David James Barnett

Self-Knowledge, Rationality, and Moore's Paradox

Abstract. Is self-knowledge a requirement of rationality, like consistency, or means-ends coherence? Many claim so, citing the evident impropriety of asserting, and the alleged irrationality of believing, Moore-paradoxical propositions of the form <p, but I don't believe that p>. If there were nothing irrational about failing to know one's own beliefs, they claim, then there would be nothing irrational about Moore-paradoxical assertions or beliefs. This paper responds to these claims by considering a few ways that the data surrounding Moore's paradox might be taken to support rational requirements to know one's own beliefs, and finds that none of these succeed in motivating self-knowledge requirements.

1. Introduction

To hear philosophers tell it, rationality requires a lot of us. It requires us to have logically consistent beliefs,[1] or (perhaps better) probabilistically coherent credences.[2] It requires us to believe the obvious deductive consequences of our existing beliefs, at least if we entertain them.[3] It requires us to intend to do the things that we believe are necessary means for achieving our ends.[4] It requires us not to believe things that we believe we shouldn't believe,[5] and not to intend things that we believe we shouldn't do.[6] It requires us to have preferences that make us representable as having a utility function, and to act in ways that maximize expected utility.[7] And so on. Now I'm not sure myself about everything on this list, and perhaps you're not either. But to a first pass, it illustrates the kind of things that rationality is supposed to require of us.

Does rationality require us to know our own minds? Or if not knowledge, does rationality at least require something in the ballpark, like accurate belief or credence? Many philosophers have thought so, and have taken Moore's paradox to support their thinking. This paper will opposes these common views.

2. Formulating Self-Knowledge Requirements

Because of its especially direct connection to Moore's paradox, my focus here will be on knowledge of one's *doxastic states*. Perhaps anyone who says that rationality requires knowledge of one's doxastic states should be prepared to say that it also requires knowledge of other types of mental states, on pain of arbitrariness. Extending self-knowledge

[1] Cf. Broome 2013, 9.2.

[2] Christensen 2004.

[3] Cf. Broome 2013, 9.3.

[4] Cf. Broome 2013, 9.4 and Kolodny 2005.

[5] Greco 2014, Horowitz 2014, and Smithies MS.

[6] E.g., Broome 2013, 9.5; Kolodny 2005; and Wedgewood 2007, Ch 1.

[7] See Buchak forthcoming for review.

requirements to other mental states might raise additional difficulties for my opponents, but I will set them aside. So, the principal claim I oppose is:

> (SELF-KNOWLEDGE) Rationality requires that one have accurate doxastic states about one's doxastic states.

Note that despite its name, SELF-KNOWLEDGE only alleges a requirement to have accurate doxastic states. This falls short of requiring knowledge, though I take it to be in the ballpark. Weakening the view I oppose in this way will if anything smooth the way for its being motivated by Moore's paradox. It might also come with fringe benefits, such as avoiding worries raised by Williamson's argument that no nontrivial states are luminous.[8] I will leave it open whether a requirement to genuinely *know* one's doxastic states is plausible, and, if not, whether the proponent of SELF-KNOWLEDGE can plausibly explain why accuracy still is required even if knowledge isn't.

SELF-KNOWLEDGE is deliberately a bit vague. I oppose all reasonable precisifications of it. Here are two examples:

> (SK+) Rationality requires that if one believes p, then one believes that one believes p.
>
> (SK-) Rationality requires that if one believes p, then one does not believe that one does not believe p.

The similarities between SK+ and SK- will matter more here than their differences. So I will mostly talk about the vague thesis SELF-KNOWLEDGE, and only mention particular precisifications when the differences matter.

There are also natural precisifications of SELF-KNOWLEDGE that trade in credences rather than beliefs, such as:

> (SK_{cr}) Rationality requires that if $Cr(p) = n$, then $Cr(Cr(p) \approx n) \approx 1$.

SELF-KNOWLEDGE and its precisifications all concern *requirements* of rationality. Some authors, like Tyler Burge (2013), speak instead of rational entitlements, or *permissions*.[9] They endorse something like:

> ($SK_{permission}$) If rationality permits one to believe that p, then rationality permits one to believe that one believes that p.

This claim is arguably weaker than SK+, at least if we suppose that rationality's requirements are consistent. Because it is weaker, it might be thought less vulnerable to objection. But my objections to SK+ and its cohort will mostly concern its alleged connection to Moore's Paradox, and I cannot see a way to argue from premises regarding Moore's Paradox to $SK_{permission}$ without at the same time arguing for SK+. So I think the specific objections that I raise here should extend to $SK_{permission}$.

[8] Williamson 2000, Ch. 4.

[9] See also de Almeida 2007, pg. 66.

3. The Significance of Self-Knowledge Requirements

The question whether rationality requires self-knowledge is important for how we think about self-knowledge, rationality, and their interrelations.

Start with self-knowledge. **Rationalism** is a loose collection of theories regarding **introspection**, our special way (or ways) of knowing our own mental states.[10] Very roughly, rationalist theories are ones that explain introspection in terms of rationality, reasoning, or responding to reasons.[11] For example, Sydney Shoemaker (1996) famously held that a belief that one believes p is constituted by the first-order belief p, plus sufficient *rationality*. And on Alex Byrne's (2005) transparency account, the transition from the belief that p to the belief that one believes p amounts to *good reasoning*.[12] And according to Christopher Peacocke (1998), Declan Smithies (2012b), and Aaron Zimmerman (2004), believing p gives one a *reason* to believe that one believes p. What these views share is the contention that failures to know one's beliefs are *rational failures*; that is, failures to be rational, to reason correctly, or to respond to one's reasons. In this way, they treat failures of self-knowledge more like failures to hold consistent beliefs than like failures to hold accurate beliefs about the external world. And so they dovetail nicely with the view that self-knowledge, like consistency, is rationally required.[13]

Consider now rationality. Looking over the usual suspects for rational requirements, some important similarities stand out. Despite recent protests from the 'knowledge first' crowd, most alleged requirements solely concern relations among a single agent's attitudes, or closely related matters like transitions between an agent's attitudes. In this respect, a requirement to have accurate beliefs about one's mental states fits the usual pattern. And the similarities don't end there. **Epistemic consequentialists** about the normativity of rationality claim that requirements are united in the effects they can be expected to produce, and that this explains their normative force. For example, consider common arguments for probabilism, the view that rationality requires probabilistic coherence among one's credences. Old-fashioned Dutch book arguments for probabilism appealed to the fact that agents with attitudes violating probabilistic coherence requirements are easy swindled. And newfangled accuracy-based arguments stress that for any set of credences violating probabilistic requirements, there is another probabilistically coherent set of credences that is a priori guaranteed to be more accurate.[14] But consequentialist intuitions are not limited to arguments for probabilism. They are present in the common claim that we in some sense pursue the aim of truth by adopting rationality as a means.[15] Self-knowledge requirements

[10] It is uncontroversial that we have a special capacity for knowing at least some of our own mental states, though Peter Carruthers (2011) and arguably Gilbert Ryle (1949) deny that capacity extends to belief.

[11] See Gertler 2011 and 2015 and Zimmerman 2008.

[12] The locus classicus is Evans 1982, Ch. 7, though similar thinking was expressed earlier by Edgley 1969. For other views like Byrne's, see Fernández 2013, Gallois 1996, and Setiya 2011. See Boyle 2011 and Moran 2001 for a different way of developing the transparency account, which is also congenial to rationalism.

[13] Cf. Gertler 2011, Ch. 5.

[14] See Pettigrew 2016 for helpful review, and Berker 2013 for critical discussion.

[15] See, e.g., Boghossian 2008, pg. 109 and BonJour 1985, pp. 7-8.

come out looking good by consequentialist lights, too. Violating them can make you vulnerable to Dutch books,[16] and trivially guarantees less accurate credences than you otherwise would have had.

Finally, consider the relations between rationality and self-knowledge. **Reflectivism** says that one's beliefs should be guided by one's higher-order knowledge of what rationality requires. Some reflectivists think reflective self-knowledge is a prerequisite for following other rational requirements.[17] But others concede that unreflective beings can follow ordinary first-order requirements. They instead regard it as an additional achievement of reflective beings that we follow higher-order requirements requirements like:

> (ENKRASIA) Rationality requires that one believe p only if one would judge on reflection that rationality permits one to believe that p.

Now I think the relationship between ENKRASIA and SELF-KNOWLEDGE is murkier than it is sometimes made out to be.[18] But I agree that it is hard to find a plausible unified conception of rationality on which it requires enkrasia but not self-knowledge. Many rational requirements concern relations among beliefs, and so errors about one's beliefs will tend to generate errors about what else one is required to believe. Thus the supporter of ENKRASIA can expect violations of SELF-KNOWLEDGE to lead to trouble.

4. Objections to Self-Knowledge Requirements

I am not the first to oppose self-knowledge requirements. But often when they are opposed, it is only especially strong requirements that are opposed, and they are opposed on the grounds that they are too strong. This can give the impression that even the opponents grant that rationality requires *some* degree of self-knowledge, and that what's debatable is merely how much.[19]

For example, one might oppose SK+ on the grounds that it is too demanding. Whenever I hold a belief, complying with the requirement alleged by SK+ will mean holding the further belief that I hold that belief. But then I will be required to hold a belief that I hold that further belief, and a further belief that I hold the further belief, and so on. This might be claimed to be metaphysically impossible, or at least psychologically unrealistic.[20]

But this is not a good reason to reject self-knowledge requirements in general. Similar worries can be raised for requirements connecting rationality and logic. Consider:

> (SINGLE-PREMISE CLOSURE) If p logically entails q, then rationality requires that if one believes that p, then one believes that q.

[16] Sobel 1987

[17] E.g., Shoemaker 1996, pp. 32-33.

[18] E.g., Burge 2013 and Smithies MS.

[19] E.g., Christensen 2007 and de Almeida 2007.

[20] Cf. Shoemaker 1996, and see de Almeida 2007 for related objections to a related claim by Williams (1994).

Some claim that sSINGLE-PREMISE CLOSURE is overly demanding, among other reasons for requiring an infinity of beliefs. Now one possible response is to reject any connection between rationality and logic.[21] But to many, this seems like overkill. Perhaps SINGLE-PREMISE CLOSURE should be qualified, or replaced by something else connecting rationality and logic in a different way. Or perhaps ideal rationality is extremely demanding.[22] Unless overdemandingness worries lead us to reject any connection between rationality and logic, they should not lead us to oppose self-knowledge requirements altogether.

The same goes for opposition to SELF-KNOWLEDGE on the grounds that higher-order beliefs are defeasible. Suppose I believe that my mother loves me, but there are psychologists lining up down the block to tell me that I don't really believe this. It could be claimed that if so, I should falsely believe that I don't believe that my mother loves me, in contrast with what SELF-KNOWLEDGE seems to say.[23]

Again, the same objections can be raised to SINGLE-PREMISE CLOSURE. If I rationally believe p, and have logicians lining up down the block to tell me (falsely) that p doesn't entail q, then intuitively I shouldn't believe q. Again, we might claim that SINGLE-PREMISE CLOSURE needs qualification. Or claim that I am faced with conflicting requirements.[24] Or bite the bullet, and claim that ideal rationality does require believing q.[25] Or something else. Again, whatever we say, it seems the same could go for self-knowledge requirements.

