David James Barnett

Self-Knowledge Requirements 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 else 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 self-knowledge, does rationality at least require something in the ballpark, like true higher-order belief? Many philosophers have thought so, and have taken Moore's paradox to support their thinking. This paper will oppose these common views.

2. Self-Knowledge Requirements

Fans of self-knowledge requirements have some flexibility in how to formulate them. But this won't matter too much, because I will argue that Moore's paradox fails to motivate even the weakest formulations.

[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 Wedgwood 2007, Ch 1.

[7] See Buchak forthcoming for review.

A first question is what the scope of the required self-knowledge is. Because of its especially direct connection to Moore's paradox, my focus here will be on knowledge of one's *beliefs*. Perhaps anyone who says that rationality requires knowledge of beliefs should be prepared to say that it also requires knowledge of other mental states, on pain of arbitrariness. Extending self-knowledge requirements to other mental states might raise additional difficulties for my opponents, but I will set them aside.

Restricting the requirement to beliefs still leaves a lot of flexibility. Here are some examples of requirements for self-knowledge, or at least something in the ballpark:

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

Of these, only SELF-KNOWLEDGE+ is a *bona fide* requirement for self-knowledge. In contrast, SELF-KNOWLEDGE does not require one to know one's beliefs, but instead only to have true higher-order beliefs about them. SELF-KNOWLEDGE- is even weaker. Instead of requiring the presence of true higher-order beliefs, it requires merely the absence of false ones.[8]

Even so SELF-KNOWLEDGE and SELF-KNOWLEDGE- are arguably close enough to SELF-KNOWLEDGE+ to retain much of its interest. A principal source of interest in SELF-KNOWLEDGE+ is that it makes self-knowledge out to be very different from knowledge of other deeply contingent topics. SELF-KNOWLEDGE+ says that by failing to know one's beliefs, one will violate a requirement of rationality. Thus SELF-KNOWLEDGE+ entails Sydney Shoemaker's thesis that **self-blindness** is impossible, where self-blindness is stipulated to be the condition of lacking knowledge of one's beliefs despite possessing idealized rationality, intelligence, and conceptual sophistication. In contrast, an agent can be fully rational and yet be ignorant of various facts about her surroundings, due for example to perceptual limitations like ordinary blindness. So SELF-KNOWLEDGE+ proposes a special relationship between rationality and self-knowledge.

SELF-KNOWLEDGE and SELF-KNOWLEDGE- likewise allege a special relationship between rationality and higher-order belief. Just as one can rationally fail to know facts about one's surroundings, one can rationally fail to have true beliefs about them. According to SELF-KNOWLEDGE, things are different regarding one's own beliefs. If you in fact believe that p, and you are rational, then you will have the true belief that you believe p. While SELF-KNOWLEDGE- takes us a step further from requiring self-knowledge, it still preserves a special relationship, at least on the plausible assumption that false beliefs about most contingent matters of fact can be rational. This assumption is plausible, because it seems

[8] Throughout, I harmlessly use 'SELF-KNOWLEDGE' and the like to denote both an alleged requirement and the proposition that there is such a requirement.

one can rationally have false beliefs about, say, one's immediate surroundings due to unwitting perceptual illusion, or other forms of misleading evidence.[9]

The question of a special relationship is important because it places constraints on our understandings of both self-knowledge and rationality. Such a special relationship would arguably be a major strike against theories of introspection that make self-knowledge out to be too much like ordinary perceptual knowledge. Indeed, this has featured prominently in recent 'rationalist' alternatives to the perceptual model.[10] A special relationship between rationality and self-knowledge also would affect our understanding of rationality. I discuss elsewhere the difficulty of reconciling such a special relationship with evidentialism about rational inference, in contrast to reliabilism and other consequentialist accounts.[11] And others have argued plausibly that self-knowledge requirements are a natural compliment to conceptions of rationality that emphasize critical reflection on one's beliefs.[12]

It matters that SELF-KNOWLEDGE and SELF-KNOWLEDGE- retain the interest of SELF-KNOWLEDGE+, because these weaker theses might be easier to defend. As we will see, Moore's paradox arguably speaks more directly to SELF-KNOWLEDGE and SELF-KNOWLEDGE- than to SELF-KNOWLEDGE+. So it matters whether an opponent of SELF-KNOWLEDGE+ like myself can get away with granting these weaker requirements, and claiming merely that they don't get us all the way to the *bona fide* self-knowledge requirement SELF-KNOWLEDGE+. I think I can't get away with it.

