David James Barnett

published version in *Philosophical Review* 130(2)

Self-Knowledge Requirements and Moore's Paradox[1]

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 considers a few ways the data surrounding Moore's paradox might be marshaled to support rational requirements to know one's beliefs, and finds that none succeed.

1. Introduction

To hear philosophers tell it, rationality requires a lot of us. It requires us to have logically consistent beliefs,[2] or else probabilistically coherent credences.[3] It requires us to believe the obvious deductive consequences of our existing beliefs, at least if we entertain them.[4] It requires us to intend to do the things that we believe are necessary means for achieving our ends.[5] It requires us not to believe things that we believe we shouldn't believe,[6] and not to intend things that we believe we shouldn't do.[7] It requires us to have preferences that make us representable as having a utility function, and to act in ways that maximize expected utility.[8] And so on. Now I'm not sure myself about everything on this list, and perhaps you're not either. But as 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 accurate 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.

[1] I am very grateful to Paul Boghossian, Brian Cutter, Brendan de Kenessey, Sinan Dogramaci, David Hunter, Uriah Kriegel, Harvey Lederman, Mohan Matthen, Jennifer Nagel, Elliot Paul, Jim Pryor, Timothy Rosenkoetter, Miriam Schoenfield, Nico Silins, Declan Smithies, Jonathan Weisberg, the editors and two anonymous referees at this journal, and audiences at the New York University Epistemology Reading Group, the Self-Knowledge and Belief Workshop at Ryerson University, and the Northern New England Philosophical Association.

[2] Cf. Broome 2013, 9.2 and Way forthcoming.

[3] Christensen 2004.

[4] Cf. Broome 2013, 9.3 and Way forthcoming.

[5] Cf. Broome 2013, 9.4; Kolodny 2005; and Way forthcoming.

[6] Greco 2014, Horowitz 2014, and Smithies 2019.

[7] E.g., Broome 2013, 9.5; Kolodny 2005; Wedgwood 2007, Ch 1; and Way forthcoming.

[8] See Buchak forthcoming for review.

2. Self-Knowledge Requirements

Whether self-knowledge is rationally required is of central importance to how we should understand its nature. In particular, it would be a major strike against theories that make introspective self-knowledge out to be too much like ordinary perceptual knowledge, as Sydney Shoemaker has argued. If lacking self-knowledge amounts to a violation of rationality's requirements, this supports Shoemaker's famous contention that it is impossible to suffer from **self-blindness**, the condition of lacking a capacity for introspective knowledge despite possessing idealized rationality, intelligence, and conceptual sophistication. And if so, this is hard to reconcile with broadly perceptual accounts of introspection.[9] An agent can be ideally rational and yet be ignorant of various facts about her surroundings, due for example to perceptual limitations like ordinary blindness. If that is impossible for one's own beliefs, it would seem our way of knowing them must be quite different from ordinary perception. Indeed, this is a central rationale behind recent **rationalist** alternatives to the perceptual model.[10]

A rational requirement for self-knowledge also would affect our understanding of rationality. I discuss elsewhere the difficulty of reconciling a self-knowledge requirement with evidentialism about rational inference, in contrast to reliabilism and other consequentialist accounts.[11] And others have argued that self-knowledge requirements are a natural complement to conceptions of rationality that emphasize critical reflection on one's beliefs.[12] Still, my hope is to stay more neutral on many recent controversies about the general form and normative significance of rational requirements.[13] Are they wide or narrow scope? Synchronic or diachronic? Strict or permissive? More or less fundamental than claims about what reasons, or about ideally rational agents? While these questions are important, my concern is the relationship between self-knowledge and rationality, not how rationality in general ought to be understood. For concreteness, I will pitch things in terms of synchronic wide-scope requirements, though my hope is that not much turns on this. In a few places where something might, I'll mention it.

An immediate question about such requirements 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

[9] Throughout, I use 'introspection' broadly to include any distinctive method we have for knowing our beliefs, including alleged methods that don't involve 'looking inward'.

[10] See, e.g., Boyle 2011, Burge 2013, Byrne 2005 and 2018, Fernández 2013, Moran 2001, Peacocke 1998, Setiya 2011, Shoemaker 1996, Smithies 2012b and 2019, 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 2019. And see Barnett MSa for discussion.

[13] See Way forthcoming for review.

mental states might raise additional difficulties for my opponents, but I won't worry about them.[14]

Limiting the scope to beliefs leaves open further questions, including whether it really should be *knowledg*e that is required, as opposed to something less demanding. But again, my hope is that the details won't matter too much. I will argue that Moore's paradox fails to motivate a wide range of potential requirements, including for instance:

> (SELF-KNOWLEDGE) Rationality requires that if one believes *p*, then one knows that one believes *p*.
>
> (HIGHER-ORDER BELIEFS) Rationality requires that if one believes *p*, then one believes that one believes *p*.
>
> (NO HIGHER-ORDER ERRORS) Rationality requires that if one believes *p*, then one does not believe that one does not believe *p*.

