performance on word completion tasks – for instance, subjects primed with the word ‘reason’ are more likely to complete the word stem ‘rea-’ as ‘reason’ than ‘reader’. But these subjects don’t have introspective justification to believe that they represent the masked stimulus. The source of justification to believe that there is unconscious visual representation of the masked stimulus derives from psychophysical experiments and not from introspection.

The simple theory is neutral on the question of which mental states are sources of introspective justification. Nevertheless, it is plausible that all conscious experience falls within the scope of the simple theory; that is, for any conscious experience E:

One has introspective justification to believe that one has E iff one has E.

The question I want to address in this paper is whether the simple theory can be plausibly extended from conscious experience to doxastic attitudes such as belief; that is, for any proposition p:

One has introspective justification to believe that one believes that p iff one believes that p.
In this section, I will summarize three central motivations for the simple theory and explain how they support extending the simple theory from conscious experience to belief. In the next section, I will add a further argument for extending the simple theory of introspection on the grounds that it explains the irrationality of believing Moorean conjunctions.

The first motivation for the simple theory derives from reflection on examples. It is plausible that if I am in pain, then I have introspective justification to believe that I am in pain just by virtue of being in pain. Similarly, if I am thinking about rhubarb, then I have introspective justification to believe that I am thinking about rhubarb just by virtue of thinking about rhubarb. And likewise for any other conscious experience. If I have some conscious experience, but I don’t believe that I have that conscious experience, then I am to that extent less than fully rational. There may be some perfectly good excuse for my rational failing – perhaps I am delusional, or distracted, or conceptually impoverished. Nevertheless, my conscious experience provides me with all the evidence that is needed to justify beliefs about my conscious experience and so my failure to form those beliefs constitutes a departure from the ideally rational response to my evidence.

Reflection on examples also motivates extending the simple theory from conscious experience to belief. If I believe that Canberra is the capital of Australia, then I have introspective justification to believe that I believe this just by virtue of believing it. To see the plausibility of this, consider the following. In fact, I do believe that Canberra is the capital of Australia, and I have believed this for many years, although I do not consciously think about the matter except on rare occasions. Similarly, I believe that I believe this – in fact, I know that I believe this – although, once again, I rarely think about the matter. So, we can ask, what justifies my belief that I believe that Canberra is the capital of Australia at a time when I’m not consciously thinking about the matter? The most plausible answer is that I have introspective justification to believe that I believe this just by virtue of believing it.

The second motivation for the simple theory is that it explains the sense in which introspection is peculiar and privileged. Introspection is peculiar in the sense that it is a way of knowing about one’s own mental states that is available only to oneself and not to anyone else. If someone else knows that I am in pain, then they know this by testimony or by making inferences from observation of my behavior or the state of my brain, whereas I am in a position to know that I am in pain just by virtue of being in pain. Furthermore, introspection is privileged in the sense that it is a way of knowing about one’s mental states that is always available to one. Others are in a position to know that I am in pain only if they have sufficient evidence from observation, inference, or testimony, whereas I am always in a position to know that I am in pain just by virtue of being in pain. In this way, the simple theory explains the sense in which introspection is epistemologically different from, and better than, other ways of knowing about the world.
This way of explaining the sense in which introspection is peculiar and privileged also motivates extending the simple theory from conscious experience to belief. One’s introspective knowledge of what one believes is peculiar and privileged in much the same way as one’s introspective knowledge of one’s conscious experience: that is, it is always available to oneself and it is never available to others. The simple theory explains these distinctive features of introspective knowledge in a way that can be straightforwardly generalized from conscious experience to belief.

The third motivation for the simple theory is that it is a consequence of an independently motivated thesis in epistemology – namely, access internalism. Access internalism is the thesis that justification is accessible in the following sense: one is always in a position to know all the epistemic facts about which doxastic attitudes one has justification to hold on the basis of introspection and a priori reflection alone. If access internalism is true, then what explains why it is true? The only plausible explanation is that the epistemic facts about which doxastic attitudes one has justification to hold are determined in an a priori way by non-epistemic facts about one’s mental states that one is always in a position to know by introspection. Therefore, the simple theory of introspection is indispensable for explaining the truth of access internalism.

Access internalism also motivates extending the simple theory from conscious experience to belief. This is because the epistemic facts about which doxastic attitudes one has justification to hold are determined in part by non-epistemic facts about one’s beliefs together with non-epistemic facts about one’s experience. After all, we cannot explain access internalism without assuming that one is always in a position to know all the relevant non-epistemic facts about one’s mental states on the basis of introspection. Therefore, one must always be in a position to know all the relevant non-epistemic facts about one’s beliefs as well as one’s conscious experience on the basis of introspection.

Can the simple theory of introspection be explained? In one sense, yes: the simple theory of introspection can be explained as a consequence of access internalism. But in another sense, no: the simple theory is a primitivist theory, rather than a reductionist theory. Reductionist theories seek to explain the epistemology of introspection by assimilating it to a more general model that includes perception, inference, and other familiar ways of knowing about the world. Primitivist theories, on the other hand, hold that the epistemology of introspection is primitive, or sui generis, and so cannot be explained as an instance of a more general model that applies elsewhere. According to the simple theory, introspective justification is a distinctive and sui generis kind of justification that cannot be explained on the model of perception, inference, or any other way of knowing about the world.

Having sketched the general motivations for the project, my goal in the next two sections is to develop an additional and more detailed argument for extending the simple theory from conscious experience to belief on the grounds that it best explains the irrationality of believing Moorean conjunctions.
2. Moore’s Paradox

As G. E. Moore observed, there is something patently “absurd” (Moore 1942, 1944), one might even say “Mooronic” (Koethe 1978), involved in asserting sentences such as the following:

(1) I went to the pictures last Tuesday but I don’t believe that I did. (1942: 543)
(2) I believe that he has gone out, but he has not. (1944: 204)

Indeed, Moore’s observation seems applicable to any sincere assertion of a sentence that has either of these syntactic forms:

(3) p, but I don’t believe that p. (The omissive form).
(4) I believe that p, but it’s not the case that p. (The commissive form.)

The problem of explaining what’s wrong with asserting these Moorean sentences (and others like them?) has become known as Moore’s paradox. The problem has an air of paradox about it because Moorean sentences are not contradictory and yet asserting them seems absurd or self-defeating in much the same way as asserting a contradiction. Moorean sentences can be true – after all, I am neither omniscient nor infallible, so I frequently fail to believe that p when p is true and believe that p when p is false. But although Moorean sentences can be true, and although often they are true, they cannot be sincerely asserted without absurdity. Moore’s paradox is the problem of explaining why this is so.