It also matters because some existing objections to self-knowledge requirements seem to turn on particularly strong formulations of them. It is not clear that they attack the *very idea* of a special requirement involving higher-order belief or knowledge. This can give the impression that even the opponents grant that rationality requires some degree of something resembling self-knowledge, and that what's debatable is merely what it is and what degree is required.[13]

[9] Though plausible, the assumption is increasingly controversial. Because I oppose a special relationship between rationality and higher-order belief or knowledge, my opponent is someone who accepts SELF-KNOWLEDGE- for reasons specific to higher-order belief, rather than a general commitment to false beliefs being irrational. Those with the more general commitment might consider some replacement for SELF-KNOWLEDGE- that they deem suitable. For example, even if one denies that false beliefs based on perceptual illusion can be rational, one still might grant that they can have some derivative normative status, like reasonableness or blamelessness. If so, my discussion of SELF-KNOWLEDGE- would need to be recast, but its central concerns would not go away. For there remains the question whether false higher-order beliefs can be reasonable or blameless.

[10] See, e.g., Boyle 2011, Burge 2013, Byrne 2005, Fernández 2013, Moran 2001, Peacocke 1998, Setiya 2011, Shoemaker 1996, Smithies 2012b and forthcoming, and Zimmerman 2004. And see also Gertler 2011 and 2015 for discussion.

[11] Barnett 2016. And see Berker 2013 for a critical review of epistemic consequentialism.

[12] E.g., Burge 2013 and Smithies forthcoming.

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

A paradigmatic example is Williamson's argument for the anti-luminosity thesis that one can believe that p without being in a position to know that one believes that p.[14] Anti-luminosity is difficult to reconcile with SELF-KNOWLEDGE+, since it is unappealing to say you can be required to know something you are in no position to know. But anti-luminosity is compatible with SELF-KNOWLEDGE, since you can have true beliefs about something you are in no position to know. This does not mean that anti-luminosity is unimportant. Indeed, it has great importance for the viability of Williamson's knowledge-first But it does mean that it falls short of denying a special relationship between rationality and something like self-knowledge.

A second objection to SELF-KNOWLEDGE+ is that it is too demanding. Whenever I hold a belief, complying with SELF-KNOWLEDGE+ will mean knowing and thus 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.[15]

Retreating to SELF-KNOWLEDGE will not help with this overdemandingness problem. But going further to SELF-KNOWLEDGE- would, since it requires only the absence of false higher-order beliefs, rather than the presence of true ones. In any case, the question whether retreat is warranted seems like a side issue. Overdemandingness problems arise for requirements having nothing to do with self-knowledge, such as:

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

SINGLE-PREMISE CLOSURE is very appealing. If p entails q, then the truth of p conclusively establishes the truth of q, making it seem incoherent to accept p as true but not q. Even so, SINGLE-PREMISE CLOSURE faces familiar overdemandingness problems of its own, among other reasons for requiring an infinity of beliefs. Now one possible response is to reject any direct relationship between rationality and logic.[16] But to many, this seems like overkill. Perhaps the closure requirement should be qualified, or replaced by something else connecting rationality and logic in a different way. Or perhaps ideal rationality is extremely demanding after all.[17] Proponents of SELF-KNOWLEDGE+ must chose from a similar range of responses, but the very idea of a self-knowledge requirement doesn't hang in the balance.

