Belief and Self-Knowledge:
Lessons from Moore's Paradox

Declan Smithies
The Ohio State University

How do we know what we believe? Gilbert Ryle is often credited with the view that we know what we believe in much the same way that we know what others believe: namely, by inference from observation of behavior. This Rylean view cannot explain all of our self-knowledge, however, since we can know what we believe even when our beliefs make no relevant causal impact on our behavior.[1]

In opposition to this Rylean view, I will assume that there is an epistemic asymmetry to be drawn between first-personal and third-personal ways of knowing what we believe. Each of us has some way of knowing what we ourselves believe that is *peculiar* in the sense that it is different from any of our ways of knowing what others believe. What is more, I will argue, this first-personal way of knowing what we believe is *privileged* in the sense that it is immune from the rational uncertainty and error that affects our ways of knowing what others believe.[2]

What is this first-personal way of knowing what we believe? A preliminary answer is that we know what we believe by *introspection*. However, the term 'introspection' is nothing more than a placeholder for an account of how we know what we believe. We certainly cannot assume that our knowledge of what we believe has its source in anything like inner perception. So the task remains for a theory of introspection to fill in this placeholder by giving an informative account of our first-personal way of knowing what we believe.

The aim of this paper is to argue that what I call the *simple theory of introspection* can be extended to account for our introspective knowledge of what

[1] See Boghossian 1989: 67 for a statement of this objection. I won't discuss whether the Rylean view is correctly attributed to Ryle himself, but see his 1949: Ch. 6.
[2] I borrow the terminology of peculiar and privileged access from Byrne 2005: 80-2.

we believe as well as what we experience.[3] In section one, I present the simple theory of introspection and motivate the extension from experience to belief. In section two, I argue that extending the simple theory provides a solution to Moore's paradox by explaining why believing Moorean conjunctions always involves some degree of irrationality. In section three, I argue that it also solves the puzzle of transparency by explaining why it's rational to answer the question whether one believes that *p* by answering the question whether *p*. Finally, in section four, I defend the simple theory against objections by arguing that self-knowledge constitutes an ideal of rationality.

1. The Simple Theory of Introspection

It is sometimes assumed that our knowledge of what we believe, like our knowledge of what others believe, must be based upon observation, inference, or nothing at all. Paul Boghossian (1989) relies on this assumption to generate a skeptical paradox by arguing that none of these options can explain how we know what we believe. According to Boghossian, an adequate response to the skeptical paradox requires showing that one of these three options can be made to work after all. But a more promising option is to reject the starting assumption that our knowledge of what we believe is based on observation, inference, or else nothing at all. This is the option taken by Christopher Peacocke, who calls it "the spurious trilemma" (1998: 83).

Peacocke argues that you can know that you're in a conscious state by forming a belief on the basis of that very conscious state. This is what he calls a "consciously-based self-ascription". For example, you can know that you're in pain by believing that you're in pain on the basis of the conscious state of being in pain. This knowledge is not based on observation or inference, but that doesn't mean that it's not based on reasons at all. According to Peacocke, merely being in pain gives you a reason to believe that you're in pain. Thus, he writes, "An experience of pain can be a thinker's reason for judging that he is in pain" (1998: 72).

[3] I first introduced the simple theory of introspection in Smithies 2012a, but see also Peacocke 1998, Pryor 2005, Zimmerman 2006, Shoemaker 2009, and Neta 2011 for closely related proposals.

The simple theory of introspection extends this proposal. It says that you know by means of introspection that you're in some mental state M when you believe that you're in M on the basis of a reason that is constituted by the fact that you're in M. On this view, the mere fact that you're in M gives you a reason to believe that you're in M and thereby puts you in a position to know that you're in M.[4]

Introspection provides knowledge of negative as well as positive facts about conscious experience. Just as I can know by introspection that I'm in pain when I am, so I can know by introspection that I'm not in pain when I'm not. Can we say that the fact that I'm not in pain is my reason for believing that I'm in pain? The problem is that my laptop is not in pain, but it doesn't thereby have a reason to believe that it's not in pain. We can say instead that my reason for believing that I'm in pain is the fact that I'm in some total mental state that includes being in pain, whereas my reason to believe that I'm not in pain is the fact that I'm in some total mental state that excludes being in pain. My laptop doesn't satisfy either of these conditions.

In general, having a reason to believe that *p* is not sufficient for being in a position to know that *p*. According to the simple theory, however, an introspective reason for believing that you are in M is constituted by the very fact you are in M. This fact is *self-intimating* in the sense that if you're in M, then you thereby have an introspective reason to believe that you're in M. Moreover, this reason is *infallible* in the sense that if you have this reason to believe that you're in M, then it's true that you're in M. I claim that this reason is *indefeasible* because it cannot be defeated by evidence that you have them reason when it's false that you're in M. And it is *immune from Gettier cases* because it cannot be that you have this reason when it's accidentally true that you're in M. The upshot is that having an introspective reason to believe that you're in a mental state puts you in a position to know by means of introspection that you're in that mental state.

Which facts about our mental states can be known by introspection? We cannot know all of our mental states by introspection. Psychophysical experiments

[4] Are reasons facts or mental states? I say that reasons are constituted by facts about your mental states, whereas Peacocke says that reasons are constituted by your mental states themselves, but the difference will not be crucial here.

show that our performance on word completion tasks is affected by unconscious representations of masked stimuli – for instance, subjects primed with the word 'reason' are more likely to complete the word stem 'rea-' as 'reason' than 'reader'. We can't know by introspection that we represent these masked stimuli, since these mental representations are unconscious. We know about them only on the basis of psychophysical experiments, and not on the basis of introspection.[5]

Plausibly, we can know our *conscious experiences* by means of introspection. If we restrict the simple theory of introspection to our conscious experiences, then we are committed to the following thesis:

> *The restricted thesis*: for any conscious experience E, one has an introspective reason to believe that one has E if and only if one has E.

My target question in this paper is whether we can extend the simple theory from conscious experience to standing belief so as to yield the thesis below:

> *The extended thesis*: for any proposition *p*, one has an introspective reason to believe that one believes that *p* if and only if one believes that *p*.

In this section, I'll briefly sketch three arguments for extending the simple theory to include the extended thesis as well as the restricted thesis. In the next section, I'll argue that the extended thesis provides a solution to Moore's paradox.

First, the extended thesis explains how we can have standing knowledge of what we believe.[6] We have standing knowledge about many things: for instance, I know that I live in Columbus, and I retain this knowledge even when I'm not consciously thinking about where I live. We also have standing knowledge about our beliefs: for instance, I know that I believe that I live in Columbus, and I retain this knowledge even when I'm not consciously thinking about what I believe. But what

[5] See Marcel 1983 and Reingold and Merikle 1990 for experimental data on effects of masked priming in unconscious perception.
[6] Compare Zimmerman 2006: 357-61 and Shoemaker 2009: 48-50.

explains our standing knowledge of what we believe? I cannot know that *p* unless I have some reason for believing that *p*, but then what is my reason for believing that I believe that I live in Columbus? The problem is that I can retain my knowledge of what I believe even when there is nothing in the stream of conscious experience that bears on the question of what I believe. The extended thesis solves this problem because my standing belief that I live in Columbus gives me a standing reason for believing that I believe that I live in Columbus. When my standing second-order belief is properly based on the reason provided by my standing first-order belief, then it constitutes introspective knowledge.[7]

Second, the extended thesis is needed for explaining *access internalism* in epistemology.[8] To a first approximation, access internalism is the thesis that one is always in a position to know which propositions it is rational for one to believe by means of introspection and a priori reflection alone. The rationale for access internalism is that the facts about which propositions it is rational for one to believe at any given time are determined by facts about one's mental states at that time. Moreover, one is always in a position to know these facts about one's mental states through introspection when they are relevant in determining which propositions it is rational for one to believe. But facts about one's beliefs, as well as facts about one's conscious experiences, are relevant in determining which propositions it is rational for one to believe. Therefore, explaining access internalism requires assuming that one is always in a position to know all the facts about which beliefs one has, as well as which conscious experiences one has, at any given time.