The literature on Moore’s paradox contains many attempted solutions – so many, in fact, that I cannot attempt an exhaustive survey. Nevertheless, my aim in this section is to enumerate various criteria of adequacy for a solution to Moore’s paradox and to argue that many of the leading proposals in the literature fail to satisfy these criteria. In the next section, I go on to argue that extending the simple theory of introspection from conscious experience to standing belief provides the resources for solving Moore’s paradox in a way that satisfies all of these criteria.

2.1. Assertion and Belief

The first constraint is that a solution to Moore’s paradox should generalize from assertion to belief. As Sydney Shoemaker (1995) observes, believing Moorean conjunctions seems absurd in much the same way as asserting them. Moreover, the absurdity does not depend upon whether or not one gives linguistic expression to one’s belief in the speech act of assertion. Thus, he writes:

What seems to me too little noticed is that there is something paradoxical or logically peculiar about the idea of someone’s believing the propositional content of a Moore-paradoxical sentence, whether or not the person gives linguistic expression to this belief. (1995: 213)
Since Moore's paradox is not a purely linguistic phenomenon, it cannot be given a purely linguistic solution. Moore (1944) suggests that in asserting a proposition, one "implies" that one believes it, and so in asserting that one does not believe it, one thereby contradicts what one has implied. Despite Moore's (1944: 542) claim that there is "nothing mysterious about this sense of 'imply'", it is not entirely clear what sense he has in mind. After all, there is no logical implication from 'p' to 'I believe that p' or 'I don't believe that not-p'. Nevertheless, many inspired by Moore have proposed that there is a contradiction between the content that is asserted and the content that is represented (Black 1952), implicated (Martinich 1980), or expressed (Rosenthal 1995) by the speech act of assertion itself. On a broadly Gricean account, for instance, the speech act of asserting that p expresses the belief that p in the sense that it is an action performed with the intention of producing in one's audience the belief that one believes that p. Without exploring the details of these linguistic proposals, however, the general problem that affects them all is that they fail to generalize in a way that explains the absurdity of Moorean belief.

A solution to Moore's paradox should not only generalize from assertion to belief; it should also explain the absurdity of Moorean assertions in terms of the absurdity of the Moorean beliefs that they express. After all, there is nothing wrong with uttering Moorean sentences when they do not express Moorean beliefs – for instance, in making a joke or a philosophical point. Wittgenstein (1980) gives the example of a railway announcer who does not believe the schedule he is obliged to report and so closes his announcement of the schedule with the disclaimer, "Personally, I don't believe it." Similarly, John Turri (2010) considers the proponent of eliminativism about folk psychology who states their thesis as follows: "Eliminativism is true: there are no beliefs; but of course, I don't believe it."

There is room for disagreement about whether these speech acts are genuine cases of assertion. If assertion is defined narrowly as a speech act that has the function of expressing belief, then Moorean assertions are always absurd, but these are not cases of Moorean assertion. If assertion is defined more broadly, on the other hand, then these are cases of Moorean assertion, but Moorean assertions are absurd only when they function to express Moorean beliefs. Either way, the absurdity of asserting Moorean conjunctions can be derived exclusively from the absurdity of believing them. The following strategy therefore recommends itself: if we can explain what is wrong with believing Moorean conjunctions, then we can derive an explanation of what is wrong with assertions that express those beliefs. As Shoemaker suggests, "An explanation of why one cannot assert a Moore-paradoxical sentence will come along for free, via the principle that what one can believe constrains what one can assert" (1995: 213).

2.2. Unifying Moorean Phenomena

The second constraint is that a solution to Moore's paradox should give a unified explanation of the full range of Moorean phenomena. In the first instance, it should
explain the absurdity of both omissive and commissive forms of Moorean belief in a more or less unified way. In addition, however, it should explain the absurdity of some related Moorean phenomena described below.

It is plausible that belief distributes over conjunction in the sense that one believes the conjuncts of any conjunction that one believes. However, it is much less plausible that belief is closed under conjunction in the sense that one believes the conjunction of every set of conjuncts that one believes. This makes room for the possibility that one might believe the individual conjuncts of a Moorean conjunction without believing the Moorean conjunction itself. Arguably, however, one cannot avoid the charge of Moorean absurdity just by failing to follow through the logical consequences of one’s beliefs in this way. Believing that $p$ while also believing that one does not believe that $p$ is not much less absurd than believing the Moorean conjunction that $p$ and one does not believe that $p$. Similarly, believing that one believes that $p$ while also believing that it’s not the case that $p$ is not much less absurd than believing the Moorean conjunction that one believes that $p$ and it’s not the case that $p$. Insofar as there is any difference between these cases in respect of absurdity, it is a difference in degree, not a difference in kind. Therefore, a solution to Moore’s paradox should explain the absurdity of believing the conjuncts of a Moorean conjunction regardless of whether or not one believes their conjunction.

Moreover, just as it is absurd to believe that $p$ while disbelieving that one believes that $p$, so it is absurd to believe that $p$ while withholding belief that one believes that $p$. And similarly, just as it is absurd to believe that one believes that $p$ while disbelieving that $p$, so it is absurd to believe that one believes that $p$ while withholding belief that $p$. This form of absurdity is more difficult to present in the form of a Moorean conjunction, since there is no speech act that naturally functions to express withholding in the way that assertion functions to express belief. But if an assertion of the form, ‘It is an open question whether $p$,’ can be used to express withholding, then we can add the following Moorean conjunctions:

\begin{align*}
(5) & \ p \text{ and it is an open question whether I believe that } p. \\
(6) & \ I \text{ believe that } p \text{ and it is an open question whether } p.
\end{align*}

Any solution to Moore’s paradox should explain why asserting or believing (5) and (6) is absurd in much the same way as asserting or believing (3) and (4) above. Moreover, this point is crucial for recognizing that the key to diagnosing Moorean absurdity is neither believing contradictions nor having contradictory beliefs, but rather the existence of rational conflict among one’s doxastic attitudes.

**2.3. Self-Conscious Belief**

The third constraint is that a solution to Moore’s paradox should explain the absurdity of all Moorean beliefs and not just those that are held self-consciously. A self-conscious belief is a belief that one also believes oneself to have; in other words, it is a matter of believing a proposition while also believing that one believes it. As
Baldwin (1990; see also Shoemaker 1995, Kriegel 2004) observes, Moorean conjunctions cannot be self-consciously believed without having contradictory beliefs or believing that one has contradictory beliefs. If I self-consciously believe a Moorean conjunction of the omissive form, ‘p and I don’t believe that p,’ then I believe that I believe that p and I believe also that I don’t believe that p. And if I self-consciously believe a Moorean conjunction of the commissive form, ‘I believe that p and it’s not the case that p,’ then I believe that I believe that p and I believe also that I believe that not-p. The problem with this strategy is that it does not explain what’s wrong with Moorean beliefs that are not held in a self-conscious way.