The same goes for opposition to SELF-KNOWLEDGE+ on the grounds that self-knowledge is 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 cannot know that that I hold this belief, arguably in contrast to SELF-KNOWLEDGE+.[18]

[14] See Williamson 2000 and Silins 2012, and see Smithies forthcoming for a reply that is compatible with an unrestricted SELF-KNOWLEDGE+.

[15] Cf. Shoemaker 1996, and see de Almeida 2007 for related objections to a related claim from Williams 1994.

[16] Harman 1986

[17] Christensen 2004 and Smithies forthcoming.

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

It is debatable whether retreating to SELF-KNOWLEDGE or even SELF-KNOWLEDGE- helps with defeasibility worries. Arguably, I rationally should withhold belief on whether I believe that my mother loves me, in contrast with SELF-KNOWLEDGE. Indeed, I arguably should outright believe that I do not believe my mother loves me, in contrast with SELF-KNOWLEDGE-.

But 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.[19] Or bite the bullet, and claim that ideal rationality does require believing q.[20] Or something else. Again, this is a problem for rational requirements in general, not just SELF-KNOWLEDGE+.

So while I oppose self-knowledge requirements, I doubt existing objections get to the heart of the matter. The real reason to oppose SELF-KNOWLEDGE+ is that it is on the losing side of very broad debates about the natures of rationality and self-knowledge. There is no quick and easy refutation. At the same time, supporters of self-knowledge requirements have claimed a quick and easy argument in their favor, and with it victory in the broader debates. The argument 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.[21] It is also a common motivation for views that take critical reflection to be central to rationality.[22] I have not seen an extended attack on evidentialism or defense of epistemic consequentialism drawing on Moore's paradox, though a tension between vaguely Moorean views and evidentialism is widely acknowledged.[23] It also is common in advancing controversial requirements of rationality on broadly consequentialist grounds to appeal to SELF-KNOWLEDGE+ as a largely undefended premise. Often a passing nod to Moore's paradox is included.[24]

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

[19] Cf. Lasonen-Aarnio 2014.

[20] Smithies forthcoming.

[21] 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.

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

[23] E.g., Byrne's (2018, Ch. 4 and Sec. 5.2.4) and Gallois' (1996, pp. 52-53) discussion of what Gallois (pp. 46-47) calls 'Moore inferences'.

[24] 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.

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.

3. Moore's Paradox

G. E. Moore observed 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 things for 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>.[25]

Like other requirements we have considered, NO MOOREAN ASSERTIONS faces defeasibility and overdemandingness problems. And perhaps unlike the others, it is not plausibly 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. One can rationally 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.

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. Some instances of this phenomenon, like those discussed in Section 5, are controversial. But here is a relatively uncontroversial 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 typical 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 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.[26] (All these accounts are usually

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

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

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.[27] 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 consider their implications for self-knowledge requirements.

There is an obvious argument that if we accept the alleged datum NO MOOREAN BELIEFS, we also should accept at least SELF-KNOWLEDGE-, 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.

Here is the argument. 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 SELF-KNOWLEDGE- holds.

Should we accept MULTI-PREMISE CLOSURE? It has *prima facie* motivation, but also some familiar problems. If you believe some argument's premises and yet withhold belief from the conclusion, then it seems you are leaving room open for the possibility that the premises are true while the conclusion is false. But if you are rational, it could be claimed, you will not do that if the argument is valid.

Of course, MULTI-PREMISE CLOSURE faces the usual overdemandingness and defeasibility problems. And it also faces an additional problem involving the accumulation of riSELF-KNOWLEDGE of error over large numbers premises.[28] If a book contains enough individual claims, then you arguably can rationally believe each claim while doubting that *all* the claims are true. But I do not want to hang my opposition to SELF-KNOWLEDGE- on these familiar

[27] 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 forthcoming; and Williams 2006 and 2007.