Third, the extended thesis explains a plausible connection between rationality and self-knowledge. On the one hand, the *new evil demon problem* for externalist theories of rationality relies on the intuition that a Cartesian demon can deceive me about the external world without thereby impugning my rationality. On the other hand, the *isolation problem* for coherentist theories of rationality relies on

---

[7] Note that introspective knowledge need not be based on any conscious activity of "introspecting" either in the present or at any past time.

[8] See BonJour 1985 for a classic defense of access internalism. I argue for access internalism and respond to objections in Smithies 2012b, 2015a, and forthcoming.

the intuition that a Cartesian demon *cannot* deceive me about my own beliefs and experiences without thereby impugning my rationality. Combining these two intuitions reveals an important asymmetry in our concept of rationality. Rationality requires knowledge of the internal world, but not the external world.[9]

In effect, my goal in what follows is to bolster this third argument by using Moore's paradox to support the connection between rationality and self-knowledge in the special case of belief. The key idea is that failing to know what you believe results in a Moorean predicament that seems quite irrational. If we assume that rationality requires self-knowledge, then we can explain why this kind of Moorean predicament is as irrational as it seems.

2. Moore's Paradox

G. E. Moore observed that there is something patently "absurd" – 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 assertion of a sentence that has one of the following 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 why it's absurd to assert Moorean sentences of these forms has become known as *Moore's paradox*.[10]

---

[9] See Cohen 1984 for the new evil demon problem and Sosa 1991: 136 for the isolation problem. Silins (forthcoming) argues that the new evil demon problem arises for the internal world too. I plan to address his argument elsewhere.

[10] The paradox was named by Wittgenstein 1953: 190. Moore mentions the paradox in his 1942: 540-3 and 1944: 204, but his most detailed discussion is in Moore 1993. See Green and Williams 2007 for a historical introduction to Moore's paradox.

There is a paradox here because asserting Moorean sentences seems absurd or self-defeating in much the same way as asserting a contradiction, and yet Moorean assertions are not contradictions; after all, they can be true. Since I am neither omniscient nor infallible, it can be true that *p* when I don't believe that *p* and it can be false that *p* when I believe that *p*. But although Moorean sentences can be true, they cannot be asserted without absurdity. Moore's paradox is the problem of explaining why this is so. As Moore says, "It is a paradox that it should be perfectly absurd to *utter assertively words* of which the *meaning* is something which may well be true – is not a contradiction" (1993: 209).

2.1. Linguistic Solutions

Since Moore, it has been widely noted that believing Moorean conjunctions is absurd in much the same way as asserting them. Moreover, it is absurd to believe Moorean conjunctions whether or not one gives linguistic expression to one's belief in the speech act of assertion. If Moore's paradox is not a purely linguistic phenomenon, then it cannot be given a purely linguistic solution. This undermines many of the earliest solutions to Moore's paradox, including those originally proposed by Moore and Wittgenstein.

Moore claims that there is a sense in which one contradicts oneself by asserting a conjunction of the omissive form. In asserting that *p*, one "implies" that one believes that *p*, and so in asserting that one does not believe that *p*, one thereby contradicts what one has implied. Of course, there is no *logical* implication from the assertion that *p* to the conclusion that one believes that *p*. Moore's claim is rather that asserting that *p* reliably indicates that one believes that *p*. On a Gricean account, asserting that *p* functions to *express* that one believes that *p* because assertion is a speech act that is performed with the intention of causing one's audience to believe that one believes that *p*. So, following Moore, one might hold that there is a contradiction between the content that is asserted and the content that is expressed in the act of making the assertion.[11]

[11] See Baldwin 1990: 228 and Rosenthal 1995 for neo-Moorean solutions.

In contrast, Wittgenstein claims that in asserting that I believe that *p*, I thereby assert that *p*.[12] On this view, one asserts a contradiction by asserting a conjunction of the commissive form. Inspired by Wittgenstein, Jane Heal (1994) claims that asserting that I believe that *p* functions to *express* and not merely to *report* the belief that *p*. On this view, I contradict myself by asserting a conjunction of the commissive form insofar as the belief that I express in reporting that I believe that *p* contradicts the belief that I express in asserting that *p*.

There are problems for both Moorean and Wittgensteinian solutions to the paradox. On the one hand, the Moorean solution cannot easily be extended from the omissive form to the commissive form. Does asserting that *p* express not only that I believe that *p*, but also that I don't believe its negation? On the other hand, the Wittgensteinian solution cannot easily be extended from the commissive form to the omissive form. Does asserting that I don't believe that *p* express that I believe its negation? A more serious problem for both accounts is that it's not clear how to extend either of them from Moorean assertion to Moorean belief. What's wrong with believing Moorean conjunctions without asserting them?

A solution to Moore's paradox should not only generalize from assertion to belief, but it should also explain the absurdity of Moorean assertions in terms of the absurdity of the Moorean beliefs that they express. There is nothing wrong with uttering Moorean sentences in performing speech acts that don't express Moorean beliefs – for instance, in making a joke or a philosophical point. Wittgenstein gives the example of a railway employee who concludes his announcement of the schedule with the skeptical disclaimer, "Personally, I don't believe it" (1980: 84). There is no absurdity here because the announcement doesn't function to express what the speaker believes. The same applies to Andre Gallois' example of an eliminativist about belief, who says, "Neurophysiology is the key to understanding the mind, but I do not believe that it is" (1996: 52). Her speech act doesn't function to express what she believes, since it is not performed with the intention of causing

[12] See Wittgenstein 1953: 190-2 and 1980: 90-6.

her audience to believe that she believes what she is saying. Indeed, the speech act is intended to have precisely the opposite effect.

If assertion is defined narrowly as a speech act that functions to express belief, then all Moorean assertions are absurd, but not all utterances of Moorean sentences are assertions. If assertion is defined more broadly, then Moorean assertions are absurd only when they have this function. Either way, the absurdity of Moorean assertions can be derived exclusively from the absurdity of the Moorean beliefs that they express. As Sydney Shoemaker writes, "An explanation of why one cannot (coherently) assert a Moore-paradoxical sentence will come along for free, via the principle that what can be (coherently) believed constrains what can be (coherently) asserted" (1996: 76).

2.2. Psychological and Epistemological Solutions

What is wrong with believing Moorean conjunctions? We can draw a distinction between *psychological solutions*, which claim that believing Moorean conjunctions is psychologically impossible, and *epistemological solutions*, which claim that believing Moorean conjunctions is epistemically irrational. I'll begin with some reasons for skepticism about the prospects for a psychological solution before exploring the options for an epistemological solution in more detail.[13]

First, it's not clear that there is any proposition so absurd that believing it is beyond the realm of human possibility. Some human beings believe some very strange things, especially those who reside in psychiatric hospitals and philosophy departments. Patients with Cotard's syndrome believe they are dead. Eliminativists believe they have no beliefs. Dialetheists believe that some contradictions are true. It's not at all clear that believing Moorean conjunctions is humanly impossible.[14]

---

[13] See Hintikka 1962: 67 and Shoemaker 1996: 85-6 for the claim that it's impossible to believe an omissive Moorean conjunction.

[14] Are eliminativists and dialetheists irrational? Surely not in the same sense as delusional patients. I claim that they are rational by non-ideal standards, although Moorean incoherence and logical incoherence always constitutes some departure from ideal rationality. See section 4 for more on the distinction between ideal and non-ideal standards of rationality.

Second, it's not clear how to motivate the claim that believing Moorean conjunctions is humanly impossible. Some philosophers have argued that there are rationality constraints that impose limits on how much irrationality is consistent with having beliefs at all.[15] One might argue that believing Moorean conjunctions is impossible on the grounds that it violates these rationality constraints. But this is a risky argument, since we know from empirical studies of human reasoning that any plausible rationality constraints must be weak enough to allow for a considerable degree of human irrationality.[16] As a result, believing Moorean conjunctions may be humanly possible even if it always involves some degree of irrationality.

Third, it's humanly possible to believe the conjuncts of a Moorean conjunction without conjoining them, but it's not rational, so an epistemological puzzle remains. Generally speaking, if it's rational to believe that *p*, and it's rational to believe that *q*, then it's also rational to believe the conjunction that *p* and *q*.[17] So if it's irrational to believe a Moorean conjunction, then it's irrational to believe the conjuncts of a Moorean conjunction. Believing a Moorean conjunction is more egregiously irrational than believing the conjuncts of a Moorean conjunction, but neither is fully rational. In much the same way, believing an explicit contradiction is more egregiously irrational than believing contradictory propositions, but neither is fully rational. Our rational failings are sometimes excusable given our psychological limitations, but we cannot avoid rational criticism just by failing to conjoin our beliefs. We need an epistemological solution to explain this.