One might respond by insisting that all beliefs are held self-consciously – that is, by endorsing the following iterativity principle:

**The iterativity principle:** necessarily, if one believes that p, then one believes that one believes that p.

Unfortunately, however, this principle is highly implausible. First, it is inconsistent with the existence of conceptually primitive believers, including young children and higher animals, who have beliefs without possessing the concept of belief. Second, it is inconsistent with normal forms of human irrationality, such as repressed belief, in which one has beliefs without believing one has them. And third, it implies an infinite regress that seems well beyond the psychological capacities of any finite creature such that every belief is a member of an infinite hierarchy of increasingly complicated higher-order beliefs.

Shoemaker (1995) endorses a qualified version of the iterativity principle on which all fully rational belief is self-conscious belief. He writes:

...believing something commits one to believing that one believes it, in the sense that...if one believes something, and considers whether one does, one must, on pain of irrationality, believe that one believes it. (1995: 214)

On Shoemaker’s view, believing a proposition rationally commits one to believing that one believes it and, conversely, believing that one believes a proposition rationally commits one to believing it. Thus, he writes, “in a rational person belief that p brings with it belief that one believes that p” and, conversely, “a rational person who believes that she believes that p thereby believes that p” (1995: 225-6).

Shoemaker’s proposal can be articulated as follows:

**Shoemaker’s principle:** necessarily, if one is fully rational, then one believes that p if and only if one believes that one believes that p.

Shoemaker’s principle avoids the objections to the iterativity principle so long as enough weight is placed on the condition of full rationality. Moreover, Shoemaker’s principle can explain the phenomenon of Moorean absurdity as a manifestation of
irrationality or, at the very least, as a deviation from full rationality. If believing a proposition rationally commits one to believing that one believes it, and vice versa, then it is irrational to believe either the proposition that \( p \) or the proposition that one believes that \( p \), while disbelieving or withholding belief in the other.

Shoemaker's insight is that Moore's paradox is a symptom of the irrationality that results from failures of self-knowledge. On this view, the connection between rationality and self-knowledge is the key to solving Moore's paradox. The problem is that Shoemaker himself does not explain this connection between rationality and self-knowledge.\(^\text{10}\) In particular, he does not explain why it is a requirement of rationality that one believes that \( p \) if and only if one believes that one believes that \( p \). For this reason, Shoemaker's principle does not really provide a solution to Moore's paradox, but merely gives a sharper statement of the problem to be solved – that is, to explain why believing Moorean conjunctions is a manifestation of irrationality. This brings us to our next constraint on a solution to Moore's paradox.

2.4. Moorean Irrationality

The fourth constraint is that a solution to Moore's paradox should explain why believing Moorean conjunctions is irrational. The absurdity of believing Moorean conjunctions is plausibly diagnosed as a manifestation of irrationality, but the problem remains to explain why it is irrational, given that it cannot easily be assimilated to the irrationality of believing a contradiction.

One influential strategy here is to explain the irrationality of believing Moorean conjunctions by appealing to the phenomenon of transparency.\(^\text{11}\) The question whether I believe that \( p \) is sometimes said to be transparent to the question whether \( p \) in the sense that I am justified in answering the first question by answering the second question. Gareth Evans gives the following well-known illustration:

> 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?' 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 \( p \). (1982: 225)

According to Evans, this procedure for answering questions about one's beliefs can be summarized in terms of the following principle:

> Whenever you are in a position to assert that \( p \), you are ipso facto in a position to assert 'I believe that \( p \).' (1982: 226)

John Williams (2004) has argued that the irrationality of believing Moorean conjunctions can be explained in terms of the following closely related principle, which he takes to be inspired by this passage from Evans:
Evans’s principle: Whatever justifies me in believing that $p$ also justifies me in believing that I believe that $p$. (2004: 348)

Williams uses Evans’s principle in arguing that I cannot have justification to believe a Moorean conjunction of the omissive form, ‘$p$ and I don’t believe that $p$’. He assumes, plausibly enough, that I have justification to believe the conjunction only if I have justification to believe each of its conjuncts. If I have justification to believe the first conjunct, then it follows by Evans’s principle that I have justification to believe the negation of the second conjunct. So, if I also have justification to believe the second conjunct, then I have justification to believe contradictory propositions. According to Williams, however, “This is logically impossible, because anything that justifies me in believing that something is the case renders me unjustified in believing that it is not the case and vice versa” (2004: 352).

The problem with this explanation of Moorean irrationality is that Evans’s principle is false: it is refuted by counterexamples in which one’s total evidence justifies believing that $p$ without thereby justifying the belief that one believes that $p$. Evidence that justifies believing that $p$ may justifying believing that one does not believe that $p$ or even that one believes its negation. As Richard Moran (2001: 84) notes, “I can well imagine the accumulated evidence suggesting both that I believe that it’s raining, and that it is not in fact raining.” In that case, one’s total body of evidence justifies believing a Moorean conjunction and the question whether one believe that $p$ is not transparent to the question whether $p$ in the sense that one is justified in answering the first question by answering the second question.

Moran’s (2001, 2012) reaction is to claim that transparency corresponds to an ideal of rationality and that failures of transparency are manifestations of irrationality. Thus, he writes:

If we assume a subject with mastery of the first-person pronoun..., it will only be in conditions of compromised rationality that a person could believe that it’s raining, possess the concept of belief, and yet be unable to know her belief through reflection on the weather. (2012: 232)

As with Shoemaker’s proposal, however, the challenge is to explain why transparency is an ideal of rationality; that is, why it is irrational to believe Moorean conjunctions in cases where transparency fails. Moreover, this challenge presents something of a puzzle. The puzzle of transparency is to explain why it is rational to make the transition from believing that $p$ to believing that one believes that $p$ in light of the fact that evidence that justifies believing that $p$ need not thereby justify believing that one believes that $p$. In the remainder of this section, I will briefly consider two prominent responses to the puzzle of transparency before proposing my own solution in the following section.
Alex Byrne (2005) proposes a solution to the puzzle of transparency using the apparatus of epistemic rules and rule-following. An *epistemic rule* is a conditional of the form: ‘If conditions C obtain, believe that p.’ One *follows* an epistemic rule of this form on a particular occasion if and only if one believes *p* because one knows (and so believes) that conditions C obtain. Byrne claims that introspective knowledge of one’s beliefs can be acquired by following the transparent rule, BEL:

**BEL**: If *p*, believe that you believe that *p* (2005: 95)

The puzzle of transparency, as Byrne articulates it, is to explain why BEL is a good rule to follow – that is, why following the rule is sufficient to yield knowledge or justified belief – in light of the fact that *p* neither deductively entails nor inductively raises the probability that one believes that *p*. Byrne’s solution is that BEL is a good rule to follow because it is *self-verifying* in the sense that “if it is followed, the resulting second-order belief is true” (2005: 95).