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

problems, whose relevance to Moore's paradox is questionable. The argument for SELF-KNOWLEDGE- invokes MULTI-PREMISE CLOSURE to rule out the possibility that one might rationally not believe a Moorean conjunction while believing each conjunct. While a Moorean conjunction can be riskier than each conjunct individually, it cannot be by enough to matter except in cases where one's beliefs in the conjuncts are borderline to begin with. Maybe the proponent of SELF-KNOWLEDGE- could allow for principled exceptions in such cases. For example, maybe one can rationally mistake a first-order credence just above the threshold for belief for one just below it.

If SELF-KNOWLEDGE- is accepted, what does this mean for the *bona fide* self-knowledge requirement SELF-KNOWLEDGE+? A proponent of SELF-KNOWLEDGE+ might suggest that it is supported by SELF-KNOWLEDGE- via inference to the best explanation. While I think this suggestion has promise, I will leave it to my opponents to develop it. Since I oppose any special relationship between rationality and self-knowledge or higher-order belief, I oppose SELF-KNOWLEDGE- on its own, regardless of whether it is a stepping stone to SELF-KNOWLEDGE+.

But should NO MOOREAN BELIEFS itself be accepted? In Section 5, 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.

4. First Route to Self-Knowledge Requirements: Banning Foreseeable Errors

Many claim that NO MOOREAN BELIEFS is an obvious datum, which should serve as a starting point for any plausible account of Moore's paradox. I don't buy it. At best, what's an obvious datum is that Moorean beliefs are guaranteed to be in error. 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.[29] Maybe this means that Moorean beliefs are guaranteed errors. And maybe this even counts as a datum. 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 knowledge-first epistemologists accept this, because they accept:

> (ONLY KNOWLEDGE) Rationality requires that you believe that q only if you know that q.[30]

[29] 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.

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

But I want to set aside defenses of NO MOOREAN BELIEFS appealing to ONLY KNOWLEDGE. First, OK is too controversial to ground the claim that NO MOOREAN BELIEFS is an obvious datum. It entails, for example, that brains in vats and the victims of misleading inductive evidence are irrational. Second, if we accepted OK, there would be no need to trifle with NO MOOREAN BELIEFS. For SELF-KNOWLEDGE- would follow immediately if false beliefs can never be rational. As I said in footnote 9, my interest here is in defenses of SELF-KNOWLEDGE- that uphold a special relationship between rationality and higher-order belief, rather than a general ban on false beliefs.

Perhaps there is another way to support NO MOOREAN BELIEFS, not by the claim that Moorean beliefs are guaranteed errors, but instead by the further claim that this is obvious. If it 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.[31]

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 SELF-KNOWLEDGE-. 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 SELF-KNOWLEDGE- and its cohort that are relevant to matters of broader philosophical concern. If Moorean beliefs are irrational only because they violate NO FORESEEABLE ERRORS, then arguably they are little more than idle curiosities. Much of the recent interest in Moore's paradox is driven by the assumption that

[31] 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.

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)**.[32] 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.[33] And the analogy makes appearances in recent discussions from Alan Hájek (2007, pg. 219), Richard Moran (2001, pg. 70), Antonia Peacocke (2017), Nico Silins (2012), Declan Smithies (2016 and forthcoming), 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

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

[33] 1996, pp. 78-79.

beliefs.[34] 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 further 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 reject 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. It will take some work to get there, but I will give an informal recap at the end of this section.

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 5. 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, where this is defined as follows:

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

Here the *K*s are **dependence hypotheses**—i.e., maximal hypotheses about how outcomes depend causally on one's actions that form a partition. The agent's probability function, *Pr*,

[34] 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.

can be understood as representing the degree to which the agent's evidence supports various propositions she might entertain. The agent's value function, *v*, is a little tricky. It is normally understood to represent the overall degree to which she values various states of affairs, balancing her various conflicting aims. But the action conception is a theory of rational judgment, which might be taken to depend solely on the agent's **alethic aims** of initiating true beliefs and avoiding error. Otherwise, the action conception must say that it is rational to make a judgment because it will further one's aim to be happy, for example. So we can here take *v* to be the agent's to be her alethic value function, which represents solely her alethic aims.