Finally, Roy Sorensen notes that there are Moorean sentences that have neither omissive nor commissive forms, although they *entail* sentences with omissive or commissive forms. For instance, each of (5) and (6) entails (7):

(5) God knows that we are atheists. (1988: 17)

---

[15] See Davidson 1973 and Lewis 1974.
[16] See Kahneman 2011 for a survey of empirical work on human reasoning.
[17] Multi-premise closure is controversial because the probability of a conjunction can be less than each of its conjuncts when they're not certain. But we need only a weakened version that applies in cases where the probability of the conjunction does not fall below the minimum threshold for rational belief.

(6) The atheism of my mother's nieceless brother's only nephew angers God. (1988: 28)

(7) God exists, but I don't believe that God exists.

In some cases, the entailments are more complex than others. If one fails to recognize these entailments, then one can have beliefs that entail omissive Moorean conjunctions. Given our psychological limitations, these beliefs are more egregiously irrational when the entailments are more obvious and less so when they are more complicated. But again, while our psychological limitations can excuse our rational failings, they do not absolve us from rational criticism altogether.

An epistemological solution to Moore's paradox should explain why there is always some irrationality – that is, some failure of epistemic rationality – associated with believing anything that entails a Moorean conjunction. I'll argue that it's always irrational to believe Moorean conjunctions of the omissive form, although it's sometimes rational to believe Moorean conjunctions of the commissive form, but only when one has inconsistent beliefs. I'll begin with the omissive form and I'll revisit the commissive form in due course.

2.3. Moorean Belief is Self-Falsifying

John Williams (1994: 165) argues that it's always irrational to believe an omissive Moorean conjunction, *p and I don't believe that p*, because it is *self-falsifying* in the sense that believing the conjunction makes it false. If I believe the conjunction, then I believe both conjuncts, but believing the first conjunct makes the second conjunct false. As a result, the whole conjunction is false whenever I believe it.[18]

Claudio de Almeida (2001: 41) rejects Williams' proposal on the grounds that one can rationally believe necessary falsehoods. His example is believing the negation of the Lowenheim-Skolem theorem on the basis of misleading testimony from experts in logic. This example is controversial, since it turns on a disputed question about whether full rationality requires omniscience and infallibility about

[18] Here, and elsewhere, I'll assume the principle that belief distributes over conjunction, but not the principle that belief collects over conjunction.

a priori truths of logic. But we can give less controversial examples: for instance, it can be rational to believe the negations of Kripkean a posteriori necessary truths – Hesperus is distinct from Phosphorus, water is not composed of H_2O, and so on.

In defense of Williams, however, these propositions are not self-falsifying. They are necessarily false, and so false whenever believed, but believing them doesn't *make* them false, since they are false whether or not they are believed. So the challenge remains to give a more convincing counterexample to the thesis that all self-falsifying beliefs are irrational. As I'll explain, Moore himself provides the materials we need for constructing a counterexample.

Moore (1993: 208) uses the following pair of sentences to illustrate an epistemic asymmetry between first-person and third-person perspectives:

(8) I don't believe it's raining, but as a matter of fact it is.
(9) Moore doesn't believe it's raining, but as a matter of fact it is.

As Moore notes, it's absurd for him to assert (8), but it's not absurd for someone else to assert (9) in making reference to Moore. Similarly, it needn't be absurd for Moore himself to assert (9) in making reference to himself, so long as he is suffering from amnesia or otherwise rationally ignorant of his own identity. In that case, Moore can rationally believe (9), despite the fact that in believing it, he thereby makes it false. This shows that not all self-falsifying beliefs are irrational.[19]

Moorean belief is irrational not merely because it's self-falsifying but because I can *know* that it's self-falsifying. After all, I can't rationally believe what I know to be false. But knowing that a proposition is self-falsifying doesn't enable me to know that it's false unless I also know that I *believe* it. What makes believing an omissive Moorean conjunction irrational is the fact that I can know, or rationally believe, the premises of the following argument:

[19] Just as not all self-falsifying beliefs are irrational, so not all self-verifying beliefs are rational. See Pryor 2006 for examples and discussion.

(1) Anyone who believes that *p and I don't believe that p* thereby believes something false.

(2) I believe that *p and I don't believe that p*.

(3) Therefore, it is false that *p and I don't believe that p*.

So it seems that we can't explain why believing an omissive Moorean conjunction is irrational without assuming that we can know what we believe. This is our first hint that the connection between rationality and self-knowledge is the key to solving Moore's paradox.

2.4. The Knowledge Rule

Timothy Williamson argues that believing omissive Moorean conjunctions is irrational because they cannot be *known* (2000: 253-4). Knowing that *p* requires truly believing that *p*, but you can't truly believe an omissive Moorean conjunction, since believing the conjunction makes it false. It follows that you can't know an omissive Moorean conjunction.

How do we get from the premise that you can't know an omissive Moorean conjunction to the conclusion that you can't rationally believe it? Williamson bridges the gap in the argument by proposing the *knowledge rule* for belief:

> *The knowledge rule*: one should believe *p* only if one knows *p*. (2000: 255-6)

If we interpret the knowledge rule in terms of the 'should' of rationality, then it implies that one rationally believes that *p* only if one knows that *p*. Since I cannot know omissive Moorean conjunctions to be true, it follows that I cannot rationally believe them either.

Like many others, I reject the knowledge rule on the grounds that one can rationally believe that *p* without knowing that *p* in deception cases in which it's false that *p* or Gettier cases in which it's accidentally true that *p*. But if that is right, then why can't I rationally believe an omissive Moorean conjunction without knowing it?

There is an important difference between these cases. We can put the point in terms of Jonathan Sutton's (2007: 8-14) distinction between "known unknowns" and "unknown unknowns". It can be rational to believe that *p* in deception cases and Gettier cases so long as one cannot know that one cannot know that *p*. These are unknown unknowns. In contrast, one can know on the basis of Williamson's argument that one cannot know omissive Moorean conjunctions. These are known unknowns. So the question arises whether it can be rational to believe that *p* while knowing that one cannot know that *p*.[20]

Can it be rational to believe an omissive Moorean conjunction while knowing that I cannot know it? Since I can know that it's self-falsifying to believe an omissive Moorean conjunction, I can know that either it's false or I don't believe it. Now, if I can know which of these disjuncts is true, then I can argue as follows:

(1) Either I can know that it's false or I can know that I don't believe it.
(2) If I can know that it's false, then I can't rationally believe it, since I can't rationally believe what I can know to be false.
(3) If I can know that I don't believe it, then I can't rationally believe it, since I can't rationally believe what I don't believe at all.
(4) Either way, I can't rationally believe an omissive Moorean conjunction.

The problem is that knowledge of a disjunction doesn't entail knowledge of either disjunct. In the case at hand, I can't know which disjunct is true unless I know whether or not I believe the omissive Moorean conjunction. But if I can't know this, then the first premise is false: I can know that either the conjunction is false or I don't believe it, but I can't know which of these disjuncts is true. If I believe the

[20] Smithies (forthcoming) argues for an RK thesis, which states that it's rational for one to believe that *p* only if it's rational for one to believe that one is in an epistemic position to know that *p*. The RK thesis doesn't undercut the rationality of believing an omissive Moorean conjunction, since it can be rational to believe you're in an epistemic position to know an omissive Moorean conjunction even if you know that you cannot convert this epistemic position into knowledge because your evidence is finkish. See section 2.6 below for more on finkish evidence.

conjunction, then it's false, but if I can't *know* that I believe it, then I can't *know* that it is false, and so there's no obstacle to the rationality of believing it.