The problem with Byrne’s solution is that not all self-verifying rules are good rules in the sense that following them is sufficient to yield knowledge or justified belief. Consider the rule: ‘If *x* is composed of H$_2$O molecules, then believe that *x* is water.’ This rule is self-verifying in the sense that if it is followed, then the resulting belief is guaranteed to be true, but following the rule is not sufficient to yield justified belief unless one has a justified belief that water is composed of H$_2$O molecules. Moran articulates the objection as follows: “Following a rule for belief...requires from the rule-follower some understanding of, and an endorsement of, the rational connection between the contents mentioned in the rule” (2012: 227).

Moran’s own response to the puzzle of transparency involves an account of the rational connection between the contents mentioned in the rule. He writes:

> The transition described in Transparency is not an inference from evidence about a particular person, but rather something more like a general presupposition of rational thought, to the effect that, from the first-person point of view, I must take what I believe about something to be the expression of my sense of the reasons relating to the content of that belief. As we have seen, it is a presupposition that may lapse in cases of compromised rationality, but at the same time, it is hard to see what could be more basic to rationality than the idea that I take the question of what I believe about X to be determinable by my reflection on X itself. (2012: 232)

The suggestion is that it is rational to make the transition from believing that *p* to believing that one believes that *p* insofar as one is defeasibly but rationally entitled to presume that one’s beliefs are formed and maintained in a way that reflects one’s assessment of the evidence that bears on the question whether *p*. In other words, the rationality of making the transition relies implicitly on a background assumption that one’s beliefs reflect one’s assessment of the evidence. Moran insists that the transition is not an inference, presumably on the grounds that one has a default
rational entitlement to make this assumption in the absence of specific reasons for doubt. Nevertheless, Moran’s account of the rationality of making the transition can be usefully reconstructed in the form of an inference of the following kind:

(1) The evidence justifies believing that \( p \).
(2) If the evidence justifies believing that \( p \), then I believe that \( p \).
(3) Therefore, I believe that \( p \).

This proposal raises many unanswered questions. Is it true that I have defeasible entitlement to presume that I am rational in the sense that my beliefs are formed and maintained in a way that appropriately reflects my assessment of the evidence? And, if so, what explains why I have this entitlement? Moreover, even if I have this entitlement, don’t I have standing defeaters in the form of background evidence that my beliefs are sensitive to all sorts of non-evidential factors, including the influence of prejudice and wishful thinking? And isn’t it rational for me to make the transition from believing that \( p \) to believing that I believe that \( p \) even in the face of background evidence of this kind?

No doubt, these proposals deserve more extended discussion than I have space to give them here. Nevertheless, the problems I have indicated motivate the search for an alternative solution to the puzzle of transparency. In the next section, I argue that the extending the simple theory of introspection from conscious experience to belief provides us with all the resources needed for solving Moore’s paradox and the closely related puzzle of transparency.

3. A Simple Solution to Moore’s Paradox

The puzzle of transparency is the problem of explaining why it is rational to make the transition from believing that \( p \) to believing that one believes that \( p \). As Shoemaker’s principle acknowledges, it is a requirement of rationality that one believes that \( p \) if and only if one believes that one believes that \( p \). But the residual problem is to explain why this is a requirement of rationality by appealing to more fundamental principles of justification. Williams proposes to explain this requirement of rationality by appealing to Evans’s principle that one’s evidence justifies believing that \( p \) if and only if it also justifies believing that one believes that \( p \). As we have seen, however, Evans’s principle is false and so an alternative explanation is needed.

Moran explains the rationality of the transition by claiming that one has a default rational entitlement to presume that one’s beliefs are responsive to one’s assessment of the evidence. This claim can be questioned, but even if it is granted, it seems too easily defeated by background evidence of one’s own irrationality. Byrne explains the rationality of the transition by citing the fact it is self-verifying in the sense that it guarantees true beliefs. Not all self-verifying transitions are rational, however, and so the rationality of the transition remains to be explained.
An alternative solution to the puzzle of transparency can be given by extending the simple theory of introspection from conscious experience to belief as follows:

One has introspective justification to believe that one believes that \( p \) iff one believes that \( p \).

The simple theory of introspection explains why the transition from believing that \( p \) to believing that one believes that \( p \) is not only reliable, but also a rational one. It is always rational for us to make the transition because believing that \( p \) ensures that one has introspective justification to believe that one believes that \( p \). On this view, the transition from believing that \( p \) to believing that one believes that \( p \) is justified in the same way as the transition from feeling pain to believing that one feels pain. It is not justified in the same way as an inference from believing that \( p \) to believing that \( q \) where \( p \) deductively entails or inductively raises the probability that \( q \). The rationality of the transition cannot be assimilated to the rationality of deductive or inductive inference, but is rather an instance of a more general pattern of rational transitions from mental states to beliefs about those mental states.

The simple theory also explains the irrationality of believing Moorean conjunctions of the omissive form, ‘\( p \) and I don’t believe that \( p \).’ If I believe the conjunction, then I believe each of its conjuncts. But if I believe the first conjunct, then I lack justification to believe the second conjunct, since I have introspective justification to believe its negation. Therefore, if I believe the conjunction, I lack justification for believing it and so I cannot believe it in a way that is justified.

In much the same way, the simple theory explains the irrationality of believing Moorean conjunctions of the commissive form, ‘I believe that \( p \) and it’s not the case that \( p \).’ If I believe the conjunction, then I believe each of its conjuncts. But I have justification to believe the first conjunct only if I believe that \( p \), so if I also believe the second conjunct, then I have inconsistent beliefs. Assuming that inconsistent beliefs cannot be held in a way that is justified, I cannot believe the conjunction in a way that is justified.