One further wrinkle concerns the Jamesian distinction between the competing alethic aims 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.[35] So I will assume only concern with the truth of one's beliefs.

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

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,

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

But (6) is false for certain values of q that include Moorean conjunctions. 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.[36] 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.

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 SELF-KNOWLEDGE-.

A familiar theorem of the probability calculus is that if p entails q, then $\Pr(p) \leq \Pr(q)$. 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 $\Pr[r] \approx \Pr[B(r)] \approx 1$, and therefore that $\Pr[r \mathbin{\&} B(r)] \approx 1$.

[36] 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 $\Pr[J(q) \Rightarrow q]$, 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 $\Pr[q|J(q)] = \Pr[q]$.

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.[37] 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? Simple evidentialism takes the rationality of believing p to depend on the evidential probability of p itself. Informally, it takes the object of doxastic deliberation to be the question *whether p*, and reasons for belief to be considerations bearing on the truth of p itself. But the action conception instead takes the object of doxastic deliberation to be the question *whether to judge that p*, and reasons for belief or judgment to be considerations bearing instead on whether p would be true if one judged that p. In special cases like Stubborn Stella, Unbelievable Consequences, and Unthinkable Consequences, one's evidence can support a self-defeating proposition without supporting that it would be true if one were to judge it to be true. This is why the action conception upholds No Foreseeable Errors, even though simple evidentialism opposes it. At the same time, it means that the action conception opposes Multi-Premise Closure in these same cases, at least if rationality's requirements are consistent. One's evidence can supports that a first conjunct would be true if judged true, and that a second would be true if judged true, but not that the conjunction would be true if judged true. Thus evidentialism and the action conception can agree that the conjuncts can be jointly believed, even though they disagree on whether the conjunction can be. In short, if we say Moorean beliefs are obviously irrational, just because they are foreseeably false, this undermines the relevance of Moore's paradox to self-knowledge requirements.

5. Second Route to Self-Knowledge Requirements: 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

[37] 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.

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- via MULTI-PREMISE CLOSURE, 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 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.[38] 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 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

[38] 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.

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.[39]

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 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 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

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

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 4, 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.[40] This marks an important difference between endorsement and judgment. We assumed in Section 4 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.

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.

[40] Here the reliance on CDT rather than EDT is essential. For plausibly, $Cr([p \& {\sim}B(p)] \mid E[p \& {\sim}B(p)]) \ll Cr[p \& {\sim}B(p)]$.

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 different way, by comparing:

> **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 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 should endorse it. But if Ira will not give up his irrational belief in a Moorean proposition, then he still should not 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 broader pattern that has been widely discussed by decision theorists, in which an agent's actions themselves amount to evidence about what their effects will be.[41] Mundane failures of self-knowledge are enough to generate cases like:

> **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.[42] 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 be sure 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 will not 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.

[41] E.g., Egan 2007, Harper 1986, Reed 1984, Skyrms 1990, Weirich 1985.

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

Epistemicism is not looking so good anymore. But what is the alternative? Answering this question is a major undertaking, since as Andy Egan (2007) argues, cases like these are apparent counterexamples to CDT. A satisfying account must arguably 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:

$$(14)\ U(B \mid A) = \sum_K Cr(K \mid A) v(KB).$$

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

But what is the relationship between ratifiablity and rational action? Here is one proposal:[43]

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

Unfortunately, ABSOLUTE RATIFIABILITY is false as it stands. Consider George 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

[43] Harper (1986, pg. 33) endorses ABSOLUTE RATIFIABILITY, while Egan ultimately rejects it.

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):

> (GRADED RATIFIABILITY) Rationality requires that if option B is more ratifiable that option A, then one prefers B to A.[44]

When A and B are one's only options, preferring B to A licenses adopting B. So if endorsing a Moorean proposition is less ratifiable than refraining, then by GRADED 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 GRADED 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 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.

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.[45]

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.

6. Third Route to Self-Knowledge Requirements: Shoemaker's Reductio

A final way of motivating self-knowledge requirements via Moore's paradox comes from Sydney Shoemaker, who argues from NO MOOREAN ASSERTIONS to:

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

[45] 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.