Once again, the key to solving Moore's paradox is the connection between rationality and self-knowledge: we need to assume that rationality requires knowing what you believe in order to explain the irrationality of believing omissive Moorean conjunctions. Sydney Shoemaker has done more than anyone to argue for this connection between rationality and self-knowledge, so I turn to his view next.[21]

2.5. The Rational Self-Intimation Thesis

Sydney Shoemaker (1996: 76) notes that one cannot self-consciously believe an omissive Moorean conjunction without thereby having contradictory beliefs.[22] We can define a *self-conscious* belief as a belief that one believes oneself to have; that is, one self-consciously believes that *p* just in case one believes that *p* while also believing that one believes that *p*. If I self-consciously believe an omissive Moorean conjunction of the form, *p and I don't believe that p*, then I believe that I believe that *p* while also believing that I don't believe that *p*. Assuming that rationality precludes having contradictory beliefs, it follows that there's always some degree of irrationality involved in self-consciously believing a Moorean conjunction.

The appeal to self-consciousness cannot explain the irrationality of all Moorean beliefs unless we assume that all beliefs are self-conscious. On this view, it's impossible to believe that *p* without also believing that one believes that *p*. This is captured by the following self-intimation thesis:

> *The self-intimation thesis*: necessarily, if one believes that *p*, then one believes that one believes that *p*.

---

[21] See Shoemaker 1996: Ch. 4 on Moore's paradox. Shoemaker gives additional arguments for the rational self-intimation thesis in his 1996: Ch. 2 & 11. These arguments are beyond the scope of this paper, but see Kind 2003, Siewert 2003, and Byrne 2005: 89-92 for critical discussion.

[22] See also Baldwin 1990: 230 and Kriegel 2004 for this argument.

However, the self-intimation thesis is false. First, non-human animals and human infants can have beliefs without possessing the concept of belief. And second, human adults who possess the concept of belief can have beliefs without having an infinite hierarchy of higher-order beliefs of infinitely increasing complexity. To avoid these problems, we need to modify the self-intimation thesis as follows:

> *The modified self-intimation thesis*: necessarily, if one believes that *p*, and one has some doxastic attitude towards the proposition that one believes that *p*, then one believes that one believes that *p*.

But even the modified self-intimation thesis is falsified by cases of compromised rationality – such as repression or self-deception – in which one believes that *p* while disbelieving or withholding belief that one believes that *p*.

If we now modify the self-intimation thesis by adding a rationality condition, as Shoemaker does, then we arrive at the rational self-intimation thesis below:

> *The rational self-intimation thesis*: necessarily, if one is rational, and one believes that *p*, and one has some doxastic attitude towards the proposition that one believes that *p*, then one believes that one believes that *p*.

On this view, it's always irrational to believe that *p* while disbelieving or withholding belief that one believes that *p*. In support of this claim, Shoemaker (1996: 78) notes that the following conversation seems absurd:

> A: Is it raining?
> B: Yes.
> A: Do you believe that it's raining?
> B: No. (Or: I don't know.)

Presumably, what explains the absurdity of this exchange is that believing that *p* rationally commits one to believing that one believes that *p*. As Shoemaker puts the

point, "if one believes something, and considers whether one does, one must, on pain of irrationality, believe that one believes it" (1996: 77).

According to Shoemaker, the rational commitment to believe self-consciously, together with the rational commitment to avoid contradictory beliefs, implies a rational commitment to avoid believing omissive Moorean conjunctions. If I believe that *p*, then I'm rationally committed to believing that I believe that *p*, and hence to refrain from believing that I don't believe that *p*. But if I believe an omissive Moorean conjunction, then I violate one or other of these rational commitments.

2.6. A Puzzle about Finkish Evidence

If the rational self-intimation thesis is true, then believing omissive Moorean conjunctions is always irrational. But this gives rise to a puzzle. Why does rationality require that if I believe that *p*, then I believe that I believe that *p*?

The puzzle arises from the fact that my total evidence can make it rational to believe *p* while also making it rational to believe I don't believe *p*. For instance, I might have meteorological evidence that it will rain, while also having psychological evidence that I don't believe it will rain. In that case, my total evidence makes it rational to believe the omissive Moorean conjunction, "It will rain, but I don't believe it will rain". But if my evidence makes it rational to believe the Moorean conjunction, then why does rationality require me to refrain from believing it?

Shoemaker (1996: 42-3) argues that one *cannot* have evidence for an omissive Moorean conjunction. Suppose my meteorological evidence that it will rain consists in the fact that the forecast says it will rain. Shoemaker argues that this is part of my evidence only if I believe it, in which case the fact that I believe it is part of my psychological evidence. In that case my psychological evidence makes it rational to believe that I believe it will rain, since I'm likely to believe it will rain when I believe the forecast says it will rain. In response to Shoemaker, however, I might have background evidence that I'm irrational and so unlikely to believe it will rain even if the forecast says it will rain. In that case, my total evidence makes it rational to believe the omissive Moorean conjunction. So the puzzle remains.

This puzzle relies on an assumption of *evidentialism*, defined as the thesis that one's evidence determines which propositions it is rational for one to believe. If we reject this assumption, then we can dissolve the puzzle by allowing for rational dilemmas in which one is rationally required to refrain from believing propositions that are supported by the evidence.[23] In my view, however, this is a last resort. Other things equal, we should prefer a simpler, more unified theory that explains the requirements of rationality in terms of facts about what the evidence supports. The challenge that remains is to defend evidentialism by solving the puzzle.

The key to solving the puzzle is to draw a distinction between propositional and doxastic senses of rationality.[24] Within the framework of evidentialism, this is the distinction between having evidence that makes it rational for one to believe a proposition and believing the proposition in a way that is properly based on the evidence. The puzzle arises because one can have evidence that makes it rational for one to believe an omissive Moorean conjunction, although one cannot believe an omissive Moorean conjunction in a way that is properly based on the evidence. But the apparent conflict can be resolved if we allow that the evidence in question is "finkish" in the sense that it is destroyed or undermined in the process of attempting to form a doxastically rational belief that is properly based on the evidence.[25]

The simple theory implies that evidence for an omissive Moorean conjunction is finkish in just this sense. After all, believing that *p* has the effect of destroying the evidence that makes it rational to believe that one does not believe that *p*. On the simple theory, psychological evidence about what one believes is constituted by the facts about what one believes. So one can have meteorological evidence that will rain, while also having psychological evidence that one doesn't believe that it will rain, so long as one doesn't believe that it will rain. But if one now comes to believe that it will rain on the basis of the meteorological evidence, then

---

[23] See Worsnip 2015 for a view of this kind.
[24] See Firth 1978. The distinction can be drawn either in terms of rationality or justification. I'll use these terms interchangeably.
[25] The allusion is to Martin's 1994 finkish dispositions, which are destroyed whenever their manifestation conditions obtain. I introduced the notion of finkish evidence in Smithies 2012b: 288.

one's psychological evidence changes: one loses one's earlier psychological evidence that one does not believe it will rain and acquires new psychological evidence that one believes it will rain. Therefore, one's evidence for an omissive Moorean conjunction is always finkish.[26]

On the simple theory, believing omissive Moorean conjunctions is sometimes propositionally rational, although it is never doxastically rational. One important consequence is that this undermines attempts to define the propositional sense of rationality (or justification) in terms of its doxastic cousin. For instance, John Turri proposes the following necessary condition for propositional justification:

> Necessarily, for all S, *p*, and *t*, if *p* is propositionally justified for S at *t*, then *p* is propositionally justified for S at *t* because S currently possesses at least one means of coming to believe *p* such that, were S to believe *p* in one of those ways, S's belief would thereby be doxastically justified. (2010: 320)

Omissive Moorean conjunctions provide a counterexample. As we have seen, one can have evidence that propositionally justifies believing an omissive Moorean conjunction. However, it's not true that one has the means for coming to believe an omissive Moorean conjunction in a way that is doxastically justified.

How should we understand the connection between propositional and doxastic justification within an evidentialist framework? A standard view is that doxastic justification is propositional justification plus proper basing:

> Necessarily, for all S, *p*, and *t*, S's belief that *p* is doxastically justified at *t* if and only if at *t*, S has some evidence *e* that makes it the case that *p* is propositionally justified for S, and S believes that *p* in a way that is properly based on evidence *e*.

---

[26] De Almeida notes that believing an omissive Moorean conjunction is "epistemically self-defeating" in the sense that "belief in the conjunction necessarily furnishes me with a reason to disbelieve the right-hand-side of the conjunction" (2001: 51-2). But he goes too far in claiming that "a Moore-paradoxical proposition is one for which the believer can have *no non-over-ridden evidence*" (2001: 44).