So even though I oppose SELF-KNOWLEDGE, I do not think it faces any *distinctive* problems involving overdemandingness and defeasibility. My objections are more fundamental. I oppose SELF-KNOWLEDGE for the same reason one might oppose:

> (KNOWLEDGE OF SURROUNDINGS) Rationality requires that you have accurate doxastic states about your immediate surroundings.[26]

Why reject KNOWLEDGE OF SURROUNDINGS? Well, it too might be thought too strong. Even if someone is looking right at a pink elephant, rationality might require her not to believe so if doctors are lining up to tell her she is hallucinating. But I take it that if we opposed KNOWLEDGE OF SURROUNDINGS merely for being too strong, we'd be missing the point. The problem with KNOWLEDGE OF SURROUNDINGS is not that it overstates the requirement to know your surroundings. It's that there is no such requirement. A brain in a vat can be perfectly rational despite wildly inaccurate beliefs about her surroundings. In a less extreme case, an agent with ordinary blindness might face a pink wall, and yet rationally fail to believe so, just because she doesn't see it. These counterexamples don't take for granted ordinary perceptual access, and then pile defeaters on top. Rather, they deny rational agents perceptual access altogether.

[21] Harman 1986

[22] Christensen 2004 and Smithies MS.

[23] Cf. Burge 2013, pp. 83-85; Moran 2001; and Shoemaker 1996.

[24] Cf. Lasonen-Aarnio 2014.

[25] Smithies MS.

[26] Gibbons 2013 accepts something approaching this.

I think SELF-KNOWLEDGE faces similar counterexamples. Following Sydney Shoemaker, let **self-blindness** be the condition of lacking ordinary introspective access to one's beliefs, despite possessing normal rationality, intelligence, and conceptual sophistication. Now consider:

> **Self-Blind George:** Fully rational George believes that it will rain, but he lacks normal introspective capacities. His non-introspective evidence supports that he does not believe it will rain.

Because George lacks introspective access to his beliefs, he should believe that he does not believe it will rain. Thus the case is a counterexample to SELF-KNOWLEDGE. And more importantly, it does not seem like the kind of counterexample that can be avoided with minor tweaks or qualifications, like the others.

But I admit it does not provide a dialectically effective objection to SELF-KNOWLEDGE. Any supporter of SELF-KNOWLEDGE will deny that cases like this are so much as possible. Indeed, Shoemaker's himself introduced the notion of self-blindness only to dispute its possibility. So while I think Self-Blind George is a counterexample to SELF-KNOWLEDGE, it persuades nobody.

So why should we be persuaded to reject SELF-KNOWLEDGE? Because we should reject cognate views about self-knowledge, rationality, and their interrelations.

First, we should reject rationalist theories of introspection in favor of competitors that treat introspection on the broad model of perception. If the actual mechanisms subserving introspection are *sui generis* and perception-like, then that sits poorly with the view that introspective self-knowledge is guaranteed simply by our general capacity for rationality.[27]

Second, we should reject epistemic consequentialism, and understand the normative significance of rationality in a way that prioritizes responsiveness to reasons.[28] This will count against self-knowledge requirements, because unless you are an expert about p, the fact that p is true is not a sufficient reason to believe that you believe that p. And so believing that p won't plausibly give you a sufficient reason to believe that you believe p.[29]

Finally, we should reject reflectivism in favor of theories that have agents responding directly to worldly reasons, rather than the the requirements imposed on them by rationality. This move could be coupled with a more general externalism, but not necessarily. One can be an internalist about reasons and rationality without being a reflectivist.

So as I see it, the merits of SELF-KNOWLEDGE cannot be decided independently of these very broad debates. But some proponents of self-knowledge requirements disagree. They think there is a quick and easy way to support SELF-KNOWLEDGE, and in doing so, to

[27] Cf. Gertler 2011, who attempts to reconcile a perceptual model of introspection with something like SELF-KNOWLEDGE.

[28] Cf. Kolodny 2005.

[29] Cf. Barnett 2016 and Byrne 2005.

support rationalism, or consequentialism, or reflectivism. This quick and easy support is supposed to stem from Moore's paradox.

Moore's paradox is often presented as a motivation for rationalism about introspection, in some discussions the primary motivation.[30] It is also a common motivation for reflectivism.[31] I have not seen an extended defense of epistemic consequentialism drawing on Moore's paradox. But it is common, in advancing controversial requirements of rationality, including on consequentialist grounds, to appeal to SELF-KNOWLEDGE as a largely undefended premise. Often a passing nod to Moore's paradox is included.[32]

So I think Moore's paradox is the official motivation for SELF-KNOWLEDGE if anything is. Even for theorists who simply find self-knowledge requirements intuitively appealing, I wonder if Moore's paradox might go some way towards bringing their attraction into sharper focus. It is often said that even though the asserter of a Moorean proposition does not contradict herself, there is some subtler kind of conflict present in her assertion. Proponents of self-knowledge requirements might similarly think that, even though an agent who violates them can remain consistent in her beliefs, there is some subtler sense in which her doxastic states are in conflict. I hope that my discussion of Moore's Paradox might connect with their thinking in some more indirect way.

5. Moore's Paradox

G. E. Moore himself observed only that it is somehow improper (or "absurd") to *assert* propositions of the form <p, but I don't believe that p>. Just what this impropriety consists in is not obvious. To streamline the arguments of my opponent, I will assume that it at least includes *irrationality* on the part of the agent, such that:

> (NO MOOREAN ASSERTIONS) Rationality requires one not to assert propositions of the form <p, but I don't believe that p>.[33]

NO MOOREAN ASSERTIONS faces the usual defeasibility and overdemandingness problems. And it arguably faces additional difficulties that do not affect SELF-KNOWLEDGE and SINGLE PREMISE CLOSURE. Unlike those claims, NO MOOREAN ASSERTIONS does not plausibly characterize a *basic* requirement of rationality. Instead, it at best is a consequence of more basic requirements. It also is implausible as a fully general requirement. It is not irrational to assert a Moorean proposition if ordered to at gunpoint, after all. What is plausible is that, given the knowledge, beliefs, and aims of a speaker in **normal circumstances**, by asserting a Moorean proposition one must violate rationality's requirements. Nothing I'll say in what follows will assume a stronger reading of NO MOOREAN ASSERTIONS than this.

[30] See, e.g., Fernández 2005 and 2013; Moran 2001, pp. 69-77; Shoemaker 1996; Silins 2012 and 2013; Smithies 2012b, 2016, and MS; and Zimmerman 2008, Sec. III.

[31] See, e.g., Gibbons 2013, Smithies 2012a and MS, Shoemaker 1996.

[32] E.g., Adler and Armour-Garb 2007; Douven 2009, pp. 367-8; Egan and Elga 2005, pg. 83; Huemer 2011; van Frassen 1995, pg. 19 and 1984, pg. 247.

[33] Cf. Chan 2008 and Fernández 2005, pg. 534.

Explaining NO MOOREAN ASSERTIONS is not so easy as explaining why, say, it would be irrational to assert a logical contradiction. As is commonly observed, Moorean conjunctions are logically consistent, and often true. In fact, the common observations undersell the difficulty of explaining NO MOOREAN ASSERTIONS. For it is moreover entirely possible for one to have sufficient evidence supporting a Moorean conjunction. This would obviously follow under the controversial assumption that self-blindness is possible, as I discuss in Section 7. But there are cases of evidentially supported Moorean conjunctions not involving self-blindness. Here is one inspired by Declan Smithies (2016):

> **Stubborn Stella:** Stella has sufficient meteorological evidence supporting that it will rain, but she stubbornly refuses to believe that it will rain. She has normal introspective capacities.

Stella can know by introspection that she does not believe it will rain. But her meteorological evidence supports that it will rain. So it seems her total evidence supports the Moorean conjunction <It will rain, but I do not believe that it will rain.>. Even so, Stella is in no position to rationally *assert* this conjunction.

Many attempts to explain NO MOOREAN ASSERTIONS have been proposed. The earliest often appealed to communicative aims or norms distinctive of assertion. Paradigmatic examples include Moore's own account in terms of what assertions "imply", Martinich's (1980) account in terms of the Gricean communicative intentions associated with assertion, and various Wittgensteinian accounts that posit a distinctive use of avowal statements like 'I don't believe that it will rain' for *expressing* a first-order lack of belief that it will rain, perhaps in addition to *describing* one's lack of belief.[34] (These accounts are usually pitched as explaining the *impropriety* of Moorean assertions, but they can be adapted to explain their *irrationality* by supposing that normally agents aim to avoid impropriety and know it when they see it.)

Recent decades have brought an expanded conception of the data that an account of Moore's paradox must explain. Consider:

> (NO MOOREAN BELIEFS) Rationality requires one not to believe propositions of the form <p, but I don't believe that p>.

Nowadays it is routine to claim NO MOOREAN BELIEFS as among the data—that is, among the obvious facts which any respectable account must seek to explain.[35] Some recent discussions go even further, and claim that NO MOOREAN BELIEFS is explanatorily prior to NO MOOREAN ASSERTIONS, in the sense that the irrationality of Moorean assertions should be explained in terms of the irrationality of Moorean beliefs. We will consider shortly whether these "epistemic accounts" of Moore's paradox should be accepted. First I want to say why they are important.

[34] See, e.g., Bar-On 2004, Heal 1994, and Rosenthal 1995.

[35] See, e.g., Chan 2010; de Almeida 2001 and 2007; Fernández 2005 and 2013, Ch. 4, pg. 112; Gibbons 2013, pp. 3 and 231; Heal 1994; Kriegel 2004; Moran 2001, pg. 70; Setiya 2011; Shoemaker 1996, Chs. 2, 4, and 11; Silins 2013, pg. 297; Smithies 2012b, 2016, and MS; and Williams 2006 and 2007.

There is an obvious argument that if we accept MOOREAN BELIEF, we also should accept at least SK-, if not stronger self-knowledge requirements. For this will follow if we accept:

> (MULTI-PREMISE CLOSURE) If p and q entail r, then rationality requires that if you believe that p and believe that q, then you believe that r.

Of course, MULTI-PREMISE CLOSURE faces the same overdemandingness and defeasibility difficulties as SINGLE PREMISE CLOSURE. And it additionally faces familiar problems involving the accumulation of risk of error over large numbers beliefs.[36] But again, I want to set these general difficulties aside. Later on, I will return to a potential problem for MULTI-PREMISE CLOSURE that I consider more relevant to its role in the argument for SELF-KNOWLEDGE. For now, let's move on.

If we accept MULTI-PREMISE CLOSURE and MOOREAN BELIEF, then we should accept SK-. If an agent were to jointly believe that p and that she does not believe that p, then by MULTI-PREMISE CLOSURE rationality would require that she either give up one of these beliefs or else believe the Moorean conjunction <p, but I don't believe that p>. By NO MOOREAN BELIEFS, she is required not to believe the Moorean conjunction. So rationality will require her to give up one of her beliefs—i.e., to either not believe that p or else not believe that she believes that p. Thus rationality will require her not to jointly believe that p and that she doesn't believe that p, as SK- holds.

6. First Route to SELF-KNOWLEDGE: Banning Foreseeable Errors

In Section 7, I will consider whether NO MOOREAN BELIEFS provides the best explanation of the obvious datum NO MOOREAN ASSERTIONS. But first, I want to consider the common claim that NO MOOREAN BELIEFS is a datum in its own right. I don't buy it. At best, what's an obvious datum is:

> (GUARANTEED ERROR) Beliefs in propositions of the form <p, but I don't believe that p> cannot be true.