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

There is an obvious affinity between SHOEMAKER'S THESIS and self-knowledge requirements. Since a self-blind agent is by stipulation rational despite lacking knowledge of her beliefs, it is plausible that SELF-KNOWLEDGE+ entails SHOEMAKER'S THESIS. The converse is less obvious, as I will explain later on. But I won't rest my opposition to self-knowledge requirements on subtle differences between them and SHOEMAKER'S THESIS. Instead, I hope to show that Shoemaker's argument unsound, considered on its own terms.

So how does Shoemaker argue for SHOEMAKER'S THESIS? His strategy is to assume for the sake of *reductio* that self-blindness is indeed possible, and then to argue that if so, a self-blind agent would evince no sign of self-blindness in her behavior. Shoemaker regards this consequence as absurd, and rejects the assumption that self-blindness is possible.

So where a **self-aware** agent is a rational agent who has introspective knowledge of her beliefs, Shoemaker's argument appeals to the following lemma:

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

This is the lemma that Shoemaker hopes to support via Moore's paradox, and it will be my main focus. 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 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:

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

But RESTRICTED BEHAVIORISM seems questionable, for two reasons. The first is that self-awareness requires *knowledge* of one's beliefs, but merely having the appropriate higher-order beliefs would generate the same behavior. Thus Shoemaker's argument at best shows that it is impossible to be rational and yet lack true higher-order beliefs. This will be of interest to anyone worried about the distinction between the *bona fide* self-knowledge requirement SELF-KNOWLEDGE+ and the weaker SELF-KNOWLEDGE, but since I also oppose SELF-KNOWLEDGE, I won't harp on the point. To my mind the more serious problem is that 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 in their (nonfactive) mental states without differing in their behavioral dispositions. Since this general behaviorism is implausible, I doubt Shoemaker can give us a well-motivated deductive argument for SHOEMAKER'S THESIS.[46]

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

[46] For related discussion, see Kind 2003.

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 has problems, too. By Shoemaker's admission, his argument for BEHAVIORAL INDISTINGUISHABILITY is highly idealized.[47] Even if successful, it shows 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.[48] 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 simple 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.[49]

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 is dangerous to speculate about their evolutionary origins without a better understanding of how they work. What can responsibly be done from the 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.

So why does Shoemaker accept this extreme claim, BEHAVIORAL INDISTINGUISHABILITY? His argument is developed in papers spanning several decades, and resists easy summary. But the basic outline is:

> (NO MOOREAN ASSERTIONS) Rationality requires one not to assert propositions of the form <p, but I don't believe that p>.
>
> (PROXY) If rationality requires one not to assert propositions of the form <p, but I don't believe that p>, then rationality requires acting like a self-aware agent.

[47] E.g., Shoemaker 2009, pp. 38-39.

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

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

> (CONFORMITY) Necessarily, any self-blind agent would conform to rationality's requirements.
>
> Therefore, (BEHAVIORAL INDISTINGUISHABILITY) Necessarily, any self-blind agent would act like a self-aware agent.

NO MOOREAN ASSERTIONS is a datum, and CONFORMITY is trivial given the definition of self-blindness. The crucial premise is PROXY. Why accept it? The rough idea is that Moorean assertions are a good general proxy for other actions that might evince an agent's self-blindness. If rationality alone enables a self-blind agent to "appreciate the logical impropriety of affirming something while denying that one believes it," then it also will enable her to "give appropriate answers to questions about what she believes," and more generally to act self-aware.[50]

In contrast, I think Moorean assertions are idiosyncratic. First, they are conjunctive, evincing both first-order and higher-order belief in a single action, which gives them their distinctive self-defeating flavor. What goes for them might not go for other assertions about one's beliefs that lack this Moorean flavor. Second, Moorean assertions are assertions. What goes for them might not go for other actions evincing self-blindness.