There is no commitment here to the claim that doxastic justification can be reductively defined in terms of propositional justification plus proper basing. In fact, I suspect that no such reductive definition is immune from counterexamples.[27] Instead, we can define proper basing non-reductively as the relation that converts propositional justification into doxastic justification. Since evidence for an omissive Moorean conjunction is always finkish, one cannot satisfy the proper basing relation that converts propositional justification into doxastic justification. However, it does not follow – as Turri's proposal implies – that it cannot be propositionally rational to believe an omissive Moorean conjunction.

2.7. The Rational Infallibility Thesis

It is sometimes assumed that a solution to Moore's paradox should give a unified treatment of omissive and commissive forms.[28] As I'll explain, however, there are logical differences between them that call for differential treatment. I've argued that believing omissive Moorean conjunctions is always irrational. In contrast, I'll argue that believing commissive Moorean conjunctions is sometimes rational, but only when one has contradictory beliefs. Assuming that rationality precludes having contradictory beliefs, it follows that believing Moorean conjunctions is always either irrational or involves some associated irrationality.

Whereas omissive Moorean conjunctions are self-falsifying, commissive Moorean conjunctions are not. They can be truly believed when, and only when, one has contradictory beliefs. Indeed, one can *know* that a commissive Moorean conjunction is true when one has contradictory beliefs. Suppose I know on the basis of good evidence that I'm a bad driver: my past is littered with wreckage. At the same time, I know that I can't shake the belief that I'm a good driver. In that case, I know a Moorean conjunction of the commissive form, "I believe I'm a good driver,

---

[27] Turri 2010 gives counterexamples to the analysis of doxastic justification as propositional justification plus basing, but his examples don't involve proper basing. I have more to say about the proper basing relation in Smithies 2015b.
[28] See, for example, Green and Williams 2007.

but I'm not". Since rational belief is required for knowledge, it follows that I can rationally believe this Moorean conjunction too.[29]

It might be objected that one cannot rationally believe a proposition while also believing its negation. On this view, rational belief requires a kind of internal coherence that precludes conflicting beliefs. But this requirement is too demanding: if I have inconsistent beliefs, then my total set of beliefs is irrational to some degree, but it doesn't follow that the irrationality can be traced to every member of the set. In particular, the rationality of my belief that I'm a bad driver need not be impugned by the recalcitrance of my irrational belief that I'm a good driver.[30]

If I can know a commissive Moorean conjunction, then I can rationally believe it when it's true. But can I rationally believe it when it's false? If so, then believing a commissive Moorean conjunction just requires *believing* that one has contradictory beliefs. The problem is that while it's irrational to *have* contradictory beliefs, it's not obviously irrational to *believe* that one has contradictory beliefs. So how can we explain the sense that believing a Moorean conjunction always involves some degree of associated irrationality?

The rational self-intimation thesis is no use here. If I self-consciously believe an omissive Moorean conjunction, then I *have* contradictory beliefs: I believe that I believe that *p* while also believing that I don't believe that *p*. If I self-consciously believe a commissive Moorean conjunction, in contrast, then I merely *believe* that I have contradictory beliefs: I believe that I believe that *p* while also believing that I believe that not-*p*. But this fails to identify any irrationality associated with believing commissive Moorean conjunctions.

In order to explain this associated irrationality, we need to combine the rational self-intimation thesis with the following rational infallibility thesis:

---

[29] Thanks to Alex Byrne for persuading me of this. See Shoemaker 1996: 89-90, Moran 2001: 85, de Almeida 2001: 43, Gertler 2010: 139-41, and Coliva 2015: 178-9 for similar examples. For some dissent, see Heal 1994: 11.

[30] Compare Arpaly 2003 on inverse akrasia: the rationality of helping Jim need not be impugned by the irrationality of my belief that I shouldn't help Jim; similarly, the rationality of my believing that *p* need not be impugned by the irrationality of my belief that I shouldn't believe that *p*.

> *The rational infallibility thesis*: necessarily, if one is rational, and one believes that one believes that *p*, then one believes that *p*.[31]

The rational infallibility thesis ensures that if one rationally believes a commissive Moorean conjunction, then one has contradictory beliefs. Assuming that rationality precludes having contradictory beliefs, it follows that believing Moorean conjunctions always involves some associated irrationality. But it doesn't follow that my Moorean belief is itself irrational. It may instead be a rational response to irrationality that lies elsewhere in my belief system.

2.8. A Simple Solution

If we combine the rational self-intimation thesis with the rational infallibility thesis, then we arrive at the following:

> *The rational biconditional thesis*: necessarily, if one is rational, and one has some doxastic attitude towards the proposition that one believes that *p*, then one believes that *p* if and only if one believes that one believes that *p*.

The rational biconditional thesis explains why there is always some degree of irrationality associated with believing Moorean conjunctions – or the conjuncts of Moorean conjunctions – of either omissive or commissive forms.

Moreover, the simple theory explains why this biconditional thesis is true. The rational self-intimation thesis is true because the fact that one believes that *p* makes it rational to believe that one believes that *p*. The rational infallibility thesis is true because the fact that one doesn't believe that *p* makes it rational to believe that one doesn't believe that *p*. The simple theory implies that one cannot rationally believe that one doesn't believe that *p* when one does believe that *p*, and one cannot rationally believe that one believes that *p* when one doesn't believe that *p*. On the

[31] Heal 1994: 22-3 endorses a psychological version of the infallibility thesis on which believing that one believes that *p* entails believing that *p*.

simple theory, one has reasons that make it rational to believe that one believes that *p* if and only if one believes that *p*.

The result is an extended argument by inference to the best explanation for extending the simple theory of introspection from conscious experience to belief. One might challenge this argument either by disputing the Moorean data to be explained or by proposing a rival explanation of the Moorean data. In section 3, I'll consider rival explanations that appeal to the transparency of belief, but I'll argue that the simple theory of introspection provides a better account of the sense in which belief is transparent. Finally, in section 4, I'll defend my solution to Moore's paradox against the objection that that there is no irrationality involved in believing Moorean conjunctions when one has misleading evidence about what one believes.

3. The Puzzle of Transparency

It is often said that belief is *transparent* in the sense that I can rationally answer the question whether I believe that *p* by answering the question whether *p*. This idea is encapsulated in the following passage from Gareth Evans:

> 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*. . . . We can encapsulate this procedure for answering questions about what one believes in the following simple rule: whenever you are in a position to assert that *p*, you are *ipso facto* in a position to assert 'I believe that *p*'. (1982: 225-6)

Several philosophers have argued that this claim that belief is transparent provides the key to solving Moore's paradox. In this section, I'll criticize some of these rival

solutions to Moore's paradox and I'll argue that the simple theory of introspection gives a better account of the sense in which belief is transparent.[32]

3.1. Evans' Principle

Drawing on this passage from Gareth Evans, John Williams (2004) proposes the following principle:

> *Evans' Principle*: Whatever justifies me in believing that *p* also justifies me in believing that I believe that *p*. (2004: 348)

Williams argues that Evans' Principle explains why I cannot have justification to believe a Moorean conjunction of the omissive form, '*p* and I don't believe that *p*'. Williams assumes, quite plausibly, that I have justification to believe the conjunction only if I have justification to believe each of its conjuncts. But if I have justification to believe the first conjunct, then it follows by Evans' Principle that I have justification to disbelieve the second conjunct. I cannot also have justification to believe the second conjunct unless 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).[33]

The main problem with this solution to Moore's paradox is that Evans' Principle is false: it is refuted by counterexamples in which one's total evidence justifies believing that *p* without also thereby justifying the belief that one believes that *p*. As we have seen, one's total evidence can justify believing that it will rain without thereby justifying the belief that one *believes* that it will rain. Indeed, one's total evidence can justify believing that it will rain while also justifying the belief

---

[32] On the connection between transparency and Moore's paradox, see Gallois 1996, Moran 2001 and 2003, Williams 2004, Byrne 2005 and 2011, Fernandez 2005 and 2013, Velleman and Shah 2005, and Silins 2012.

[33] To explain the commissive form, Williams proposes a variant of Evans's principle: "Whatever justifies me in believing that *p* also justifies me in believing that I do not believe that *not-p*" (2004: 352).

that one *doesn't believe* that it will rain. In that case, one's total body of evidence justifies believing an omissive Moorean conjunction.