Believing a conjunction of the form <p, but I don't believe that p> arguably entails believing the first conjunct. And if one believes the first conjunct, then the second conjunct will be false.[37] Maybe this makes GUARANTEED ERROR obvious. But NO MOOREAN BELIEFS would follow only given something like:

> (NO GUARANTEED ERRORS) Rationality requires that if you cannot have a true belief that q, then you do not believe that q.

Now some epistemologists accept this, because they accept:

[36] See, e.g., Christensen 2004.

[37] Note that it is less clear that beliefs in "commissive" conjunctions of the form <p, but I believe that not-p> cannot be true. So it might be objected that even if my opponent succeeds at motivating NO MOOREAN BELIEFS via GUARANTEED ERROR, she cannot motivate that it is irrational to believe propositions of the form <p, but I believe that not-p>. But I think something weaker still is plausible: that rational agents should consider such beliefs unlikely to be true. Since this weaker claim is arguably all my opponent needs, I set aside this objection to her view.

> (ONLY KNOWLEDGE) Rationality requires that you believe that q only if you know that q.[38]

But ONLY KNOWLEDGE flies in the face of philosophical orthodoxy. It entails, e.g., that brains in vats and the victims of misleading inductive evidence are unjustified in their beliefs. Even if it is true, it could hardly support the claim that NO MOOREAN BELIEFS is an obvious datum.

Perhaps there is another way to support NO MOOREAN BELIEFS, not by GUARANTEED ERROR itself, but instead by the claim that GUARANTEED ERROR is obvious. If GUARANTEED ERROR is obvious, then a rational agent with the relevant concepts is in a position to know it. So NO MOOREAN BELIEFS will arguably follow if we accept:

> (NO FORESEEABLE ERRORS) Rationality requires that if you can know that you cannot have a true belief that q, then you do not believe that q.[39]

But accepting NO FORESEEABLE ERRORS is not a good dialectical move for the proponent of self-knowledge requirements. Consider:

> **Unbelievable Consequences:** Sylvie knows that the Oracle's past predictions all turned out true, so she rationally believes that today's prediction will be true. The Oracle then predicts: "You, Sylvie, will not just now come to believe any new conjunctions." This new prediction seems plausible enough, so hearing it does not affect the rationality of Sylvie's belief that the prediction is true.

According to NO FORESEEABLE ERRORS, rationality requires Sylvie not to believe the conjunction <Today's prediction is that I won't now believe any new conjunctions, and today's prediction is true>. For it is foreseeable that this conjunction will be false if Sylvie believes it. (We can suppose Sylvie can know that she doesn't already believe it.)

At the same time, MULTI-PREMISE CLOSURE says that Sylvie is required to believe the conjunction, since she rationally believes each conjunct. So assuming rationality's requirements are consistent, accepting NO FORESEEABLE ERRORS forces us to reject MULTI-PREMISE CLOSURE.

But here is the problem. MULTI-PREMISE CLOSURE was a crucial premise in the argument from NO MOOREAN BELIEFS to SK-. Without it, the alleged irrationality of believing Moorean conjunctions would not impugn the rationality of jointly believing each conjunct. Thus even if NO FORESEEABLE ERRORS succeeds in motivating NO MOOREAN BELIEFS, it undermines the case for self-knowledge requirements in another way.

This is important, because it is SK- and its cohort that are relevant to broader issues concerning self-knowledge, rationality, and their interrelations. If Moorean beliefs are irrational only because they violate NO FORESEEABLE ERRORS, then arguably they are little

[38] E.g., Williamson 2014, pp. 989-991 and Littlejohn 2010.

[39] Cf. Silins 2013, pg. 297 and esp. Smithies 2016, to which this discussion is indebted. For further wrinkles involving commisive Moorean conjunctions, see de Almeida 2007 and Williams 1994.

more than idle curiosities. Much of the recent interest in Moore's paradox is driven by the assumption that it has something to teach us about rationality and self-knowledge. But under the present view, it's not so clear what it would be.

A proponent of self-knowledge requirements might reply that rationality's requirements are not mutually satisfiable for Sylvie. She is both required to believe the conjunction and required not to believe it. This view is unattractive in many ways. But in any case, it does not help with the present difficulty. For if we allow rationality's requirements to be inconsistent, then an opponent of self-knowledge requirements also could allow that sometimes one is both required to believe Moorean propositions and required not to believe them. And this allows her to claim that one can be required both to believe p and believe that one does not believe it—despite an alleged conflicting requirement not to believe one does not believe p. Although technically consistent with self-knowledge requirements, it is hard square this claim with an overall satisfying view connecting self-knowledge and rationality. So no matter what the response to Unbelievable Consequences, the broader significance of NO MOOREAN BELIEFS is questionable if we accept it only because we think foreseeable errors are irrational.

If this is right, then *any* proposed motivation of NO MOOREAN BELIEFS that appeals to NO FORESEEABLE ERRORS severs the connection between Moore's paradox and self-knowledge requirements. We can reinforce the point by considering some particular proposals, which appeal to an alleged connection between belief and the conscious mental act of *judgment*.

There are at least two ways to to defend NO FORESEEABLE ERRORS via the connection between belief and judgment. The first invokes the **higher-order thought theory of consciousness (HOT)**.[40] Proponents of HOT claim that a belief is conscious only if one believes that one has the belief. So if judgment is (or entails) conscious belief, then under HOT judging that q entails both believing that q and believing that one believes q. This makes NO FORESEEABLE ERRORS come out looking pretty good. For suppose one judges that q knowing that one cannot have a true belief that q. Under HOT, one will be guilty of straightforward logical inconsistency, by believing that q, that one believes q, and that if one believes q, then not-q.

The other defense of NO FORESEEABLE ERRORS is less widely discussed, though I suspect often covertly assumed. It takes judgment to be a kind of internal analogue to the public action of assertion. Shoemaker stresses an analogy between judgment and assertion in an influential passage where he introduces the now common claim that NO MOOREAN BELIEFS is as much a datum as NO MOOREAN ASSERTIONS.[41] And the analogy makes appearances in recent discussions from Alan Hájek (2007, pg. 219), Richard Moran (2001, pg. 70), Nico Silins (2012), Declan Smithies (2016 and MS), Timothy Williamson (2000, pp. 255-6), and Mitchell Green and John Williams (2007, pg. 3).

If judgments are analogous to assertions, how would this support NO FORESEEABLE ERRORS? I take the rough idea to be this. Just as assertions are actions governed by the aim of truthful assertion, judgments are mental acts governed by the aim of initiating (only) true

[40] Kriegel 2004; Shoemaker 1996, pp. 76-77; and Williams 2006.

[41] 1996, pp. 78-79.

beliefs.[42] So, it is irrational for an agent to judge as true a proposition which she knows she cannot have a true belief in. For in general, it is irrational to adopt an action that one knows will fail to meet the aims that motivate it. Call this view the **action conception of judgment**, because it takes judgments to be somehow analogous to the ordinary voluntary action of assertion.

These defenses of NO FORESEEABLE ERRORS are committed to rejecting MULTI-PREMISE CLOSURE. In addition to Unbelievable Consequences, this is brought out by examples like:

> **Unthinkable Consequences:** Robin has known for a long time that he only thinks about the one-hit wonder band Nena when he hears their song '99 Luftballons'. Today he is at the library, where he knows it is very quiet.

MULTI-PREMISE CLOSURE entails that Robin is required to believe that he is not currently thinking about Nena. But anyone who explains NO FORESEEABLE ERRORS by appealing to a tight connection between belief and judgment must deny this, regardless of their account of judgment. For it is obvious that Robin cannot initiate a true belief by *judging* that he is not thinking about Nena, simply because judging this involves thinking about Nena. Thus if the explanation of NO FORESEEABLE ERRORS is that one cannot rationally hold a foreseeably erroneous belief because one cannot rationally initiate the belief via judgment, then we must reject MULTI-PREMISE CLOSURE.

Indeed, I think proponents of these views should regard the rejection of MULTI-PREMISE CLOSURE as a welcome consequence. I'll explain why, focusing on the action conception of judgment. Although I oppose the action conception, I will try to develop it as sympathetically as I can, to see how it supports NO FORESEEABLE ERRORS, and opposes MULTI-PREMISE CLOSURE.

Judgments differ in many ways from ordinary actions, for example in their voluntariness. The action conception's explanation of NO FORESEEABLE ERRORS need not assume otherwise. It assumes only that judgments are subject to the same rational requirements as ordinary actions. Whether that means judgments must be voluntary, or in some other sense "up to us," is a further question.

To what requirements are ordinary actions subject? This is controversial, but fortunately the major points of disagreement won't seriously affect our discussion until later, in Section 7. For now I will adopt **causal decision theory (CDT)**, which holds that one is rationally required to select an action iff its **causally expected utility** is greater than each of one's other options. The causally expected utility of an action is a function of the agent's probability function, Pr, and utility function, v. I understand the probability function as representing the degree to which the agent's evidence supports various propositions she might entertain. Where the Ks are **dependence hypotheses**—i.e., maximal hypotheses about how outcomes depend causally on one's actions that form a partition—the causally expected utility of A-ing, or U(A), is defined as follows:

(1) $U(A) = \sum_K \Pr(K)v(KA).$

[42] Perhaps better: with the aim of initiating (only) a true belief at this very moment, by making the judgment, in the proposition judged. But see Berker (2013) for further worries.

Thus when one's options are to judge or to refrain, the action conception says that one is rationally required to adopt the option with the highest causally expected utility. Before considering how this explains NO FORESEEABLE ERRORS, let's consider some preliminary worries.

As it stands, the action conception is committed to pragmatism about judgment. Suppose my evidence supports a depressing possibility, and I value not being depressed above above having true beliefs. The action conception says that it is irrational for me to judge according to my evidence. But this commitment can be avoided or at least softened. Let an agent's **alethic values** be the subset of her overall values that concern attaining true belief and avoiding error. The action theorist could say that epistemic rationality is distinct from all-things-considered rationality. Whether epistemic rationality requires a judgment depends not on the agent's overall utility function, but instead on her alethic utility function, which represents solely her alethic values.

A related worry is that some agents, such as young children might lack alethic values altogether. There are a number of possible responses. One denies that agents without alethic values qualify as making judgments, even if they have mental episodes that resemble judgments in other ways. Another appeals to idealized versions of these agents, and the alethic values they allegedly would have. Yet another claims that the epistemic rationality of one's judgments depends on the *correct* alethic utility function, regardless of whether one accepts it.