In this way, the simple theory explains the irrationality of believing Moorean conjunctions of the omissive and commissive forms. Given Shoemaker’s principle linking belief and assertion, it also explains the irrationality of asserting them. Moreover, the explanation generalizes to the full range of Moorean phenomena. In particular, it explains the irrationality of believing the conjuncts of Moorean conjunctions regardless of whether or not one believes their conjunction. And parallel arguments can be given to explain the irrationality of believing Moorean conjunctions of the following forms:

\[
\begin{align*}
(5) \ p \text{ and it is an open question whether I believe that } p. \\
(6) \ I \text{ believe that } p \text{ and it is an open question whether } p.
\end{align*}
\]
In short, the simple theory provides the materials for a solution to Moore’s paradox that satisfies all the criteria enumerated in the previous section.

The proposed solution to Moore’s paradox also satisfies an additional constraint that any adequate solution must meet. It explains why believing Moorean conjunctions is always unjustified, but without implying that one cannot have evidence that justifies believing Moorean conjunctions. To articulate this point more effectively, let us introduce a distinction between two senses of justification: a belief is *ex ante* justified just in case one has evidence that justifies holding the belief, while a belief is *ex post* justified if and only if one holds the belief in a way that is justified on the basis of the evidence. These are clearly distinct senses of justification, since one may have evidence that justifies holding a belief even if one does not hold the belief on the basis of one’s evidence in a way that is justified.

A solution to Moore’s paradox should explain why believing Moorean conjunctions is always *ex post* unjustified, but without implying that it is always *ex ante* unjustified. As Moran observes, there are cases in which one’s total evidence provides *ex ante* justification to believe Moorean conjunctions. For instance, suppose that one’s total evidence justifies believing that it will rain while also justifying the belief that one does not believe that it will rain. In that case, one has *ex ante* justification to believe the Moorean conjunction, ‘It will rain and I don’t believe that it will rain’. Nevertheless, one cannot believe this Moorean conjunction on the basis of one’s evidence in a way that is *ex post* justified. This is because the evidence that *ex ante* justifies believing the Moorean conjunction is destroyed in the act of coming to believe it. In other words, one’s evidence is *finkish* in the sense that it is destroyed in the act of trying to use it.14

The existence of finkish evidence has important consequences for understanding the relationship between *ex ante* and *ex post* senses of justification. Alvin Goldman suggests that *ex ante* justification can be analyzed in terms of *ex post* justification in the following way:

\[ S \text{ is } \textit{ex ante} \text{ justified in believing that } p \text{ at } t \text{ just in case his total cognitive state at } t \text{ is such that from that state he could come to believe } p \text{ in such a way that this belief would be } \textit{ex post} \text{ justified.} \ (1979: xx) \]

But Moorean conjunctions constitute a counterexample to this analysis. If S is *ex ante* justified in believing a Moorean conjunction, it does not follow that he could come to believe the Moorean conjunction in a way that would be *ex post* justified. The connection between *ex ante* and *ex post* justification is more accurately described as follows:

\[ S \text{ is } \textit{ex ante} \text{ justified in believing that } p \text{ at } t \text{ just in case his evidence at } t \text{ is such that if } S \text{ were to believe that } p \text{ in a way that is appropriately based on the evidence that provides his } \textit{ex ante} \text{ justification to believe that } p, \text{ then his belief would be } \textit{ex post} \text{ justified.} \]
A few observations. First, this is not a reductive analysis of *ex ante* justification in terms of *ex post* justification, since the notion of *ex ante* justification appears on the right-hand side. Second, the order of analysis is plausibly reversed, since *ex post* justification is analyzed in terms of *ex ante* justification plus appropriate basing, rather than the other way around. And third, we avoid counterexamples involving finkish evidence, since these are cases in which appropriate basing is impossible and so the analysis is vacuously true.

Many otherwise promising solutions to Moore’s paradox fail to satisfy this final constraint. For instance, Timothy Williamson claims that Moorean assertions are defective because they violate the knowledge rule for assertion:

One must: assert $p$ only if one knows $p$. (2000: 243)

Similarly, he suggests, Moorean beliefs are defective because they violate an analogous knowledge rule for belief:

It is plausible, nevertheless, that occurrently believing $p$ stands to asserting $p$ as the inner stands to the outer. If so, the knowledge rule for assertion corresponds to the norm that one should believe $p$ only if one knows $p$. Given that norm, it is not reasonable to believe $p$ when one knows that one does not know $p$. (2000: 255-6)

One cannot know a Moorean conjunction of the omissive form, ‘$p$ and I don’t believe that $p$,’ because one knows a conjunction only if one knows all of its conjuncts, but if one knows the first conjunct, then one believes it, and hence one cannot know the second conjunct, since it is false.\(^{15}\)

One might protest that this doesn’t explain the irrationality of believing Moorean conjunctions, since a belief can be rationally held and yet nevertheless fail to be knowledge. In fact, Williamson himself acknowledges this point:

The rule makes knowledge the condition for permissible assertion, not for reasonable assertion. One may reasonably do something impermissible because one reasonably but falsely believes it to be permissible. (2000: 256)

But Williamson also supplies a principle for deriving claims about what is reasonable from claims about what is permissible:

If one must ($\phi$ only if $p$ is true), then one should ($\phi$ only if one has evidence that $p$ is true). The transition from ‘must’ to ‘should’ represents the transition from what a rule forbids to what it provides a reason not to do. (2000: 245)

So, if it is permissible to believe that $p$ only if one knows that $p$, then it is reasonable to believe that $p$ only if one has evidence that one knows (or one is in a position to
know) that \( p \). Believing a Moorean conjunction is unreasonable, then, since one knows, or at least one is in a position to know, by means of a simple deductive argument that one cannot know a Moorean conjunction to be true. The problem with Williamson’s explanation is that it implies that it is never rational to believe a Moorean conjunction and yet, as we have seen, there are cases in which one’s evidence does make it rational to believe a Moorean conjunction. Therefore, Williamson’s explanation over-generates.

A similar problem affects the simple explanation proposed by John Williams (1996). If I believe a Moorean conjunction of the omissive form, ‘\( p \) and I don’t believe that \( p \),’ then I believe each of its conjuncts; but if I believe the first conjunct, then the second conjunct is false, and so I cannot believe the conjunction without thereby falsifying it. Similarly, if I believe a Moorean conjunction of the commissive form, ‘I believe that \( p \) and it’s not the case that \( p \),’ then I believe each of its conjuncts; but if I believe the second conjunct, then either the first conjunct is false or I have inconsistent beliefs, and so I cannot believe the conjunction without either falsifying it or having inconsistent beliefs. Thus, Williams concludes, “A Moorean believer is irrational either because his belief cannot be correct, or because it entails contradictory beliefs” (1996: 137).