Jordi Fernandez (2013) advances a related proposal, which he calls the Bypass View:

> *The Bypass View*: Normally, if S believes that she believes that *p*, then there is a state E such that (a) S's (higher-order) belief has been formed on the basis of E [and] (b) E constitutes grounds for the belief that *p* in S. (2013: 49)

On this view, the same evidential state E can do "double duty" in justifying not only the belief that *p*, but also the belief that one believes that *p*. According to Fernandez, this happens when E reliably indicates that *p*, while also reliably indicating that one believes that *p*. However, the Bypass View avoids the objection to Evans' Principle because these reliable indications can be dissociated. If E reliably indicates that *p*, but does not reliably indicate that one believes that *p*, then one's evidence can justify believing that *p* without also justifying the belief that one believes that *p*.[34]

Although the Bypass View avoids the objection to Evans' Principle, this comes at the cost of a fully general solution to Moore's paradox.[35] If one's evidence does double duty in justifying the belief that *p* while also justifying the belief that one believes that *p*, then it cannot justify believing the omissive Moorean conjunction that *p and I don't believe that p*. But the problem is that one's evidence doesn't *always* do this kind of double duty even if it *normally* does. If one's evidence reliably indicates that *p*, while also reliably indicating that one does not believe that *p*, then it justifies believing the omissive Moorean conjunction that *p and I don't believe that p*. The Bypass View can't explain why believing an omissive Moorean conjunction is irrational in such cases.

---

[34] See Fernandez 2013: 63-6 in reply to Zimmerman 2004.

[35] See Fernandez 2013: 126-38 for his proposed solution to Moore's paradox and a comparison with Evans' principle.

3.2. Rational Entitlement

Richard Moran (2001) claims that it is a normative ideal of rationality that one's beliefs are transparent in the sense that one can answer the question whether one believes that *p* by answering the question whether *p*. At the same time, he notes that there are failures of transparency in which one knows one has recalcitrant beliefs that are not supported by the evidence. For instance, he writes, "I can well imagine the accumulated evidence suggesting both that I believe that it's raining, and that it is not in fact raining" (2001: 84). At the same time, he notes that this possibility "clashes with the conception of oneself as a rational agent" (2001: 84). The claim is that transparency is a rational ideal that we don't always satisfy.

The challenge here is to explain why transparency is an ideal of rationality. The *puzzle of transparency* (as it has become known) is to explain why it is rational to answer the question whether one believes that *p* by answering the question whether *p*. This puzzle is pressing because, as we have seen, the evidence that justifies believing that *p* need not thereby justify believing that one believes that *p*. In the following passage, Moran articulates this puzzle and his solution:

> What right have I to think that my reflection on the reasons in favor of P (which is one subject-matter) has anything to do with the question of what my actual *belief* about P is (which is quite a different subject-matter)? . . . I *would* have a right to assume that my reflection on the reasons in favor of rain provided an answer to the question of what my belief is, if I could assume that *what* my belief here is was something determined by the conclusion of my reflection on those reasons. (2003: 405)

Moran claims that it's rational to answer the question whether one believes that *p* by answering the question whether *p* when, and only when, it's rational to assume that one's beliefs about whether *p* are settled by one's reflection on the reasons that bear on the rationality of believing that *p*. In other words, the rationality of making the transition relies on the rationality of assuming that one's beliefs are settled by one's reflection on the evidence.

Quassim Cassam (2014: Ch. 9) raises several problems for Moran's proposal. First, the Generality Problem: it cannot be extended to explain self-knowledge of mental states, such as conscious sensations, which are not responsive to one's reflection on reasons. Second, the Immediacy Problem: it cannot explain why knowledge of one's beliefs should count as immediate or non-inferential, rather than inferentially mediated, insofar as it depends on the rationality of assuming that one's beliefs are settled by one's reflection on evidence. And third, the Matching Problem: it cannot explain our introspective knowledge of what we believe when we know that our beliefs are not settled by our reflection on the evidence.

The third problem is particularly damaging to the prospects for solving Moore's paradox. Suppose I know my beliefs about my own abilities tend to be *recalcitrant* in the sense that they are impervious to the conclusions that I draw on the basis of my reflections on the evidence. I conclude that the evidence justifies believing that I'm a bad driver, but I suspect rationally – though, as it happens, mistakenly – that this fails to have any impact on what I believe. In that case, it is not rational, all things considered, to assume that my belief is settled by my reflection on the evidence. On Moran's account, then, it is not rational for me to answer the question whether I believe that I'm a bad driver by answering the question whether I'm a bad driver. But then what explains the irrationality of believing the omissive Moorean conjunction, "I'm a bad driver, but I don't believe I'm a bad driver"?

3.3. Self-Verifying Rules

Alex Byrne (2005, 2011) claims that I can know what I believe by making "transparent inferences" from premises about the world to conclusions about my own beliefs. For instance, I can know that I believe there will be a third world war by inferring this conclusion from the premise that there will be a third world war. More generally, I can know what I believe by making inferences in accordance with what Andre Gallois (1996: 46-7) calls the doxastic schema:

The doxastic schema: *p*, therefore I believe that *p*.

As Gallois notes, the doxastic schema is neither deductively nor inductively valid: it "does not fit any standard pattern of good inference" (1996: 47). So the puzzle of transparency, within this framework, is to explain why reasoning in accord with the doxastic schema is capable of yielding knowledge and justified belief.

Byrne's solution is that the doxastic schema is *strongly self-verifying* in the sense that reasoning in accord with it is guaranteed to yield true beliefs. He writes, "If one reasons in accord with the doxastic schema, and infers that one believes that *p* from the premise that *p*, then one's second-order belief is *true*, because inference from a premiss entails belief in that premiss" (2011: 206).

I'll consider two objections to Byrne's solution.[36] The first objection is that my knowledge of what I believe is not based on inference from what I believe, since I can know what I believe even when my belief is false. The key assumption here is that I can't acquire knowledge by inference from a false premise. Gilbert Harman (1973: 47) justifies this assumption by its role in explaining Gettier's counterexamples to the analysis of knowledge as justified true belief. In each of Gettier's original examples, the subject lacks knowledge because his justified true belief is inferred from a false premise.

Byrne (2011: 206-7) replies that one can acquire knowledge by inference from a false premise so long as one's belief is *safe from error* in the sense that it could not easily have been false given the way it is formed.[37] We can explain why subjects lack knowledge in Gettier cases because their beliefs are not safe from error. In contrast, reasoning in accord with the doxastic schema yields beliefs that are safe from error even when they are inferred from a false premise.

The second objection is that I can know what I believe when my beliefs are not only false, but also unjustified. The key assumption here is that I cannot acquire knowledge by inference from a premise that is not only false, but also unjustified.

---

[36] Both objections are raised by Shoemaker 2009: 36 and Silins 2012: 304-5.
[37] See Warfield 2005 for alleged cases of knowledge by inference from false premises. Arguably, however, in each of his examples, there is some true premise that is dispositionally believed and that is causally relevant in explaining how the subject acquires inferential knowledge.

This is because justified belief is required for knowledge, but a belief cannot be justified by inference from unjustified premises.

Byrne's appeal to safety is no use here, since safety from error is not sufficient for justified belief. Suppose I make inferences in accordance with the following schema: if *x* is water, then *x* is composed of H_2O molecules. My conclusions are no less safe than my premises, but if I know nothing about chemistry, then my conclusions are not justified even when my premises are. We can make the same point using standard counterexamples to externalist theories of justification: for instance, BonJour's (1985: Ch. 3) clairvoyant, Norman, is not justified in believing that the President is in New York even if his belief is formed on the basis of a reliable clairvoyant mechanism and is therefore safe from error. If safety is not sufficient for justified belief, then we need an alternative to Byrne's inferential account of how we can know what we believe.

3.4. A Simple Solution

All of these proposals deserve more extended discussion than I can give them here. Even so, the problems I have raised are serious enough to motivate the search for an alternative solution to the puzzle of transparency. I'll now argue that the simple theory solves the puzzle of transparency in a way that avoids these problems.

A solution to the puzzle of transparency must explain how it can be rational to answer the question whether one believes that *p* by answering the question whether *p*. Unlike Byrne's account, the simple theory explains why the transition from believing that *p* to believing that one believes that *p* is not merely reliable or safe from error, but also rational or justified. The transition is rational because the fact that one believes that *p* is an introspective reason to believe that one believes that *p*. Moreover, this introspective reason puts one in a position to know by means of introspection that one believes that *p*.