A final preliminary point concerns the Jamesian distinction between the competing alethic values of adopting true beliefs and avoiding false ones. Suppose one considers whether to judge that q. Initiating (only) true beliefs is the only aim represented by one's alethic value function. So where *T* is that one initiates a true belief and *F* is that one initiates a false belief, when evaluating the causally expected utility of judging that p, the relevant partition of the dependency hypotheses is $\{\mathrm{J(q)} \Rightarrow \mathrm{T}, \mathrm{J(q)} \Rightarrow \mathrm{F}\}$. Thus judging that q is rational iff:

$$(2)\ \Pr\left[J(q)\Rightarrow T\right]v\left[T\right]+\Pr\left[J(q)\Rightarrow F\right]v\left[F\right]\geq U\left[\sim J(q)\right].$$

Since withholding judgment initiates no beliefs, the expected utility of withholding is a constant, which I hereby set at 0. Thus (2) reduces to:

$$(3)\ \Pr\left[J(q)\Rightarrow T\right]v\left[T\right]\geq -\Pr\left[J(q)\Rightarrow F\right]v\left[F\right].$$

Since $\mathrm{J(q)} \Rightarrow \mathrm{T}$ iff not-$[\mathrm{J(q)} \Rightarrow \mathrm{F}]$, (3) reduces to:

$$(4)\ \Pr\left[J(q)\Rightarrow T\right]v\left[T\right]\geq -\left(1-\Pr\left[J(q)\Rightarrow T\right]\right)v\left[F\right],$$

and therefore to:

$$(5)\ \frac{\Pr\left[J(q)\Rightarrow T\right]}{1-\Pr\left[J(q)\Rightarrow T\right]}\geq\frac{-v\left[F\right]}{v\left[T\right]}.$$

Thus the action conception should say that (5) is the condition for rationally judging that q. Note that insofar as one's alethic values affect the rationality of a judgment, what matters is the ratio of the disvalue of false belief to the value of true belief. For simplicity, I will assume this is a constant. But one could allow it to vary between agents, if one follows James' apparent permissivism about how "trigger happy" one should be with beliefs, or between contexts, if one wants the threshold for belief to vary with practical stakes. One could even replace an alethic value function with an epistemic value function, which evaluates beliefs not just by their truth, but by their status as knowledge. This modification might be necessary to accommodate the alleged fact that one should not believe one will lose the lottery. But I think it is an idle wheel in the explanation of Moore's paradox.[43] So I will assume only concern with the truth of one's beliefs.

This ends the preliminaries. The real explanatory work regarding Moorean judgment depends on how the action conception has the rationality of judgment depend on one's probabilities. It has the rationality of judging q depend not on the probability of q itself, but instead on the probability that if one were to judge that q, then one would initiate a true belief. This would be unimportant if for any q,

$$(6)\quad \Pr\left[J(q) \Rightarrow q\right] = \Pr(q).$$

But (6) is false for certain values of q that include Moorean conjunctions. For one's judging to be true a Moorean conjunction will cause it to be false, or perhaps even constitute its being false. Thus the probability of the Moorean conjunction can differ from that of the subjunctive conditional that if one were to judge it true, then it would be true.

Recall stubborn Stella, whose evidence supports that it will rain, but who knowingly refuses to believe it will rain. Stella's epistemic probability for the Moorean conjunction <It will rain, but I do not believe it will rain> is high. But the probability that the conjunction would be true if she judged it true is low. And under the action conception, it is the latter epistemic probability that matters.

This is how the action conception yields the desired result that it is irrational to judge to be true Moorean conjunctions, or any other proposition in which one knows one cannot have a true belief.[44] And this will entail NO FORESEEABLE ERRORS and NO MOOREAN BELIEFS given:

> (BELIEF→JUDGMENT) Rationality requires that one not believe that p unless one is willing to judge that p.

[43] Cf. Williamson 2000, Ch. 11 and Littlejohn 2010.

[44] Notice that the success of the action conception does not depend on its being developed using CDT rather than **evidential decision theory (EDT)**. Whereas CDT has the rationality of judging that q depend on Pr[J(q) ⇒ q], EDT has it depend on Pr[q|J(q)]. But this does not harm the explanation of NO MOOREAN BELIEFS. Just as Moorean conjunctions are exceptions to (5), they are exceptions to the claim that Pr[q|J(q)] = Pr[q].

This ends my attempt to sympathetically develop the action conception and its explanation of NO MOOREAN BELIEFS. I now will explain why it is committed to rejecting MULTI-PREMISE CLOSURE, which is a crucial premise in the argument for SK-.

A familiar theorem of the probability calculus is that if p entails q, then Pr(p) ≤ Pr(q). This makes SINGLE-PREMISE CLOSURE hard to deny under the simple picture that belief in a proposition is rational iff its probability exceeds an invariant threshold. Of course MULTI-PREMISE CLOSURE faces additional difficulties, since, e.g., a conjunction can have a lower probability than each of its conjuncts. Roughly, this is because the conjunction accumulates the error risk of each conjunct. For very long conjunctions, the accumulation of risk can be dramatic, and the probability of the conjunction can be far below that each conjunct. But the accumulation is more limited with only two conjuncts. So anyone who accepts the simple picture should accept particular instances of MULTI-PREMISE CLOSURE, involving only a small number of sufficiently probable conjuncts. Consider Stella, for example. Where *r* is that it will rain, and *B(r)* that she believes it will rain, we can suppose that Pr[r] ≈ Pr[B(r)] ≈ 1, and therefore that Pr[r & B(r)] ≈ 1.

The action conception, in contrast, admits more dramatic failures of MULTI-PREMISE CLOSURE, by divorcing the believability of a proposition from its probability. Even if it is probable that if one judged p one would initiate a true belief, and that if one judged q one would initiate a true belief, it can still be improbable that if one judged that p and q, then would initiate a true belief.[45] For example in Stella's case, even though

$$\text{(7)}\ \Pr\left[J(r)\Rightarrow r\right]\approx 1,$$

and

$$\text{(8)}\ \Pr\left[J(\sim B(r))\Rightarrow\sim B(r)\right]\approx 1,$$

it still is true that

$$\text{(9)}\ \Pr\left[J(r\&\sim B(r))\Rightarrow(r\&\sim B(r))\right]\ll 1.$$

Proponents of the action conception should welcome the rejection of MULTI-PREMISE CLOSURE. For it is closely related to how their view accommodates NO MOOREAN BELIEFS. By divorcing the believability of a proposition from its probability, the action conception allows for the violations of MULTI-PREMISE CLOSURE that allow its supporters to uphold NO MOOREAN BELIEFS in tricky cases like Stella's.

What's the upshot? The claim that NO MOOREAN BELIEFS is a datum is closely aligned with NO FORESEEABLE ERRORS, whether motivated by the action conception of judgment, or in some other way. But accepting NO FORESEEABLE ERRORS means rejecting MULTI-PREMISE CLOSURE, at least if rationality's requirements are consistent. And that means giving up on NO MOOREAN BELIEFS as a motivation for SK-. In short, if we say Moorean beliefs are obviously irrational, just because they are foreseeably false, this undermines the relevance of

[45] Fans of EDT should note that, similarly, Pr[p|J(p)] and Pr[q|J(q)] can both be high even when Pr[p&q|J(p&q)] is low.

Moore's paradox to self-knowledge requirements. Maybe some can get away with the claim that NO MOOREAN BELIEFS is a datum, but not supporters of self-knowledge requirements.

7. Second Route to SELF-KNOWLEDGE: Inference to the Best Explanation

Even if NO MOOREAN BELIEFS is not included among the data surrounding Moore's paradox, it still might be supported by the data. More specifically, NO MOOREAN BELIEFS might be supported via an inference to the best explanation from NO MOOREAN ASSERTIONS. This IBE motivation for NO MOOREAN BELIEFS has the potential to motivate SELF-KNOWLEDGE, since it is free of troublesome assumptions about the relations between rational belief, judgment, and foreseeable error.

Recall that **epistemic accounts** of Moore's paradox take the irrationality of Moorean assertion to be explained by the prior irrationality of Moorean belief. The guiding idea is this. Among the aims of agents in normal situations is the alethic aim not to assert falsehoods. Given this aim, it will *ceteris paribus* be irrational for an agent to assert a proposition unless she believes it to be true. And so if Moorean beliefs are irrational, so too are Moorean assertions.[46] The details will take some filling in. But this general line of explanation seems plausible.

I will argue, however, that the IBE strategy fails. I do not deny that epistemic accounts can explain NO MOOREAN ASSERTIONS. But I think another explanation—the ratifiability account—is also available. And moreover, there are further data surrounding Moore's paradox that only the ratifiability account can explain. This does not automatically show that epistemic accounts are false. It could be that the irrationality of Moorean assertions is overdetermined, so that multiple explanations of their irrationality are true. But it does undermine any support that NO MOOREAN BELIEFS might derive via inference to the best explanation.

We observed in Section 5 the wide range of explanations of Moore's paradox. In addition to epistemic accounts, there are **pragmatic accounts**, including Gricean accounts, Wittgensteinian expressivist accounts, and more. I don't know of any generally accepted criterion for an account's qualifying as pragmatic. But one salient feature of many such accounts is an appeal to aims (or norms) for assertion that go beyond an alethic aim to assert only truths. For example, some appeal to an aim to persuade one's audience in a certain way, and others to using avowals to express one's beliefs.

Although I oppose epistemic accounts, I want to say a word in their favor, and against pragmatic accounts. If pragmatic accounts were true, then Moorean assertions would be irrational even if one were unconcerned with whether the propositions one asserts are true. But agents surely are concerned with the truth of their assertions. And this concern alone seems sufficient to explain the irrationality of Moorean assertions. Just compare typical Moorean assertions with the following:

> **Sadie's Exam:** Sadie is taking a true or false exam, and aims to get as high a score as possible. For each statement on the exam, Sadie can mark it as true or refrain. She will receive a heavy penalty for each marked falsehood and a

[46] See de Almeida 2001 and 2007, Chan 2010, Kriegel 2004, Shoemaker 1996, and Williams 2006 and 2007. See also Green and Williams 2007 for review.

> small bonus for each marked truth, with the ratio of penalty to bonus equaling the ratio of the disvalue of false assertion to the value of true assertion. The first statement on the exam is 'It will rain, but I, Sadie, don't believe that it will rain.'

It seems irrational for Sadie to mark the statement. But she has no Gricean aims to convince an audience, or Wittensteinian aims to express herself. Nor does she have any other relevant non-alethic aims, such as a Williamsonian aim to mark only statements that she knows. (It might for example be entirely rational for her to mark 'My lottery ticket will lose'.) This does not automatically show that pragmatic accounts are false. Perhaps the irrationality of Moorean assertions is overdetermined, making multiple explanations of their irrationality true. But it does show that no pragmatic account tells the full story. Since it is irrational for Sadie to mark the statement given her (by stipulation) purely alethic aims, whatever explains this ought also to explain the irrationality of Moorean assertion for an agent in a normal situation, who also has alethic aims (perhaps among other aims).

In contrast to pragmatic accounts, epistemic accounts easily generalize to cover Sadie's Exam. These accounts say that one cannot rationally assert what one cannot rationally believe, so long as one has the appropriate alethic aim not to assert falsehoods. And a corresponding alethic aim is present in in Sadie's case. More generally, let **endorsement** be a general category covering both assertion, marking as true on Sadie's exam, or any other similar action regarding some statement which is governed by alethic aims, and where the ratio of disvalue of false endorsement to the value of true endorsement equals that for ordinary assertion. If the epistemic account is right that an agent with normal alethic aims cannot rationally assert what she does not believe, then it should be true more generally that

> (ENDORSEMENT→BELIEF, first pass) Rationality requires that one not endorse that p unless one believes that p.[47]

Now ENDORSEMENT→BELIEF is subject to the same difficulties and qualifications as NO MOOREAN ASSERTIONS itself. But that goes with the territory. The important thing is that with it, the epistemic account can explain something very close to NO MOOREAN ASSERTIONS. By NO MOOREAN BELIEFS, one cannot rationally believe a Moorean conjunction, and by ENDORSEMENT→BELIEF, one cannot rationally assert (or otherwise endorse) it without believing it. It follows that one cannot assert a Moorean conjunction without violating a requirement of rationality. This comes pretty close to an explanation of the datum NO MOOREAN ASSERTIONS.