Shoemaker in effect addresses the first idiosyncrasy, arguing that just as self-blind agents recognize the impropriety of asserting 'It will rain, but I don't believe it will rain', they should recognize the impropriety of separately asserting 'It will rain' and 'I don't believe it will rain'. If so, self-blind agents will aim to coordinate their first-order and higher-order assertions. But even if this is granted, it falls short of what Shoemaker needs to support PROXY. Aiming to coordinate doesn't entail success. Consider:

> **Permanent Marker:** Self-blind George is taking a true or false exam in permanent marker. The first statement is 'I, George, do not believe (right now) that it will rain.' George must decide now whether to endorse the statement, and he cannot change his answer later. Later on, he will encounter the statement '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.

A self-aware agent in George's situation would believe it will rain, and, knowing this, refrain from endorsing the first statement. Will George? His behavioral evidence supports the statement's truth, and he lacks introspective knowledge to the contrary. So he will endorse, unlike a self-aware agent. To be sure, George might aim to coordinate first-order and higher-order endorsements, and he surely will in fact endorse the first-order statement 'It will rain' later. But that does not mean he will succeed at coordinating by refraining from the higher-order endorsement now. Since George falsely believes that he does not believe it will rain, he believes that he would not endorse that it will rain. So when George makes the higher-order endorsement, he will falsely believe that he *is* coordinating.

In Section 5, I claimed that in George's Exam, George would refrain from endorsing a Moorean conjunction, despite his self-blindness. Why is Permanent Marker different? It is because one's endorsing a Moorean conjunction is self-defeating in a way that endorsing

[50] 1996, pg. 237.

either conjunct alone is not. The act of endorsing a Moorean conjunction amounts to evidence that the conjunction is false, making endorsing unratifiable. But endorsing that it will rain is no evidence that it won't rain, and endorsing that one lacks the belief it will rain is no evidence that one has the belief. This is where Moorean assertions get their self-defeating flavor, which piecemeal assertions of their conjuncts lack.

Now I don't take this objection to be decisive, since it leans on the ratificationist account of Moore's paradox, which Shoemaker could reject. But the second idiosyncrasy of Moorean assertions—their being assertions—is harder to dismiss. To make the point independent of the first idiosyncrasy, I focus on a special class of actions I call **Moorean actions**, which are actions known to pay off just in case some Moorean proposition is true. I think Moorean assertions aren't just a poor general proxy for actions evincing self-blindness, but a poor proxy even for this special class of actions devised to resemble them.

I argued in Section 5 on ratificationist grounds that some apparent idiosyncrasies of assertions are not essential to explaining NO MOOREAN ASSERTIONS. The right account must cover the broader category of Moorean endorsements, such as in George's Exam. On this much, ratificationism agrees with PROXY. But to vindicate PROXY, we cannot stop there. Moorean endorsements will have to be a good proxy for other Moorean actions.

By stipulation, two features distinguish Moorean endorsements from other Moorean actions. First, for a Moorean action to count as an endorsement, it must involve an explicit articulation of the relevant Moorean conjunction. Second, it must have a certain risk profile; the ratio of disvalue of false endorsement to the value of true endorsement has to be quite high, as it is for ordinary assertion. This is why one's endorsing a proposition amount to strong evidence that one believes it.

I agree that Moorean endorsements can be good proxies for Moorean actions lacking the first feature, 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.

Taking the drug is a Moorean action, since George knows it will be to his overall benefit iff he has an ingrown toenail but does not believe that he does. And taking the drug is arguably irrational. 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 because unratifiable. 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 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 the PROXY.

But Moorean endorsements are not good proxies for Moorean actions lacking the second feature, such as:

> **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 such 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 Moorean umbrella that only opens if the agent did not believe it would rain at the time of rental.

Renting the umbrella is a Moorean action because George knows it will pay off just in case it rains but he does not believe that it will rain. But does that make renting it irrational? A self-aware agent in George's situation should not rent the umbrella. Since she will believe it will rain, she will know introspectively that she does, and so will know the umbrella is 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.