Claudio de Almeida (2001) raises the objection that the mere fact that a proposition cannot be truly believed does not thereby make it irrational to believe it. For instance, it is not obviously irrational to believe a complex logical falsehood on the basis of misleading testimony from a recognized authority or in the light of an apparently persuasive, but subtly mistaken, proof. In order to explain the irrationality of believing a Moorean conjunction, Almeida claims, we need to appeal to a further principle of the following kind:

You rationally believe that \( p \) only if there is no simple and compelling argument that we can reasonably expect you to be aware of to the effect that your believing that \( p \) is either self-falsifying or ensures the presence of contradictory beliefs in your doxastic system. (2007: 54)

De Almeida’s objection is that it’s not clear that Williams’s argument is simple enough to warrant a charge of irrationality in every case. But I want to set that concern aside and press a different one. If the suggestion is that you cannot have justification to believe a Moorean conjunction because an argument of this kind is always available through reflection, then the suggestion runs counter to Moran’s observation that I do sometimes have justification to believe Moorean conjunctions. What is needed, and what this account fails to deliver, is an account of why believing Moorean conjunctions is always irrational \textit{ex post} despite the fact that it is sometimes \textit{ex ante} rational to believe them.
4. The Role of Conscious Experience

What is the role of conscious experience in explaining our introspective knowledge of what we believe? On many theories, the answer is: none. Beliefs are standing dispositional states, rather than occurrent conscious experiences. Indeed, this is one major source of skepticism about the prospects for extending the simple theory of introspection from conscious experience to belief. The simple theory is most plausible in the context of explaining our introspective knowledge of conscious experience and so the extension to belief requires further motivation. Moreover, extending the simple theory from conscious experience to belief imposes constraints on one’s theory of belief and its relationship with conscious experience.

On the account that I propose (see also Smithies 2012a & 2012d), conscious experience plays an indispensable role in explaining our introspective knowledge of what we believe. Although beliefs are standing dispositional states, rather than occurrent conscious experiences, they are disposed to cause occurrent conscious experiences of judgment. These conscious experiences of judgment can be individuated wholly by their phenomenal character. Meanwhile, beliefs can be individuated wholly by their dispositions to cause the phenomenal character of judgment. And I claim that it is in virtue of being phenomenally individuated in this way that beliefs fall within the scope of the simple theory of introspection.

One line of resistance to this proposal is skepticism about the claim that beliefs are disposed to cause conscious experiences of judgment that are individuated by their phenomenal character. But suppose this claim is granted. Another line of resistance is to the claim that beliefs are individuated wholly by their phenomenal dispositions, as opposed to their behavioral dispositions, where these are assumed to come apart. The central argument for this claim is that it is only if one’s beliefs are phenomenally individuated that they are introspectively accessible in the sense that one is always in a position to know by introspection what one believes. Given access internalism, one’s beliefs are suited to play an epistemic role in justifying other beliefs only if they are introspectively accessible in this sense. Since beliefs play an epistemic role in justifying other beliefs, it follows that all beliefs are phenomenally individuated. This argument can be summarized as follows:

(1) An intentional state is a belief only if it plays an epistemic role.
(2) An intentional state plays an epistemic role only if it is introspectively accessible.
(3) An intentional state is introspectively accessible only if it is phenomenally individuated.
(4) Therefore, an intentional state is a belief only if it is phenomenally individuated.

My aim in this section is to compare and contrast two competing accounts of the role of conscious experience in explaining introspective knowledge of belief. On my account, believing that \( p \) is the source of one’s introspective justification to believe
that one believes that \( p \). On an alternative account, judging that \( p \) is the source of one’s introspective justification to believe that one believes that \( p \). Against this alternative account, I will argue that judging that \( p \) is neither necessary nor sufficient for having introspective justification to believe that one believes that \( p \). One need not judge that \( p \), so long as one is disposed to judge that \( p \). Moreover, judging that \( p \) is not sufficient unless it expresses a stable disposition to judge that \( p \). So, I conclude, the source of one’s introspective justification is not judging that \( p \), but rather believing that \( p \) in a way that disposes one to judge that \( p \).

First, judging that \( p \) is not *necessary* for having introspective justification to believe that one believes that \( p \). Aaron Zimmerman puts the point as follows:

> Again consider my mental state on June 10th while I am eating and thinking of nothing but the food before me. Do I then know that I believe that it is June? Of course I do. Am I then justified in believing that I believe it is June? Of course. But because I am not then occurrently judging that it is June, I seem to lack what [this account] says I need if I am to be justified in believing that I believe that it is June. (2006: 357-8)

What justifies my second-order knowledge that I believe that it is June while I am preoccupied with thinking about my food? There need be no judgment that I am making at that time, or at any earlier time, that justifies my second-order belief. So, there is pressure to acknowledge that standing second-order beliefs can be introspectively justified by standing first-order beliefs without those first-order beliefs manifesting themselves in judgment. One might deny that one has introspective justification for second-order beliefs about one’s first-order beliefs; instead, perhaps, one is merely disposed to have introspective justification for second-order beliefs under conditions in which one’s first-order beliefs are manifested in judgment. But since knowledge requires justification, this implies a more general skepticism about my standing knowledge of what I believe.

As Christopher Peacocke observes, there are also cases in which one makes a second-order judgment that expresses knowledge of what one believes, although it is not based on any first-order judgment that expresses what one believes. Peacocke calls these “NICS” cases because the second-order judgment is based on no intermediate conscious state. He writes:

> There are occasions on which a person expresses a first-order belief, or indeed makes a self-ascription of a belief, and in which these are not consciously-based in the way I have been discussing. Most of us, when it becomes conversationally appropriate to say ‘I know my name is NN’, or ‘I know my address is such-and-such’, have no need to wait upon its surfacing in consciousness what our names and addresses are. We make these utterances intentionally and knowledgeably, but not because it has just occurred to us that our names and addresses are such-and-such. (1998: 91)
Peacocke argues that a requirement of first-order ratifiability is needed to explain why NICS self-ascriptions are justified and hence can amount to knowledge:

A NICS self-ascription of the belief that \( p \) is knowledge only if it is made in circumstances in which the thinker is also willing to make the first-order judgement that \( p \). (1998: xxx)

According to the requirement of first-order ratifiability, what is required for knowing that one believes that \( p \) is not actually making the judgment that \( p \), but merely being disposed to make it. That is exactly the view I am recommending.