But this epistemic account of NO MOOREAN ASSERTIONS faces two problems. The first is that ENDORSEMENT→BELIEF fails even in some normal situations, where the agent's aims are alethic. And the second is that it what it explains subtly falls short of the target datum. The second of these problems is the more serious, but it will be clearer after examining the first. Consider:

> **Ned's Exam:** Neutral Ned is taking an exam like Sadie's. When he reaches the final statement, he realizes that he has not yet endorsed a statement. So Ned is doubtful that he will endorse any of the statements on the exam. He

[47] See, e.g. Shoemaker 1996, pp. 76 and 213.

> then reads the final statement, which says 'I, Ned, will endorse a statement on this exam.'

When Ned reads this statement, he does not believe it. ENDORSEMENT→BELIEF implies that he should not endorse the statement, but this implication seems false. Ned can recognize that if he endorses it, then it will be true. So endorsing the statement is a safe way to add some points to his score.

An ENDORSEMENT→BELIEF supporter might reply that once Ned learns what the final statement is, he should believe that he will endorse a statement, even if he did not believe it previously. But it is hard to see why Ned should believe this, if not because he knows that he is rational, and that it is rational to endorse the statement. So this reply really presupposes my point.

A better reply seeks to contain the damage. This reply claims that ENDORSEMENT→BELIEF is *usually* true, despite special exceptions like Ned's Exam. To do the trick, this reply needs to explain why cases involving Moorean assertions are not among the exceptions.

Fortunately for epistemicists, CDT provides such an explanation. Well, sort of. For the explanation to work, we need to look past some incidental difficulties. The difficulties stem from the fact that ENDORSEMENT→BELIEF trades in belief, while standard formulations of CDT trade in either credences or probabilities. So reducing ENDORSEMENT→BELIEF to CDT requires some bridge principles connecting these notions. I will address them briefly, before getting on with things.

In Section 6, I tried to sympathetically develop the views of opponents who accept the action conception of judgment. So I invoked a version of CDT involving probabilities, which best suited their view. But now I have new opponents, epistemicists. And I think they are better off with another version of CDT, directly involving credences. This version also says that an action is rationally permissible iff it no other options exceed its causally expected utility. But it defines causally expected utility in terms of an agent's rational credence function, Cr, as follows:

$$(10)\ U(A)=\sum_K Cr(K)v(KA).$$

With some additional assumptions, CDT then yields a partial vindication of ENDORSEMENT→BELIEF. Where *E(q)* is that one endorses that q, CDT entails that endorsing q will be rational only if

$$(11)\ \frac{Cr\left[E(q)\Rightarrow q\right]}{1-Cr\left[E(q)\Rightarrow q\right]}\geq\frac{-v\left[E(q)\&\sim q\right]}{v\left[E(q)\&q\right]}.$$

Now consider:

$$(12)\ Cr\left[E(q)\Rightarrow q\right]=Cr(q).$$

In a wide range of cases (12) will be satisfied. And if those cases are otherwise normal, such that the agent's operative aims are merely to endorse truths and avoid endorsing falsehoods, then asserting q will be rational only if:

$$(13)\ \frac{Cr[q]}{1-Cr[q]} \geq \frac{-v[F]}{v[T]}.$$

So endorsing q is rational only if one's credence that q meets or exceeds a threshold that is determined by the ratio of the disvalue of false assertion to the value of true assertion. Call this the **threshold for assertion**.

This gets us pretty close to ENDORSEMENT→BELIEF, but not quite. For ENDORSEMENT→BELIEF trades in belief, and (13) trades in credences. To bridge the gap, the epistemicist must assume that belief is rational whenever credence above a threshold is rational, and that the threshold for belief is the same as (or at least no greater than) the threshold for assertion. I think this assumption is less controversial that it might at first appear. For the proponent of ENDORSEMENT→BELIEF does not need to assume that the threshold for belief is invariant, nor that the threshold for assertion is. What she must assume instead is that if they vary, then they vary together, along with variation in the ratio of the disvalue of falsehood to the value of truth. So she can accommodate the common claim that in a high stakes case one cannot rationally believe or assert that the bank will be open. (Still, I think she will struggle to explain the alleged unbelievability and unassertability of <My lottery ticket will lose>.)

The upshot is that given a number of assumptions, CDT entails ENDORSEMENT→BELIEF. Of these, (12) is the important one. It says that the agent's credence in q equals his credence in the proposition that if she were to endorse q, then q. Note that this is not satisfied in the case of Neutral Ned. He does not believe that he will endorse any statement. But he should believe that if he were to endorse a statement, then he would endorse one. So Ned provides no counterexample to the following refined claim:

> (ENDORSEMENT→BELIEF, final pass): Rationality requires that one not endorse that p unless one believes that if one were to endorse that p, then p.

This does not incorrectly entail that Ned should endorse. But it retains the original formulation's intuitive plausibility, and is supported by CDT, a widely accepted theory of rational decision. It also is still strong enough to entail that if Moorean beliefs are irrational, then so too are Moorean assertions. For (12) is satisfied where q is replaced by a conventional Moorean conjunction.[48] This marks an important difference between endorsement and judgment. We assumed in Section 6 that judgments typically cause one's beliefs. This was why (6) was violated by Moorean conjunctions. But one's own assertions and other endorsements are typically the effects of one's beliefs, rather than their causes. This is why (12) is not violated by Moorean conjunctions.

[48] Here the reliance on CDT rather than EDT is essential. For plausibly, Cr([p & ~B(p)] | E[p & ~B(p)]) ≪ Cr[p & ~B(p)].

All of this supports an epistemic account of Moore's paradox. So things are looking good for an IBE from NO MOOREAN ASSERTIONS to NO MOOREAN BELIEFS.

But now we are ready for the main problem with epistemic accounts: the explanation they offer is insufficiently general. Consider things first from my point of view, as one who rejects NO MOOREAN BELIEFS. I think cases like this are possible:

> **George's Exam:** Self-blind George is taking an exam like Sadie's. The first statement on the exam is 'It will rain, but I, George, do not believe it will rain.' George's meteorological evidence supports that it will rain, but his behavioral evidence supports that he does not believe that it will rain. So he rationally believes that the statement is true.

It is stipulated that George rationally believes the statement. And yet it still seems irrational for him endorse it. For even though his endorsing it would not cause it to be false, it still would be strong evidence that it already is false. And so George can reason that if he *does* endorse it, then by doing so he probably will incur a heavy penalty. At the same time, George rationally believes the Moorean conjunction, and thus should believe that if here *were to* endorse it, then it would be true. (These are consistent, because George should expect himself not to endorse the statement.) And thus an account relying on ENDORSEMENT→BELIEF cannot handle George's Exam.

Now proponents of NO MOOREAN BELIEFS will deny that George's Exam is possible, since they deny the possibility of self-blindness. So I will make the point in a few different ways. Consider:

> **Sigmund's Exam:** Sigmund has normal introspective capacities, and he believes his mother loves him. But the psychologists are lining up down the block to tell him he doesn't really believe this. On his exam, he encounters 'My mother loves me, but I do not believe that she loves me.'

Sigmund arguably can rationally believe the statement. NO MOOREAN BELIEFS appears to imply otherwise, and should be accepted only if this implication is avoided. This is no major problem for NO MOOREAN BELIEFS, since many attractive requirements face similar problems of defeasibility. But it is a problem for epistemic accounts, which go further in taking NO MOOREAN BELIEFS to explain the data surrounding Moore's paradox. For even though Sigmund can rationally believe the statement, he should refrain from endorsing, for the same reason as George. The epistemicist cannot explain why, unless they claim that it is irrational for Sigmund to believe the psychologists' testimony.

The epistemicist's problems do not end there. He must deny not only the potential rationality of Moorean beliefs, but their mere possibility. Compare:

> **Ira's Exam:** Ira irrationally believes the Moorean conjunction that it will rain, but that he does not believe that it will rain, despite normal introspective access to his beliefs. On his exam, the first statement is 'It will rain, but I, Ira, do not believe it will rain.'

> **Sonny's Exam:** Sonny irrationally believes that it will be sunny, despite strong evidence that it will rain. On his exam, the first statement is 'It will be sunny.'

Here we stipulate that Ira's belief is irrational, just like Sonny's. And yet there is an important difference between the cases. Although it is all-things-considered irrational for Sonny to endorse the statement 'It will be sunny', this is only because his belief is antecedently irrational. Endorsing the statement is no worse than believing it without endorsing it. Indeed, if he won't give up the irrational belief, the rationally least bad next move is to endorse. Otherwise he is merely compounding the irrationality of believing against his evidence with the irrationality of refusing to endorse a statement he believes.

Meanwhile, the irrationality of endorsing a Moorean proposition goes above and beyond the irrationality of believing it. For even though Ira believes the Moorean conjunction, he should still recognize that he would only endorse it if he believed it. (At least, that's some knowledge that any proponent of ENDORSEMENT→BELIEF should be willing to grant to Ira.) And Ira also should recognize that the statement is false if he believes it. Thus even if Ira believes the statement, he should accept that if he endorses it, then in so doing he will probably be endorsing a falsehood. This makes endorsing a Moorean conjunction an *additional* rational failure, even when it is already believed.

So if Sonny will not give up his irrational belief in an ordinary proposition, then he ought to endorse it. But if Ira will not give up his irrational belief in a Moorean proposition, then he still shouldn't endorse it. So even if the epistemicist explains one source of irrationality for Moorean assertions, there is some additional source he cannot explain.

The epistemicist promised us an explanation of the datum that by making a Moorean assertion, one violates rationality's requirements. But what she really explains is that one who makes a Moorean assertion either violates a rational requirement by doing so, or already has violated a requirement. Ira's Exam illustrates the difference.

The epistemicist could reply by denying the possibility of Moorean beliefs altogether. Or he might say that an agent with a Moorean belief is too far gone for us to meaningfully assess whether he should endorse. For comparison, suppose an agent believes that it will rain all day and be sunny all day. Should an agent with this belief endorse that it will not be sunny? It's hard to say. This agent's prior irrationality is so extreme that we are at a loss as to what next move is rationally least bad.

But even if these replies are granted, the underlying problem with ENDORSEMENT→BELIEF remains. These cases are part of a pattern highlighted recently by Andy Egan (2007), in which an agent's actions themselves amount to evidence about what their effects will be. Agents with entirely mundane failures of self-knowledge can find themselves in such positions, in ways that cause further trouble for ENDORSEMENT→BELIEF. Consider:

> **Rachel's Exam:** Rachel is rationally less than fully certain that she is ideally rational. In particular, she suspects herself of irrational risk aversion. On the exam, she encounters the statement: "I am not irrationally risk-averse." Rachel's credence in this statement falls short of belief, but a little more evidence would push her over the threshold. She is rationally quite certain

> that an irrationally risk averse person in her position would not endorse. And she thinks someone who is not risk averse might.

Rachel's Exam requires us to accept only that a rational agent could suspect herself of a common rational failing. This is hard to deny. There are some holdouts who think this is impossible for ideally rational agents.[49] But it's enough for us to claim that self-doubts are psychologically possible for ordinary agents, and that it does not make one too far gone to consider what one should do, given the self-doubts.

So what should Rachel do? It seems at least permissible for her to endorse. While she has doubts about the statement's truth, these go along with doubts that she will endorse. She is quite certain that she will not endorse the statement if it is false. So she can endorse, and rest assured that she is endorsing a truth.