Unlike Moran's account, the simple theory does not rely on the assumption that we are rationally entitled to assume that our beliefs are settled by our reflection on the evidence. The simple theory can explain how we know what we believe even when we know that our beliefs are unresponsive to reflection. On the simple theory,

rationality requires self-knowledge, but self-knowledge doesn't require that one is rational, or that one is rationally entitled to assume so.

The simple theory also has the advantage of generality: it explains our introspective knowledge of what we believe on a more general model that applies equally to our introspective knowledge of conscious experience. On the simple theory, the transition from believing that *p* to believing that one believes that *p* is justified in much the same way as the transition from feeling pain to believing that one feels pain. On Moran's account, in contrast, our introspective knowledge of rational attitudes cannot be assimilated to the same model as our introspective knowledge of conscious sensations.[38]

Moreover, the simple theory explains why the transition from believing that *p* to believing that one believes that *p* is non-inferential. It's not justified in the same way as an inference where the premise deductively entails or inductively raises the probability of the conclusion. The justification for the transition cannot be explained on the model of deductive or inductive inference, but is rather an instance of a more general pattern of non-inferentially justified transitions from mental states to beliefs about those mental states.

The simple theory, like Evans' principle, explains why rationality requires that one believes that *p* if and only if one believes that one believes that *p*. However, the form of the explanation is quite different. According to the simple theory, one's introspective reason to believe that one believes that *p* has its source in one's belief that *p*, rather than in whatever justifies one's belief that *p*. This has at least three advantages. First, it means that the simple theory can be extended to account for our introspective knowledge of unjustified as well as justified beliefs. Second, it avoids counterexamples in which one's total evidence justifies believing that *p* without thereby justifying the belief that one believes that *p*. And third, it explains how there can be finkish evidence for believing an omissive Moorean conjunction.

---

[38] See Moran 2001: xxxiii. Compare Boyle 2009, who argues that there are two fundamentally different kinds of self-knowledge: an active kind through which we know our own beliefs and judgments, and a passive kind through which we know our conscious sensations.

Finally, the simple theory explains the sense in which introspection involves looking outwards towards the world, rather than looking inwards towards one's own mind. As Evans writes, "in making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward – upon the world" (1982: 225). On the simple theory, one's belief that one believes that *p* is rationally based directly on the belief that *p*, rather than any introspective representation of one's belief that *p*. The transition from believing that *p* to believing that one believes that *p* is world-directed insofar as the belief that *p* is itself world-directed. More generally, the transition from being in a mental state to believing that one is in that mental state is world-directed insofar as the mental state in question is world-directed. There is no further explanatory work to be done by an appeal to transparency that cannot be accommodated within the simple theory of introspection.

4. Ideal and Non-Ideal Rationality

In this section, I'll respond to an objection to the simple theory. The objection is that I can have misleading evidence about what I believe, just as I can have misleading evidence about what others believe. Moreover, misleading evidence can rationally support false beliefs about what I believe, just as it can rationally support false beliefs about what others believe. Since the simple theory rules out the possibility of rational false beliefs about what you believe, it thereby seems to ignore the rational force of misleading evidence about what you believe.

There are two kinds of cases to consider. In cases of the first kind, I believe that *p*, but I have misleading evidence that I don't believe that *p*. Many cases of self-deception fit this description. So, for example, I might know "deep down" that I'm an addict although I won't admit this to myself or anyone else. Suppose my knowledge is so deeply repressed that the relevant belief is blocked from playing its normal functional role in action and reasoning. In that case, it might seem rational for me to believe that I don't believe that I'm an addict when in fact I do believe this.

In cases of the second kind, I don't believe that *p*, but I have misleading evidence that I do believe that *p*. Shoemaker (1996: 90) gives an example in which a normally reliable psychiatrist mixes up her files and mistakenly informs me that I

have a repressed belief that I was adopted. At the same time, I know full well that I wasn't adopted. In that case, it might seem rational for me to believe that I believe I was adopted when in fact I don't believe this at all.

These cases pull in two different directions. On the one hand, it seems irrational to ignore misleading evidence about what you believe. On the other hand, taking this kind of evidence into consideration can lead one into a Moorean predicament that seems irrational. For instance, in the first case, I might believe the conjuncts of an omissive Moorean conjunction, "I'm an addict, but I don't believe I'm an addict." Similarly, in the second case, I might falsely believe a commissive Moorean conjunction, "I believe that I was adopted, but I wasn't adopted".

One reaction is to deny that there is always some degree of irrationality associated with believing propositions that entail Moorean conjunctions. But this conflicts with the intuitive reaction that generates Moore's paradox in the first place. At a minimum, there remains a challenge to explain when believing Moorean conjunctions involves some associated irrationality and when it doesn't. I remain skeptical that a principled and well motivated account can be given.

My own reaction is to maintain that there is always some irrationality associated with believing Moorean conjunctions, while explaining away the rational force of misleading evidence in terms of a distinction between ideal and non-ideal standards of rationality. The basis of the distinction is that non-ideal standards take into consideration one's human limitations, whereas ideal standards abstract away from them. As a result, these standards can conflict: Moorean incoherence can be prohibited by ideal standards of rationality even if it is sometimes permitted or even required by non-ideal standards of rationality.

The distinction between ideal and non-ideal standards of rationality is most familiar from discussion of the thesis that rationality requires logical omniscience. Our best formal theories of rationality imply that rational agents are logically omniscient and infallible in the sense that they are certain of all logical truths. On this view, rationality is inconsistent with uncertainty or error about logic. Plausibly, however, there can be misleading evidence that makes it rational to be uncertain or mistaken about logic, such as expert testimony or evidence that one has taken

reason-distorting drugs. The thesis that rationality requires logical omniscience therefore seems to ignore the rational force of this misleading evidence.[39]

We can use the distinction between ideal and non-ideal standards of rationality to defend the thesis that rationality requires logical omniscience. Ideal standards require that one is perfectly responsive to the logical facts, and hence that one is never mistaken or uncertain about logic. But since non-ideal agents cannot satisfy these ideal standards, we can evaluate them by non-ideal standards of rationality that take their limited capacities into consideration. These non-ideal standards sometimes require non-ideal agents to depart from ideal standards by being uncertain or mistaken about logic.

We can use the same distinction to defend the thesis that rationality requires *doxastic omniscience*. Ideal standards require that one is perfectly responsive to the psychological facts, and hence that one is never mistaken or uncertain about what one believes. But since non-ideal agents cannot satisfy these ideal standards, we can evaluate them by non-ideal standards of rationality that take their limited capacities into consideration. These non-ideal standards sometimes require non-ideal agents to depart from ideal standards by being uncertain or mistaken about their beliefs.

Ideal agents are perfectly responsive to the evidence that is constituted by logical and psychological facts, but we non-ideal agents are not. Given our non-ideal predicament, it is not advisable for us to try to imitate ideal agents. Sometimes it is more advisable to do what we know ideal agents would never do. In particular, sometimes the best strategy for us to adopt is to form our beliefs in response to empirical proxies that indicate the logical or psychological facts with some degree of reliability. When these proxies are reliable, or rationally believed to be reliable, responding to them can serve as an imperfect and indirect way of responding to the logical or psychological facts that constitute our evidence.

So the claim is that doxastic omniscience is a rational ideal in much the same way as logical omniscience. We arrive at this conclusion by treating Moorean incoherence on a par with logical incoherence. Moorean incoherence, like logical

[39] See Christensen 2007 for this challenge and Smithies 2015b for my response.

incoherence, is never rational by ideal standards, but it can be rational by non-ideal standards that take our human limitations into account.

I'll close with two big picture questions that deserve more extended discussion elsewhere:

(1) If rationality does not require knowing about the external world, then why does it require knowing about the internal world?
(2) Does rationality require knowing about all of our internal states, or just some of them? If some, but not all, then how can we demarcate the boundary?