Second, judging that \( p \) is not sufficient for having introspective justification to believe that one believes that \( p \). On this point, I disagree with Peacocke. On the one hand, he claims that judging that \( p \) makes it rational to self-ascribe the belief that \( p \). On the other hand, he claims that this way of self-ascribing beliefs is fallible, since one’s judgments do not always reflect one’s beliefs:

Someone can make a judgement, and for good reasons, but it not have the effects that judgement normally do – in particular, it may not result in a stored belief which has the proper influence on other judgements and on action.... So the ways of coming to make self-ascriptions which I have been discussing are by no means infallible. (1998: 90)

One might object to Peacocke’s account, as Zimmerman (2006) does, on the grounds that if judging that \( p \) is not sufficient for believing that \( p \), then it cannot provide immediate justification to believe that one believes that \( p \). But Nicholas Silins (2012) defends this account by arguing that immediate justification can be fallible and defeasible in the introspective domain as it is in the perceptual domain. He proposes the following sufficient condition for introspective justification:

**Silins's principle:** If you judge that \( p \), then your judgment that \( p \) gives you immediate fallible justification to believe that you believe that \( p \). (2012: 309)

Silins argues for this principle on the grounds that it explains why Moorean judgments tend to be irrational. Suppose I judge a Moorean conjunction of the omissive form, ‘\( p \) but I don’t believe that \( p \).’ In judging that \( p \), I acquire justification to believe that I believe that \( p \) and so I lack justification to believe that I do not believe that \( p \). Similarly, suppose I judge a Moorean conjunction of the commissive form, ‘I believe that \( p \) but not-\( p \).’ In judging that not-\( p \), I acquire justification to believe that I believe that not-\( p \) and so either I lack justification to believe that I believe that \( p \) or I have justification to believe that I have inconsistent beliefs. Either way, the irrationality of Moorean judgment can be explained.

Nevertheless, I want to argue that Silins’s principle cannot explain the full range of Moorean phenomena. Silins’s principle departs from the simple theory in allowing for cases in which one has justification to believe that one believes that \( p \), although
one does not believe that $p$; and cases in which one believes that $p$, although one does not have justification to believe that one believes that $p$. As I argue below, this makes room for Moorean phenomena that Silins’s principle cannot explain.

Here is the first problem case. Silins claims that judging that $p$ provides *defeasible* justification to believe that you believe that $p$. So, he allows for cases in which you judge that $p$, but you have misleading evidence that your judgment does not express what you believe, and so your defeasible justification to believe that you believe that $p$ is defeated. In that case, if you are rational, then you believe that $p$, but you do not believe that you believe that $p$. What then explains the irrationality of believing a Moorean conjunction of the omissive form, ‘$p$ and I don’t believe that $p’$ or ‘$p$ and it’s an open question whether I believe that $p’$?

Here is the second problem case. Silins claims that judging that $p$ provides *fallible* justification to believe that you believe that $p$. So, he allows for cases in which you judge that $p$, and so you have justification to believe that you believe that $p$, although in fact your judgment does not express what you believe. In that case, if you are rational, then you believe that you believe that $p$, but you do not believe that $p$. What then explains the irrationality of believing a Moorean conjunction of the commissive form, ‘I believe that $p$ and it’s not the case that $p$’ or ‘I believe that $p$ and it’s an open question whether $p’$?

Silins’s principle is designed to explain the irrationality of *judging* Moorean conjunctions. Plausibly, however, there is always some degree of irrationality involved in *believing* a Moorean conjunction, or believing its conjuncts, even if these beliefs are not actually manifested in judgment. Perhaps Silins would maintain that believing a Moorean conjunction is not always irrational. But this is a bitter pill to swallow. There always seems to be some degree of irrationality involved when one’s dispositions to judge that $p$ come apart from one’s dispositions to judge that one believes that $p$. If one is rational, then one’s dispositions to judge that $p$ march in step with one’s dispositions to judge that one believes that $p$; in other words, one’s first-order beliefs march in step with one’s second-order beliefs. An adequate solution to Moore’s paradox should therefore explain the irrationality of Moorean belief and not merely its expression in Moorean judgment or assertion.

To sum up, I have argued that judging that $p$ is neither necessary nor sufficient for having introspective justification to believe that one believes that $p$. In one case, one has a disposition to judge that $p$ that is not manifested in judgment. In the other case, one judges that $p$ in a way that fails to reflect any stable disposition towards judgment. The obvious moral to draw is that a stable disposition to judge that $p$ is both necessary and sufficient for having introspective justification to believe that one believes that $p$; after all, this is present in the counterexample to necessity and absent in the counterexample to sufficiency. Moreover, if we identify believing that $p$ with having a stable disposition to judge that $p$, then we arrive at the simple theory – that is, believing that $p$ is both necessary and sufficient for having introspective justification to believe that one believes that $p$. This, in turn, explains why it is a
requirement of rationality that one’s dispositions to judge that $p$ march in step with one’s dispositions to judge that one believes that $p$.

5. Epistemic Idealization

In this final section, I want to conclude by defending the simple theory of introspection against some objections. I am primarily concerned with objections that take the following form: first, various consequences of the simple theory are identified; and second, they are rejected on grounds of implausibility. For instance, the simple theory of introspection implies all of the following commitments:

(1) One can have introspective justification for beliefs about one’s beliefs that one cannot use in forming justified beliefs about one’s beliefs.
(2) One cannot have introspective justification for false beliefs about one’s beliefs.
(3) One cannot have introspective justification for agnosticism about one’s beliefs.
(4) One’s introspective justification for beliefs about one’s beliefs cannot be defeated by non-introspective sources of justification.
(5) One cannot have non-introspective sources of justification for beliefs about one’s beliefs.

One might reject the simple theory of introspection on the grounds that these are highly implausibly commitments. But before doing so, it is worth noting that theories of a priori justification that impose logical omniscience requirements incur structurally parallel commitments:

(1) One can have a priori justification for beliefs about logic that one cannot use in forming justified beliefs about logic.
(2) One cannot have a priori justification for false beliefs about logic.
(3) One cannot have a priori justification for agnosticism about logic.
(4) One’s a priori justification for beliefs about logic cannot be defeated by a posteriori sources of justification.
(5) One cannot have a posteriori sources of justification for beliefs about logic.

Moreover, logical omniscience requirements are built into all of our most sophisticated formal theories of belief revision. Therefore, we shouldn’t be too quick to regard objections of this kind as decisive.