If so, Rachel's Exam is a counterexample to ENDORSEMENT→BELIEF. For it says Rachel should not endorse. Rachel does not believe that if she were to endorse, then the statement would be true. If she is risk-averse, endorsing anyway wouldn't change that. Instead, she regards endorsing as evidence that the statement already is true. ENDORSEMENT→BELIEF doesn't allow this to make endorsing rational, and thus must be rejected.

Epistemicism is not looking so good anymore. But what is the alternative? Answering this question is a major undertaking, since all these cases, like those Egan proposes, are apparent counterexamples to CDT. A satisfying account must reject CDT, and provide another general theory of rational decision. We could adopt EDT, which gives accurate predictions in these cases. (I leave this as a take-home exercise.) But I follow Egan in worrying about its apparently false predictions for Newcomb and smoking lesion cases.

Egan's tentative suggestion appeals to the notion of **ratifiability**. Roughly speaking, an option is ratifiable if it still seems better than the alternatives on the assumption that one performs it. For a more precise statement, let U(B|A) be the expected utility of B-ing conditional on the assumption that one As, in the following (stipulative) sense:

$$(14)\quad U(B \mid A) = \sum\nolimits_K Cr(K \mid A)\, v(KB).$$

Thus an option A is ratifiable iff one has no other option O such that U(O|A) > U(A|A).

But what it the relationship between ratifiablity and rational action? Here is one proposal:

> (ABSOLUTE RATIFIABILITY) Rationality requires that if A-ing is unratifiable, then one does not A.

Endorsing a Moorean proposition is normally unratifiable. On the assumption that one endorses, the proposition endorsed is probably false—in which case refraining would have been in one's interest. Thus ABSOLUTE RATIFIABILITY can explain the irrationality of Moorean endorsement, including in tricky cases like George's and Rachael's. And it can explain why when Ira irrationally believes a Moorean conjunction, it would make things worse for him to endorse—for he will be violating an additional requirement of rationality

[49] E.g., Smithies forthcoming, Ch. 9.

by endorsing, unlike Sonny. So ABSOLUTE RATIFIABILITY seems promising as an explanation of NO MOOREAN ASSERTIONS.

Unfortunately, ABSOLUTE RATIFIABILITY is false as it stands, as Egan himself observes. Consider George, Sigmund, and Ira. It is not only irrational for agents in this cohort to endorse Moorean conjunctions, but also rationally permissible to refrain. And ABSOLUTE RATIFIABILITY says otherwise. For example, if George refrains, he will still believe the Moorean conjunction to be true. So imposing ratifiability as a necessary condition generates a rational dilemma. It says correctly that George is required not to endorse, but incorrectly that he is required not to refrain.

I still think the irrationality of Moorean assertions has something to do with their unratifiability. But if we reject ABSOLUTE RATIFIABILITY, we need another account of the connection between ratifiability and rational action. I won't here give a full defense of my account. But I'll tell you what I think, and how, if I am right, it would succeed at explaining NO MOOREAN ASSERTIONS.

I think ratifiability comes in degrees. Suppose one's options are to A or to B. Then A-ing's **degree of ratifiability** is defined as U(A|A) - U(B|A). Roughly and intuitively, the greater the degree to which A is ratifiable, the greater the degree to which A-ing has greater expected utility than refraining does, conditional on the assumption that one As. (Unratifiable options will thus have a negative degree of ratifiability.)

I propose (and more fully defend elsewhere):

> (COMPARATIVE RATIFIABILITY) Rationality requires that if option B is more ratifiable that option A, then one prefers B to A.[50]

When A and B are one's only options, preferring A to B licenses adopting A. So if endorsing a Moorean proposition is less ratifiable than refraining, then by COMPARATIVE RATIFIABILITY it will be irrational to endorse Moorean conjunctions, and thus permissible to refrain, assuming rationality's requirements are consistent.

I further propose the following **ratificationist** explanation of NO MOOREAN ASSERTIONS: Rationality requires one not to assert Moorean propositions because doing so violates COMPARATIVE RATIFIABILITY.

The ratificiationist proposal can handle all of the cases above. It handles Sadie, since any unratifiable action will have a lower degree of ratifiability than a ratifiable alternative. It also handles George and his cohort. (I leave Rachel as a another take-home exercise.) On the assumption that George refrains, refraining will not change his score, while endorsing will likely incur a small bonus—making U(endorse|refrain) - U(refrain|refrain) positive but low. On the assumption that George presses, refraining will not change his score, while endorsing will likely incur a heavy penalty—making U(refrain|endorse) - U(endorse|endorse) high. Again, although refraining is not ratifiable in an absolute sense for George, in graded terms it more ratifiable than endorsing is. So refraining is preferable to endorsing.

[50] Barnett MS. See also Wedgwood 2011 for a related proposal.

Maybe this is not quite the right account of the connection between ratifiability and rationality. But it is plausible that something like it is, and that it can explain NO MOOREAN ASSERTIONS. For even self-blind agents can appreciate that if they assert a Moorean conjunction, then they will probably be asserting a falsehood. Indeed, it seems that some explanation appealing to ratifiability or a closely related notion must be true, since other accounts cannot handle the data regarding George, Sigmund, and Ira.[51]

This does not automatically show that other accounts of NO MOOREAN ASSERTIONS are false, because the irrationality of ordinary Moorean assertion might be overdetermined. But without independent reason to accept a given epistemic or pragmatic account, we should not accept it simply because it explains NO MOOREAN ASSERTIONS. Thus NO MOOREAN BELIEFS, which is assumed by epistemic accounts, is not supported via an inference to the best explanation from NO MOOREAN ASSERTIONS.

8. Third Route to SELF-KNOWLEDGE: Shoemaker's Reductio

A final way of motivating SELF-KNOWLEDGE via Moore's paradox comes from Sydney Shoemaker, who argues:

> (SHOEMAKER'S THESIS) Self-blindness is impossible.

Shoemaker's strategy is to assume for the sake of *reductio* that self-blindness is indeed possible, and then argue that if so, a self-blind agent would show no sign of self-blindness in any of her behavior. Shoemaker regards this consequence as absurd, and rejects the assumption that self-blindness is possible.

This seems like the sort of thing an opponent of SELF-KNOWLEDGE like me should oppose. But the relationship between SHOEMAKER'S THESIS and SELF-KNOWLEDGE is tricky. After explaining why I oppose Shoemaker's argument, I'll return to some other motivations for SHOEMAKER'S THESIS that are compatible with rejecting SELF-KNOWLEDGE.

Let's say that a **self-aware** agent is a rational agent who who has introspective access to his or her own beliefs. Shoemaker's argument thus appeals to the following claim:

> (BEHAVIORAL INDISTINGUISHABILITY) Necessarily, any self-blind agent would act like a self-aware agent.

This is the premise that Shoemaker hopes to support via Moore's paradox, and it will be my main focus below. But first I want to register some doubts about whether, even if it is granted, it would help to support SHOEMAKER'S THESIS.

If it is possible for a self-blind agent to act just like a self-aware agent, then BEHAVIORAL INDISTINGUISHABILITY will be true while SHOEMAKER'S THESIS is false. So any deductive argument for SHOEMAKER'S THESIS would need supplementary premises strong enough to entail:

[51] Alternatively, if we adopt an error theory about our intuitions in these cases, as Ahmed 2012 urges, then ratifiability explains why Moorean assertions misleadingly appear to be irrational.

> (RESTRICTED BEHAVIORISM) Necessarily, any rational agent who acts like a self-aware agent is self-aware.

But it is hard to see what would motivate RESTRICTED BEHAVIORISM if not a more general behaviorism that says it is impossible for two agents to differ mentally without differing in their behavioral dispositions. I take general behaviorism to be implausible. So I am skeptical that Shoemaker can give us a well-motivated deductive argument for SHOEMAKER'S THESIS.[52]

Maybe the prospects are better for a probabilistic argument supporting SHOEMAKER'S THESIS. If BEHAVIORAL INDISTINGUISHABILITY is true, then our being rational is sufficient to give us the behavioral dispositions we in fact have. But presumably self-awareness cannot confer reproductive advantages without affecting our behavior in some way. And arguably we, as products of natural selection, would be unlikely to have a *sui generis* capacity for self-awareness that offers no reproductive advantages. As Shoemaker puts it, "[f]rom an evolutionary perspective it would certainly be bizarre to suppose that, having endowed creatures with everything necessary to give them a certain very useful behavioral repertoire…Mother Nature went through the trouble of instilling in them an *additional* mechanism…whose impact on behavior is completely redundant" (1996, pp. 239-240). Thus if we assume BEHAVIORAL INDISTINGUISHABILITY, the fact that we are self-aware probabilistically supports SHOEMAKER'S THESIS.

But this probabilistic argument faces serious objections. As we will see, Shoemaker's argument for BEHAVIORAL INDISTINGUISHABILITY is highly idealized. Even if successful, it would show only that a highly idealized self-blind agent could devise to act like a self-aware agent through elaborate lines of reasoning. This leaves open various possibilities for the evolutionary origins of introspection, even if Shoemaker is right. First, it could be that introspection offered reproductive advantage in our less sophisticated ancestors, and is vestigial in humans. This possibility should not be dismissed out of hand, because the existence introspection in non-humans, and its potential contribution to their behavioral repertoire, remains controversial.[53] More importantly, Shoemaker's argument leaves open the possibility that introspection confers behavioral advantages on ordinary humans, because of our own distance from idealized agents. For comparison, it might be that given my existing goals, and enough time for reflection, I can reason my way to a decision to duck when a heavy object flies at my head. This does not show that a reflex to duck confers no behavioral advantages. This too is no mere idle concern, since much empirical work on introspection involves situations where reaction times matter.[54]

So what advantages are conferred by our capacity for introspection? While this is an important question, I will avoid amateur speculation. If we reject SHOEMAKER'S THESIS, then plausibly the psychological mechanisms explaining introspection are a matter for empirical investigation. It it is dangerous to speculate about their evolutionary origins without a better understanding of how they work. What can responsibly be done from the

[52] For related discussion, see Kind 2003.

[53] See, e.g., Carruthers 2008 and 2011, Chs. 8-9; and Proust 2013, Ch 5.

[54] See, e.g., Metcalfe and Shimamura 1994.

armchair, in my view, is to reject Shoemaker's extreme claim that a contingent quasi-perceptual faculty for self-knowledge would have no behavioral effects at all.

It is time to take a closer look at BEHAVIORAL INDISTINGUISHABILITY. Why should we accept it? Shoemaker's argument for it is developed in papers spanning several decades, and resists easy summary. But I think the following valid argument exemplifies the most important moves:

> (NO MOOREAN ASSERTIONS) Rationality requires one not to assert propositions of the form <p, but I don't believe that p>.
>
> (CONFORMITY) Any self-blind agent would conform to rationality's requirements.
>
> (SEPARABILITY) If any self-blind agent would refrain from asserting <p, but I don't believe that p>, then any self-blind agent would be willing to assert "I believe that p" iff he believed that p.
>
> (PROXY) If any self-blind agent would be willing to assert "I believe that p" iff he believed that p, then any self-blind agent would act like a self-aware agent.
>
> Therefore, (BEHAVIORAL INDISTINGUISHABILITY) Necessarily, any self-blind agent would act like a self-aware agent.

Is this argument sound? NO MOOREAN ASSERTIONS is a datum, and CONFORMITY is trivial given the definition of self-blindness. Since the argument is plainly valid, an opponent must reject either SEPARABILITY or PROXY. I think we should reject both.