The same goes for actions that share only the first feature with Moorean endorsements, such as:

> **Alternate Scoring:** George is again taking a true or false exam. But this time, the penalty for incorrectly marking a statement as true is no greater than the bonus for correctly doing so. The first statement on the exam is 'It will rain, but I, George, do not believe it will rain.' As usual, George's meteorological evidence supports that it will rain, but his behavioral evidence supports that he does not believe that it will rain.

Here George expressly considers a Moorean conjunction. Assuming he is rational but self-blind, he will mark it as true. This is in contrast to a self-aware agent in the same situation. Since she will believe it will rain, and know that she does, she should not mark the statement as true.

The upshot is that we should reject PROXY. Even under ratificationism, which is friendlier to PROXY than competitors like CDT are, the irrationality of Moorean assertions depends on their idiosyncratic risk profile. This makes Moorean assertions not just bad proxies for some actions, but even for actions that are contrived to closely resemble them, like marking a Moorean conjunction in Alternate Scoring, or renting a Moorean umbrella in Umbrella Rental. And if PROXY is false, then self-blind agents will not act like self-aware ones in general, even though they resemble them in avoiding Moorean conjunctions. So BEHAVIORAL INDISTINGUISHABILITY is false, and Shoemaker's argument for SHOEMAKER'S THESIS unsound.

Some philosophers might respond that it does not matter whether Shoemaker's own argument is sound, since SHOEMAKER'S THESIS is independently motivated. For example, some might suggest the impossibility of imagining *being* self-blind as evidence of its impossibility.[51] Others might appeal to independent commitments that, e.g., some mental states are necessarily self-intimating.[52]

Now I am skeptical that these are sufficient grounds for SHOEMAKER'S THESIS. First, many *actual* psychological conditions are hard to imagine, showing imaginability to be a poor guide to possibility here. And while necessary self-intimation has some plausibility for phenomenal states, I think it's less plausible for beliefs. But the more important point is that if we accept SHOEMAKER'S THESIS for *these* reasons, it is 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. To support SELF-KNOWLEDGE+ or the like, we need a motivation for SHOEMAKER'S THESIS that trades on a self-blind agent's rationality in a way that it doesn't on, say, his moral character. Shoemaker's argument did that, which is why it matters that it fails.

7. 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 whether self-knowledge is proposed as a basic requirement, or as a consequence of familiar requirements for consistency and the like. On the former view, a self-blind agent might satisfy the familiar requirements, but still be irrational in violating a further basic requirement for self-knowledge. On the latter, an agent cannot even be rational in familiar, uncontroversial ways without satisfying self-knowledge requirements.[53]

I think we have seen good reason to reject the latter view. In discussing Shoemaker's *reductio* we examined the behavior of a hypothetical agent who satisfied 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. This alone does not show that self-blindness is possible, or that rationality fails to require self-knowledge. But it does support 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 motivation. And the main candidate source of motivation, Moore's paradox, doesn't seem to provide any.[54]

[51] Cf. Stoljar 2018, Sec. 4.

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

[53] Shoemaker arguably endorses this view. See, e.g., 1996, pp. 32-33.

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

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.

Harper, William L. (1986) 'Mixed Strategies and Ratifiability in Causal Decision Theory' *Erkenntnis* 24: 25-36.

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, Antonia (2017) 'Embedded Mental Action in Self-Attribution of Belief' *Philosophical Studies* 174: 353-377.

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.

Richter, Reed (1984) 'Rationality Revisited' *Australasian Journal of Philosophy* 62(4): 392-403.

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.

Skyrms, Brian (1990) 'Ratifiability and the Logic of Decision' *Midwest Studies in Philosophy* 15: 44-56.

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.

——— (forthcoming) *The Epistemic Role of Consciousness*. OUP.

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

Stoljar, Daniel (2018) 'Introspection and Necessity' *Noûs* 52(2): 389-410.

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

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

Weirich, Paul (1985) 'Decision Instability' *Australasian Journal of Philosophy* 63(4): 465-472.

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.