In my view, these objections are flawed because they overlook the role of epistemic idealization in a theory of justification. One of the goals of a theory of justification is to capture a certain kind of epistemic ideal. Although we often (perhaps always) fall short of these ideal standards for justification, our beliefs may nevertheless satisfy ordinary standards of justification, so long as they approximate sufficiently closely towards ideal standards to meet some contextually determined threshold that takes into account the limitations on our capacities. So, we need to distinguish between

22
ideal standards of justification which abstract away from our limited capacities, and ordinary standards of justification that take these limited capacities into consideration, where the latter are best understood as closer or more distant approximations towards the former. 17

The commitments of the simple theory of introspection can be mitigated by appealing to this distinction between ideal and non-ideal standards of justification. Consider the following example from Frank Jackson:

A retired university dean, in the course of writing his biography, may notice a pattern in his appointment and promotion decisions that reveals to him that he is biased against women. He may be shocked and surprised by this discovery; he has always believed that he believes that women are the intellectual equals of men. Examples like these tell us that sometimes we discover what we ourselves believe by observing and making inferences from our behavior, and also that our beliefs about what we ourselves believe can be mistaken. (2013: 60)

A complication of this example is that one’s behavioral biases need not always reflect one’s beliefs: someone might sometimes act as if they believe that something even if they do not really believe it. But let this be a case in which the Dean genuinely believes and is disposed to judge that women are intellectually inferior, although the belief is repressed and so it is not manifested in judgment except in unusual circumstances – perhaps after therapy or after too many drinks. The simple theory implies that the Dean has introspective justification to believe that he believes this. Is this verdict on the case defensible?

I would claim that this verdict is in fact crucial for explaining a certain kind of Moorean irrationality. Here, it is irrelevant that the Dean’s belief is unjustified by the evidence; we could easily switch examples to a different belief that is well supported by the evidence, but repressed in a way that blocks its disposition to manifest itself in judgment. The relevant defect in rationality consists in the fact that the Dean is disposed to judge each of the conjuncts of the Moorean conjunction, ‘p and I don’t believe that p.’ Admittedly, the Dean may not be disposed to judge them in the same context, and so he may not be disposed to judge the Moorean conjunction, but that is not enough to absolve him entirely of irrationality.

On the simple theory, the Dean has introspective justification to believe that he believes that women are intellectually inferior. But the fact that his first-order belief is repressed creates a “rational obstacle” (in the sense of Pryor 2004) that prevents him from forming an introspectively justified second-order belief. This second-order belief is justified by ideal standards, but the Dean is unable to satisfy these ideal standards because his first-order belief is repressed. Nevertheless, we can also recognize ordinary standards of justification that take these limitations into account. Given the fact that his belief is repressed, the Dean may be justified by ordinary standards in disbelieving or withholding belief that he has this belief. Moreover,
since the repression imposes a rational obstacle that prevents him from exploiting introspective sources of justification, other sources of justification (such as observation of behavior or the testimony of a therapist) may become relevant in determining what he is justified in believing by ordinary standards. So, the distinction between ideal standards and ordinary standards of justification provides the resources needed for defending the simple theory against objections of this kind.

Acknowledgements

I have learned from discussion of this material with participants in a seminar on introspection and self-knowledge at the Ohio State University in Winter 2011 and with audiences at the Philosophy of Mind Work in Progress Group at the University of Oxford in June 2013 and the SPAWN conference on the transparency of belief and perception at the University of Syracuse in August 2013. Thanks especially to David Barnett, Matthew Benton, Matthew Parrott, and Nico Silins for comments and discussion.

References


1 Does Ryle himself endorse this Rylean view? Some passages suggest that he does, but others point in the opposite direction, including his discussion of “silent soliloquy” or inner speech. In any case, I will side-step this exegetical question by remaining neutral on whether Ryle himself is a Rylean.
2 Smithies (2013) argues that the simple theory is consistent with Schwitzgebel’s (2011, 2012) claims about the unreliability of introspection and the absence of any dedicated psychological mechanism that subserves introspection.
3 This is an example of what Alston (1971) calls the truth-sufficiency account of privileged access, according to which S has privileged access to the fact that p if p is itself a justification for S to believe that p. See also Peacocke (1998), Zimmerman (2006), Shoemaker (2009), and Neta (2011).
4 See Smithies (2012a) for some brief discussion of these assumptions.
5 See Smithies (2012b, 2014a) for arguments for access internalism and Smithies (2012c) for a defence of access internalism against Williamson’s (2000: Ch.4) anti-luminosity argument.
6 See Smithies (2012a) for this distinction. Paul Boghossian (1989) endorses reductionism in claiming that introspective knowledge of our own minds is based either upon observation, inference, or nothing at all, while Christopher Peacocke (1998) gives expression to primitivism in rejecting this as a “spurious trilemma”.
7 Sorensen (1988) gives examples of Moorean sentences, such as ‘there are no beliefs’ or ‘God knows that I’m an atheist’, that do not have the syntactic features of either omissive or commissive forms.
8 See the introduction to Green and Williams (2007) for a historical overview of Moore’s paradox and a survey of proposed solutions.
9 Conceptual impoverishment and repression can be regarded as departures from full rationality. The regress problem remains, but the infinite regress of higher-order beliefs can be regarded as an ideal requirement of full rationality that finite creatures can approximate towards in varying degrees. See Smithies (2012b, 2014a) for discussion of a similar regress problem for access internalism.
10 Shoemaker actually takes a somewhat deflationary attitude towards the rationality condition: “I am inclined to think that the rationality required here is no more than the minimal rationality needed in order to be the subject of beliefs and desires; and if that is so the qualifier ‘rational’ can be omitted.” (1995: 226)
12 Williams explains the commissive form by relying on a variant of Evans’s principle that he claims to hold given the assumption that I am rational and reflective: “Whatever justifies me in believing that p also justifies me in believing that I do not believe that not-p” (2004: 352).
13 The argument here relies on the assumption that if I have justification to believe a proposition, then I do not also have justification to disbelieve or to withhold belief in the very same proposition. This is what Kelly (forthcoming) calls an ‘intrapersonal’ version of uniqueness, which he distinguishes from an ‘interpersonal’ version of uniqueness that is more controversial.
14 Compare Martin (1994) on finkish dispositions that are destroyed when their manifestation conditions obtain.
15 How does this account generalize to the commissive form? Here is a suggestion: one cannot know a Moorean conjunction of the commissive form, ‘I believe that p and it’s not the case that p;’ because if one knows the second conjunct, then either the first conjunct is false or one has inconsistent beliefs, and neither of a pair of inconsistent beliefs can be knowledge.
16 See Silins (2012: 303, n.11) for the claim that not all Moorean judgments are irrational.
17 See Smithies (2014a, 2014b) for more detailed discussion of epistemic idealization.