Defending SEPARABILITY, Shoemaker claims that the same rationality and conceptual sophistication allowing a self-blind agent to "appreciate the logical impropriety of affirming something while denying that one believes it" will enable her to "give appropriate answers to questions about what she believes" more generally. But is this true? In Section 7, I suggested the ratifiability account of NO MOOREAN ASSERTIONS, which (roughly) takes Moorean assertions to be irrational because (comparatively) unratifiable. But even when a compound action is unratifiable, its component actions can be individually ratifiable. So just because one cannot rationally endorse the conjunction <p, but I don't believe that p>, that does not mean one cannot rationally endorse each conjunct separately. Consider:

> **Permanent Marker:** George is taking a true or false exam in permanent marker. He believes that it will rain, but his behavioral evidence supports that he does not believe that it will rain. He encounters 'I, George, do not believe (right now) that it will rain.' He is required to decide now whether to endorse the statement, and he cannot change his answer later. Unbeknownst to George, he will subsequently encounter the statement 'It will rain'.

George believes that the first statement he encounters is true, and his endorsing it would not amount to evidence to the contrary. So he should endorse it. Then when he encounters the second statement, he will believe that it is true, and his endorsing it would be no evidence to the contrary. So he should endorse it, too.

It might seem like there is something sneaky going on. Being rational, George would know better than to endorse the Moorean conjunction 'It will rain, but I do not believe it will rain.' Shouldn't he also know better than to endorse each conjunct separately? As Shoemaker emphasizes, George should recognize the general truth that an agent's interests are best served by coordinating his answers to first-order questions about the world and higher-order questions about his own beliefs. Why is this not enough for him to refrain from endorsing the false higher-order statement that he does not believe it will rain?

The answer is that when George endorses the higher-order statement, he will think that he *is* coordinating. For he thinks he does not believe it will rain. And so he thinks that if he later encounters the first-order statement 'It will rain', then he will not endorse it. In general, knowing that his interests are served by coordinating won't enable George to coordinate, since he won't know what to do now in order to coordinate with what he will do later. This is why it matters whether George is presented with a Moorean conjunction, rather than with its conjuncts separately.

A possible rejoinder holds that George could overcome this problem with judicious contingency planning. There is nothing special about George being presented with a single Moorean conjunction, as opposed to being jointly presented with each of its conjuncts. If George knew in advance that he'd later be asked whether it would rain, perhaps he could make a joint decision, with the options of endorsing both, neither, the first but not the second, and the second but not the first. And just as he should avoid endorsing a Moorean conjunction, he should avoid the joint option of endorsing both. But there also is nothing special, the rejoinder holds, about cases where George knows in advance which statements lie in store. Regardless whether George is asked whether p or whether he believes that p, George can decide on an overall plan about how to answer both questions going forward.

I think this view faces a lot of tricky technical problems. For example, it is unclear how it can deal with cases where George changes his mind about whether p after making a plan. But set aside the technical problems. Even if the rejoinder succeeds in supporting BEHAVIORAL INDISTINGUISHABILITY, it does further harm to the broader argument for SHOEMAKER'S THESIS. For the rejoinder must help itself to a great deal of idealization. If George can succeed in acting self-aware, it is only via elaborate contingency planning that nobody really engages in. And as we saw above, the more idealization needed to support BEHAVIORAL INDISTINGUISHABILITY, the weaker the probabilistic argument from BEHAVIORAL INDISTINGUISHABILITY to SHOEMAKER'S THESIS. If faking self-awareness requires super-human rationality, then a *sui generis* faculty of introspection can still confer advantages on mere humans.

Perhaps the defender of SEPARABILITY should change tacks, and reject ratificationism in favor of an alternative account of Moore's paradox, such as Wittgensteinian expressivism. Shoemaker comes close to this, suggesting that the words "I believe" function as an "assertion sign".[55] It's not clear exactly what this claim comes to. On an extreme reading, it is that an avowal of "I believe it will rain" has the same truth conditions as an assertion that it will rain. But it is hard to accept that I am barred from using "I believe" to make assertions regarding my own psychology, rather than the weather. A moderate alternative suggested by Richard Moran holds merely that one who avows belief that p is committed to

[55] 1996, pg. 36.

the truth of p.[56] If so, then perhaps an agent in normal circumstances should aim to avoid avowing false beliefs. Understood the right way, this might support:

> (AVOWAL→BELIEF) Rationality requires that one avow belief that p only if one believes that p.

Now AVOWAL→BELIEF is strong enough to entail that self-blind agents will act like self-aware agents in cases like:

> **Ersatz Belief:** George's behavioral evidence supports that he believes that it will rain, but he does not believe that it will rain.

If AVOWAL→BELIEF is assumed, then it would follow that George should not avow belief that it will rain, despite his third-person evidence. But AVOWAL→BELIEF cannot handle other cases. We also need something like:

> (BELIEF→AVOWAL) Rationality requires that one be willing to avow belief that p if one believes that p.

This claim can handle cases like:

> **Covert Belief:** George believes that it will rain, but lacks behavioral evidence supporting that he has this belief.

A self-aware agent in George's situation would be willing to avow the belief, since she could know introspectively that she has the belief. If BELIEF→AVOWAL is true, then so would George.

Now I am skeptical that we should jointly accept AVOWAL→BELIEF and BELIEF→AVOWAL. The rationale for AVOWAL→BELIEF is that rational agents should aim to avow belief that p only when p is true, because in avowing belief that p, one commits oneself to the truth of p. A related rationale for BELIEF→AVOWAL would hold that rational agents should aim to avow belief whenever p is true (and when one is asked whether one believes p, etc.), presumably because in avowing the belief, one commits oneself *only* to the truth of p. But now this is starting to look like the extreme view that the truth conditions for an avowal do not involve the speaker's psychology, but instead only the worldly facts the avowed belief is about.

But there is a further problem for any attempt to support Shoemaker's argument via AVOWAL→BELIEF and BELIEF→AVOWAL. For even if these claims support SEPARABILITY, they do so at the expense of PROXY. For AVOWAL→BELIEF and BELIEF→AVOWAL make assertions a poor proxy for other actions. Consider:

> **Anxiety Drug:** George believes that he has cancer, but his third-person evidence supports that he does not believe this. An anxiety drug is known to

[56] 2001, pg. 70-77.

> substantially benefit people who believe themselves to have cancer, regardless of whether they do have cancer. It offers no benefit to those without the belief, and it has minor side effects, which are somewhat exacerbated by cancer. Even so, the side effects are outweighed by the benefits for all those with the belief.

A self-aware agent in George's position should take the drug, and also be willing to avow belief that he has cancer. So Shoemaker must claim that George would do the same. The Wittgensteinian claims that George should aim to avow belief that he has cancer whenever the has cancer, and thus should avow. So far, so good. But George should not have an aim of taking the drug whenever he has cancer. He knows that having cancer only makes the side effects worse! Thus the Wittgensteinian defense of SEPARABILITY simply makes avowals a poor proxy for other actions evincing self-awareness.

This is related to a more general problem we saw in Section 7 for Wittgensteinian and other pragmatic accounts of Moore's paradox. Because these accounts explain the irrationality of Moorean assertions in terms of distinctive properties of assertion, they are unable to explain closely related phenomena, such as the irrationality of endorsing Moorean statements on a true or false exam. Likewise, the Wittgensteinian is unable to extend his account of avowals to actions like taking the anxiety drug.

It might seem like the ratificationist account that I favor is friendlier to treating avowals as a proxy for other actions. It explains the irrationality of Moorean assertions in terms of more general features they share with other Moorean endorsements. Indeed, these features are clearly present in some cases involving actions other than endorsements, such as:

> **Ingrown Toenail Drug:** George believes that he has an ingrown toenail, but his third-person evidence supports that he lacks this belief. A drug is known to cure ingrown toenails, but it has side effects. First, it causes cancer in anyone without an ingrown toenail. Second, it causes extreme anxiety in anyone who believes himself to have a medical problem, no matter how minor.

Suppose George's evidence that he has an ingrown toenail is so conclusive that it outweighs the risk of the first side effect. Even so, taking the drug is irrational, for the same reason Moorean endorsements are. Because of the first side effect, only one who believed himself to have an ingrown toenail would take the drug. And hence one's taking it is strong evidence that the drug will cause the second side effect. So ratificationism says that George, being rational, won't take the drug. And that is just what a self-aware agent, who introspectively knows that he believes he has an ingrown toenail, would do. Score this as a win for PROXY, at least if we accept ratificationism.

But it's a Pyrrhic victory. Ratificationism still would not have George act like a self-aware agent in general. We already saw that ratificationsists should reject SEPARABILITY. In cases like Permanent Marker, when George faces a solitary question about high beliefs, his rationality will not allow him to answer accurately. The same point holds in Anxiety Drug, where the advisability of an action likewise depends on what George believes. Here too, ratificationism arguably says that George should not take the drug. His evidence supports that he lacks a belief that he has cancer, and thus that the drug will be useless. And his taking the drug would be no evidence to the contrary. So ratificationism would not have George act like a self-aware agent.

Indeed, even if some rejoinder to these problems for SEPARABILITY is successful, another problem for PROXY remains. Consider:

> **Umbrella Rental:** George believes that it will rain, but his third-person evidence supports that he is uncertain whether it will rain. An umbrella vendor offers short-term umbrella rentals. The cost of the rentals are balanced with the unpleasantness of getting wet so that a rational agent will rent an umbrella unless she believes it will be useless, in which case she won't rent. But today they are all out of ordinary umbrellas. Instead, they have a special umbrella that only opens if the agent did not believe it would rain at the time of rental.

A self-aware agent in George's situation should not rent the umbrella. For she believes it will rain, and can know that she does introspectively. And thus she should believe it will be useless. Meanwhile, George should rent. For he should believe that the umbrella will open, and his renting would provide little evidence to the contrary.

The crucial difference between Umbrella Rental and Ingrown Toenail is this. Taking the drug is strong evidence of belief in an ingrown toenail, because the potential costs are so high. But the relative costs of renting a useless umbrella are low. And so renting is weak evidence of belief that it will rain, and strong evidence only of a lack of belief to the contrary.

This raises a fundamental problem for PROXY. Assertions and endorsements have a crucial feature that many actions lack. Namely, they are strong evidence of belief. The ratificationist account has this feature do all the work in explaining the irrationality of Moorean assertions. Many "precautionary" actions, like taking an umbrella, are not strong evidence of belief. So even if George could feign self-awareness for endorsement-like actions, he won't for precautionary actions. Moorean precautionary actions, like renting the umbrella, are under ratificationism entirely rational for a self-blind agent.

So much for Shoemaker's argument for SHOEMAKER'S THESIS. Are there other grounds for accepting it? Jonathan Simon stressed to me the difficulty of imagining *being* self-blind as evidence of its impossibility. Others might appeal to independent commitments that, e.g., some mental states are necessarily self-intimating.[57]

I am skeptical that imaginability is a good guide to possibility here. Many *actual* psychological conditions are hard to imagine. And while necessary self-intimation has some plausibility for phenomenal states, I think it's less plausible for beliefs. But in any case, if we accept SHOEMAKER'S THESIS for *these* reasons, it's not clear that commits us to SELF-KNOWLEDGE.

Look at it this way. Even if rationality does not require self-knowledge, perhaps it is independently impossible for any agent, regardless of their rationality, to lack self-knowledge. But if so, it would be just as impossible for a *morally perfect* agent to lack self-knowledge. That would hardly show that we are morally required to know our minds.