My answers to both questions draw upon a background theoretical commitment to access internalism. According to access internalism, rationality requires knowing about the internal states that determine what rationality requires of you. It doesn't require knowing about the external world. And it doesn't require knowing about internal states that play no role in determining what rationality requires of you. Rationality requires knowing about your beliefs and conscious experiences because of their role in determining what rationality requires you to believe and do.[40]

[40] I am grateful to audiences at the University of Oxford in June 2013, the University of Syracuse in August 2013, the New York Institute of Philosophy in November 2013, the University of Bergen in May 2015, and the University of Oslo in July 2015. Many thanks especially to David Barnett, Matthew Benton, Alex Byrne, Paul Egre, Ole Koksvik, Jack Lyons, Ram Neta, Matthew Parrott, Christopher Peacocke, Nicholas Silins, and Sydney Shoemaker for helpful comments and discussion.

References

Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. New York: Oxford University Press.

Baldwin, T. 1990. *G. E. Moore*. London: Routledge.

Boghossian, P. 1989. Content and Self-Knowledge. *Philosophical Topics* 17: 5-26.

BonJour, L. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Boyle, M. 2009. Two Kinds of Self-Knowledge. *Philosophy and Phenomenological Research* 78.1: 133-64.

Byrne, A. 2005. Introspection. *Philosophical Topics* 33: 79-104.

Byrne, A. 2011. Transparency, Belief, Intention. *Proceedings of the Aristotelian Society Supplementary Volume* 85: 201-21.

Cassam, Q. 2014. *Self-Knowledge for Humans*. Oxford: Oxford University Press.

Christensen, D. 2007. Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals*. Oxford Studies in Epistemology* 2: 3-31.

Cohen, S. 1984. Justification and Truth. *Philosophical Studies* 46.3: 279-95.

Coliva, A. 2015. How to Commit Moore's Paradox. *The Journal of Philosophy* 112.4: 169-92.

Davidson, D. 1973. Radical Interpretation. *Dialectica* 27.1: 314-28.

De Almeida, C. 2001. What Moore's Paradox is About. *Philosophy and Phenomenological Research* 62.1: 33-58.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Fernandez, J. 2005. Self-Knowledge, Rationality and Moore's Paradox. *Philosophy and Phenomenological Research* 71.3: 533-56.

Fernandez, J. 2013. *Transparent Minds: A Study of Self-Knowledge*. Oxford: Oxford University Press.

Firth, R. 1978. Are Epistemic Concepts Reducible to Ethical Concepts? In *Values and Morals*, edited by A. Goldman and J. Kim. Dordrecht: Kluwer.

Gallois, A. 1996. *The World Without, the Mind Within: An Essay on First-Person Authority*. Cambridge: Cambridge University Press.

Gertler, B. 2011. Self-Knowledge and the Transparency of Belief. In *Self-Knowledge*, ed. A. Hatzimoysis. Oxford: Oxford University Press.

Green, M. & Williams, J. 2007. Introduction. In *Moore's Paradox: New Essays on Belief, Rationality, and the First Person*. Oxford: Oxford University Press.

Harman, G. 1973. *Thought*. Princeton, NJ: Princeton University Press.

Heal, J. 1994. Moore's Paradox: A Wittgensteinian Analysis. *Mind* 103: 5-24.

Hintikka, J. 1962. *Knowledge and Belief*. Ithaca, NY: Cornell University Press.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York, NY: Farrar, Straus, and Giroux.

Kind, A. 2003. Shoemaker, Self-Blindness, and Moore's Paradox. *Philosophical Quarterly* 53: 39-48.

Koethe, J. 1978. A Note on Moore's Paradox. *Philosophical Studies* 34.3: 303-10.

Kriegel, U. 2004. Moore's Paradox and the Structure of Conscious Belief. *Erkenntnis* 61: 99-121.

Lewis, D. 1974. Radical Interpretation. *Synthese* 27: 331-44.

Marcel, A. 1983. Conscious and Unconscious Perception: Experiments on Visual Masking and Word Recognition. *Cognitive Psychology* 15: 197-237.

Martin, C. B. 1994. Dispositions and Conditionals. *Philosophical Quarterly* 44: 1-8.

Moore, G. E. 1942. A Reply to My Critics. In *The Philosophy of G. E. Moore*, ed. P. A. Schlipp. La Salle, IL: Open Court.

Moore, G. E. 1944. Russell's Theory of Descriptions. In *The Philosophy of Bertrand Russell*, ed. P. A. Schlipp. La Salle, IL: Open Court.

Moore, G. E. 1993. Moore's Paradox. In *G.E. Moore: Selected Writings*, ed. T. Baldwin. London: Routledge, 207-12.

Moran, R. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Cambridge, MA: Harvard University Press.

Moran, R. 2003. Responses to O'Brien and Shoemaker. *European Journal of Philosophy* 11.3: 402-19.

Neta, R. 2011. The Nature and Reach of Privileged Access. In *Self-Knowledge*, ed. A. Hatzimoysis. Oxford University 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.

Pryor, J. 2005. There is Immediate Justification. In *Contemporary Debates in Epistemology*, ed. M. Steup and E. Sosa. Oxford: Blackwell.

Pryor, J. 2006. Hyper-Reliability and Apriority. *Proceedings of the Aristotelian Society* 106.3: 327-44.

Reingold, E. and P. Merikle. 1990. On the Inter-Relatedness of Theory and Measurement in the Study of Unconscious Processes. *Mind and Language* 5: 9-28.

Rosenthal, D. 1995. Self-Knowledge and Moore's Paradox. *Philosophical Studies* 76: 196-209.

Ryle, G. 1949. *The Concept of Mind*. Chicago, IL: University of Chicago Press.

Shoemaker, S. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Shoemaker, S. 2009. Self-Intimation and Second Order Belief. *Erkenntnis* 71.1: 35-51.

Siewert, C. 2003. Self-Knowledge and Rationality: Shoemaker on Self-Blindness. In *Privileged Access*, ed. B. Gertler. Farnham: Ashgate.

Silins, N. 2012. Judgment as a Guide to Belief. In *Introspection and Consciousness*, eds. D. Smithies and D. Stoljar. New York, NY: Oxford University Press.

Silins, N. Forthcoming. The New Evil Demon Inside. In *The New Evil Demon Problem*, eds. J. Dutant and F. Dorsch. Oxford: Oxford University Press.

Smithies, D. 2012a. A Simple Theory of Introspection. In *Introspection and Consciousness*, eds. D. Smithies and D. Stoljar. New York, NY: Oxford University Press.

Smithies, D. 2012b. Moore's Paradox and the Accessibility of Justification. *Philosophy and Phenomenological Research* 85.2: 273-300.

Smithies, D. 2015a. Why Justification Matters. In *Epistemic Evaluation: Point and Purpose in Epistemology*, eds. J. Greco and D. Henderson. New York, NY: Oxford University Press.

Smithies, D. 2015b. Ideal Rationality and Logical Omniscience. *Synthese* 192.9: 2769-93.

Smithies, D. Forthcoming. The Irrationality of Epistemic Akrasia.

Sorensen, R. 1988. *Blindspots*. Oxford: Clarendon Press.

Sosa, E. 1991. *Knowledge in Perspective*. Cambridge: Cambridge University Press.

Sutton, J. 2007. *Without Justification*. Cambridge, MA: MIT Press.

Turri, J. 2010. On the Relationship Between Propositional and Doxastic Justification. *Philosophy and Phenomenological Research* 80.2: 312-26.

Velleman, D. and N. Shah. 2005. Doxastic Deliberation. *Philosophical Review* 114.4: 497-534.

Warfield, T. 2005. Knowledge from Falsehood. *Philosophical Perspectives* 19: 405-16.

Williams, J. 1994. Moorean Absurdity and the Intentional 'Structure' of Assertion. *Analysis* 54.3: 160-6.

Williams, J. 2004. Moore's Paradoxes, Evans's Principle and Self-Knowledge. *Analysis* 64: 348-53.

Williamson, T. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Wittgenstein, L. 1980. *Remarks on the Philosophy of Psychology*. Chicago, IL: University of Chicago Press.

Worsnip, A. 2015. The Conflict of Evidence and Coherence. *Philosophy and Phenomenological Research*.

Zimmerman, A. 2004. Unnatural Access. *Philosophical Quarterly* 54.2: 435-38.

Zimmerman, A. 2006. Basic Self-Knowledge: Answering Peacocke's Criticisms of Constitutivism. *Philosophical Studies* 128: 337-379.