[57] See Block (1995) for critical discussion and review

So, rejecting SELF-KNOWLEDGE does not automatically oblige us to reject SHOEMAKER'S THESIS. Perhaps there are reasons for accepting SHOEMAKER'S THESIS that are neutral on the relationship between self-knowledge and rationality. Now I'm not saying that Shoemaker's reasons fall into that category. They seem to trade on a self-blind agent's rationality in a way that they don't to, say, her moral character. This is why I think opponents of SELF-KNOWLEDGE should reject Shoemaker's argument for SHOEMAKER'S THESIS as unsound. Even so, we could perhaps accept its conclusion on other grounds.

9. Conclusion

Supporters of self-knowledge requirements hold that self-knowledge is required by rationality, like avoiding inconsistency, or adopting means apparently conducive to one's ends. We have seen little support for this claim from Moore's paradox. But have we seen positive reason to reject it? That depends on the proposed relationship between self-knowledge and other requirements or rationality.

Self-knowledge requirements might be proposed as *basic* requirements, alongside familiar requirements of consistency and the like. This proposal allows an agent to satisfy the familiar requirements while lacking self-knowledge. But it maintains that such an agent would be irrational, simply by violating a further basic requirement. We have seen little support for such a view from Moore's paradox. But it is difficult to rule out other sources of support, such as broader commitments to reflectivism or epistemic consequentialism.

But there is a more extreme view about the relation of self-knowledge requirements to familiar requirements which we have seen sufficient reason to reject. This view takes self-knowledge requirements to be consequences of familiar requirements. Once an agent satisfies familiar requirements, this view says, he automatically satisfies a requirement to know his mind. Shoemaker is plausibly a supporter of this extreme view. When he says that self-blindness is impossible, he doesn't just mean that an otherwise rational agent would not count as fully rational if he violated self-knowledge requirements. He means that you cannot be rational in familiar, uncontroversial ways without satisfying self-knowledge requirements.[58]

This extreme view should raise eyebrows. Casually reviewing the usual list of rational requirements, like the one that began this paper, it is hard to see any that plausibly entail self-knowledge requirements. But in any case, we have seen more direct reason to reject it. In discussing Shoemaker's *reductio* we examined the behavior of a hypothetical agent who satisfied the familiar requirements of epistemic and prudential rationality, but who had mistaken beliefs about his own beliefs. This hypothetical assumption generated intelligible predictions for his behavior, and led to no obvious contradictions.

There is thus no apparent contradiction in supposing that an agent violates alleged self-knowledge requirements and yet has beliefs which are are coherent and responsive to her evidence, and actions which conform to familiar requirements of prudential rationality such as those proposed by CDT and its competitors. This does not show that self-blindness is possible, or that rationality fails to require self-knowledge. But it does show that self-knowledge requirements do not follow from ones we already accept. Any requirement to know one's own mind must be a further basic requirement, which requires independent

[58] See, e.g., 1996, pp. 32-33.

motivation. And the main candidate source of motivation, Moore's paradox, doesn't seem to provide any.[59]

[59] For helpful discussion, I wish to thank Arif Ahmed, Paul Boghossian, Sinan Dogramaci, Elliot Paul, Jim Pryor, Timothy Rosenkoetter, Nico Silins, Jonathan Simon, Declan Smithies, Jonathan Weisberg, and an audience at the Northern New England Philosophical Association.

References

Ahmed, Arif (2012) 'Push the Button' *Philosophy of Science* 79(3): 386-395.

Bar-On, Dorit (2004) *Speaking My Mind: Expression and Self-Knowledge* Oxford University Press.

Barnett, David James (2016) 'Inferential Justification and the Transparency of Belief' *Noûs* 50(1): 184-212.

——— (MS) 'Graded Ratifiability'

Berker, Selim (2013) 'Epistemic Teleology and the Separateness of Propositions' *Philosophical Review* 122(3): 337-393.

Block, Ned (1995) 'On a Confusion About a Function of Consciousness' *Brain and Behavioral Sciences* 18(2): 227-247.

Boghossian, Paul (2008) *Content and Justification: Philosophical Papers.* OUP.

BonJour, Laurence (1985) *The Structure of Empirical Knowledge.* Harvard.

Boyle, Matthew (2011) 'Transparent Self-Knowledge' *Supplementary Proceedings of the Aristotelian Society* 85(1): 223-241.

Broome, John (2013) *Rationality Through Reasoning* Wiley-Blackwell.

Buchak, Lara (forthcoming) 'Decision Theory' in *Oxford Handbook of Probability and Philosophy*, Christopher Hitchcock & Alan Hájek (eds.), Oxford University Press.

Burge, Tyler (2013) *Cognition Through Understanding.* OUP.

Byrne, Alex (2005) 'Introspection' *Philosophical Topics* 33: 79-104.

Carruthers, Peter (2008) 'Metacognition in Animals: A Skeptical Look' *Mind and Language* 23(1): 58-89.

——— (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford.

Chan, Timothy (2008) 'Belief, Assertion, and Moore's Paradox' *Philosophical Studies* 139(3): 395-414.

——— (2010) 'Moore's Paradox is Not Just Another Pragmatic Paradox' *Synthese* 173(3): 211-229.

Christensen, David (2004) *Putting Logic in its Place: Formal Constraints on Rational Belief* Oxford: Oxford University Press.

——— (2007) 'Epistemic Self-Respect' *Proceedings of the Aristotelian Society* 107(1pt3): 319-337.

de Almeida, Claudio (2001) 'What Moore's Paradox Is About' *Philosophy and Phenomenological Research* 62(1): 33-58.

——— (2007) 'Moorean Absurdity: An Epistemological Analysis' in *Moore's Paradox: New Essays on Belief, Rationality, and the First-Person*, Mitchell Green and John Williams, eds. Oxford: Oxford University Press.

Edgley, Roy (1969) *Reason in Theory and Practice*. London: Hutchison & Co.

Egan, Andy (2007) 'Some Counterexamples to Causal Decision Theory' *Philosophical Review* 116(1): 93-114.

Egan, Andy and Elga, Adam (2005) 'I Can't Believe I'm Stupid' *Philosophical Perspectives* 19: 77-93.

Evans, Gareth (1982) *The Varieties of Reference*. Oxford: Oxford University Press.

Fernández, Jordi (2005) 'Self-Knowledge, Rationality, and Moore's Paradox' *Philosophy and Phenomenological Research* 71(3): 533-556.

——— (2013) *Transparent Minds,* OUP.

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

Gertler, Brie (2011) *Self-Knowledge*. London: Routledge.

——— (2015) "Self-Knowledge", *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2015/entries/self-knowledge/>.

Gibbons, John (2013) *The Norm of Belief.* OUP.

Greco, Daniel (2014) 'A Puzzle About Epistemic Akrasia' *Philosophical Studies* 167(2): 201-219.

Green, Mitchell and John Williams (2007) 'Introduction' in *Moore's Paradox: New Essays on Belief, Rationality, and the First-Person*, Mitchell Green and John Williams, eds. OUP.

Hájek, Alan (2007) 'My Philosophical Position Says p and I Don't Believe p' in *Moore's Paradox: New Essays on Belief, Rationality, and the First-Person*, Mitchell Green and John Williams, eds. OUP.

Harman, Gilbert (1986) *Change in View.* MIT.

Heal, Jane (1994) 'Moore's Paradox: A Wittgensteinian Approach' *Mind* 103(409): 5-24.

Horowitz, Sophie (2014) 'Epistemic Akrasia' *Noûs* 48(4): 718-744.

Huemer, Michael (2011) 'The Puzzle of Metacoherence' *Philosophy and Phenomenological Research* 82(1): 1-21.

Kolodny, Niko (2005) 'Why Be Rational?' *Mind* 114(455): 509-563.

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

Kind, Amy (2003) 'Shoemaker, Self-Blindness, and Moore's Paradox' *Philosophical Quarterly* 53(210): 39-48.

Lasonen−Aarnio, Maria (2014) 'Higher-Order Evidence and the Limits of Defeat' *Philosophy and Phenomenological Research* 88(2): 314-345

Littlejohn, Clayton (2010) 'Moore's Paradox and Epistemic Norms' *Australasian Journal of Philosophy* 88(1): 79 – 100.

Martinich, A. P. (1980) 'Conversational Maxims and Some Philosophical Problems' *Philosophical Quarterly* 30(120): 215-228.

Metcalfe, Janet and Shimamura, Arthur (1994) *Metacognition: Knowing about Knowing*. MIT.

Moran, Richard (2001) *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.

Peacocke, Christopher (1998) 'Conscious Attitudes, Attention, and Self-Knowledge' in *Knowing Our Own Minds*, Crispin Wright, Barry Smith, and Cynthia Macdonald eds., Oxford: Oxford University Press.

Pettigrew, Richard (2016) 'Epistemic Utility Arguments for Probabilism' *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.) URL = <https://plato.stanford.edu/archives/spr2016/entries/epistemic-utility/>.

Proust, Joëlle (2013) *The Philosophy of Metacognition*. OUP.

Rosenthal, David (1995) 'Self-Knowledge and Moore's Paradox' *Philosophical Studies* 77(2/3): 195-209.

Ryle, Gilbert (1949) *The Concept of Mind,* University of Chicago Press

Setiya, Kieran (2011) 'Knowledge of Intention' in *Essays on Anscombe's Intention*. Anton Ford, Jennifer Hornsby, and Frederick Stoutland, eds., Cambridge, MA: Harvard University Press.

Shoemaker, Sydney (1996) The First-Person Perspective and Other Essays. Cambridge: Cambridge University Press.

——— (2009) 'Self-Intimation and Second-Order Belief' *Erkenntnis* 71(1): 35-51.

Silins, Nicholas (2012) 'Judgment as a Guide to Belief' in Declan Smithies & Daniel Stoljar (eds.), *Introspection and Consciousness*. OUP.

——— (2013) 'Introspection and Inference' *Philosophical Studies* 163(2): 291-315.

Smithies, Declan (2012a) 'Moore's Paradox and the Accessibility of Justification' *Philosophy and Phenomenological Research* 85(2): 273-300.

——— (2012b) 'A Simple Theory of Introspection' in *Introspection and Consciousness*, Declan Smithies and Daniel Stoljar eds., New York: Oxford University Press.

——— (2016) 'Belief and Self−Knowledge: Lessons From Moore's Paradox' *Philosophical Issues* 26(1): 393-421.

——— (MS) *The Epistemic Role of Consciousness*

Sobel, Jordan Howard (1987) 'Self-Doubts and Dutch Strategies' *Australasian Journal of Philosophy* 65(1): 56-81.

Wedgwood, Ralph (2007) *The Nature of Normativity*. OUP.

——— (2011) 'Gandalf's Solution to the Newcomb Problem' *Synthese* 14: 1-33.

Williams, John (2006) 'Moore's Paradoxes and Conscious Belief' *Philosophical Studies* 127: 383-414.

——— (2007) 'Moore's Paradox, Evans's Principle, and Iterated Beliefs' in *Moore's Paradox: New Essays on Belief, Rationality, and the First-Person*, Mitchell Green and John Williams, eds. Oxford: Oxford University Press.

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

Zimmerman, Aaron (2004) 'Unnatural Access' *Philosophical Quarterly* 54(216): 435-438.

——— (2008) 'Self-Knowledge: Rationalism vs. Empiricism' *Philosophy Compass* 3(2): 325-352.