I refer to these and related requirements collectively as self-knowledge requirements, though only SELF-KNOWLEDGE is a *bona fide* requirement for knowledge.[15] I group them together because I think even the weaker requirements retain much of the theoretical interest of SELF-KNOWLEDGE. Like it, HIGHER-ORDER BELIEFS and NO HIGHER-ORDER ERRORS make self-knowledge and higher-order belief out to be very different from knowledge and belief about other deeply contingent matters. Just as one can rationally fail to know facts about one's surroundings due for example to ordinary blindness, one can rationally fail to have true beliefs about them. According to HIGHER-ORDER BELIEFS, things are different regarding one's beliefs. If you in fact believe that it will rain, and you do what rationality requires, then you will have the true belief that you believe it. While NO HIGHER-ORDER ERRORS takes us a step further from requiring self-knowledge, it still preserves a special rational significance for higher-order beliefs. Deceived agents like brains in vats plausibly can have false but rational beliefs about most any contingent matters of fact. But NO HIGHER-ORDER ERRORS says this is impossible when it comes to your own beliefs.

This last point could be denied by some objectivist epistemologists, who might accept:

> (NO ERRORS) Rationality requires that one believe that *p* only if *p*.[16]

Under NO ERRORS, brains in vats and other deceived agents hold their false beliefs irrationally, and so would any agent deceived about her beliefs. But I don't want to pick a fight with someone who thinks false higher-order beliefs are irrational solely because all false beliefs are irrational. What I want to know is whether there is a distinctive requirement involving self-knowledge or higher-order belief, something that belongs in the same category

[14] For difficulties, see Adler and Armour-Garb 2007. And see Byrne 2018 for proposals that could be adapted by the Moorean.

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

[16] C.f. Gibbons 2013, Chs. 4-5; Littlejohn forthcoming; and Williamson 2017 and forthcoming, though of these only Williamson 2017 pitches the view as one about which beliefs are *rational*.

as conventional non-truth-involving requirements like those glossed in Section 1. Perhaps on some objectivist views, there is nothing positive to say about the false beliefs of a deceived agent who follows these subjectivist requirements. If so, maybe the whole category of non-truth-involving requirements needs to be tossed out, in favor of objectivist requirements like NO ERRORS. I won't have much to say here to that view. But a moderate objectivist might allow the deceived agent's beliefs to qualify as rational (or "reasonable") in some derivative sense, despite violating more fundamental objective requirements like NO ERRORS.[17] If so, my discussion here might need to be recast, but its animating concerns would not go away. For there remains the question whether there could be an agent whose violations of NO HIGHER-ORDER ERRORS are reasonable, in whatever sense a brain in a vat's false beliefs about its environment can be reasonable.

Part of why I want to remain noncommittal about how to formulate self-knowledge requirements is to avoid the pitfall of having my opposition turn on needlessly strong formulations. I worry that some common objections to self-knowledge requirements are guilty of this. Even if successful, it is not clear that they undermine 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.[18]

One such objection, raised in different contexts by Alex Byrne, Quassim Cassam, and others, is that a requirement like SELF-KNOWLEDGE is too demanding.[19] Whenever I hold a belief, complying with SELF-KNOWLEDGE will mean knowing and thus holding the further belief that I hold it. But then I will be required to hold a belief that I hold that further belief, and so on. This might be claimed to be metaphysically impossible, or at least psychologically unrealistic.

Retreating to HIGHER-ORDER BELIEFS will not help, since it too imposes a 'positive' requirement to have the right higher-order beliefs. But going further to NO HIGHER-ORDER ERRORS would, since it imposes only the 'negative' requirement not to have the wrong ones. Maybe this is a reason for retreat, or maybe not. But in any case, the question whether retreat is warranted seems like a side issue. For similar problems involving the application of idealized standards to non-ideal agents arise for requirements unrelated to 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*.

To some of us, SINGLE PREMISE CLOSURE has the ring of truth.[20] 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. But as it stands SINGLE PREMISE CLOSURE is arguably too demanding. Among other things, it resembles SELF-KNOWLEDGE in requiring an infinity of beliefs.

[17] Cf. Lasonen-Aarnio 2010 and Williamson 2017 and forthcoming.

[18] E.g., Christensen 2007a and de Almeida 2007.

[19] Byrne 2018, pp. 7-8 and Cassam 2014, pp. 42-43. See also Gertler 2011, pp. 61-64 for review, and de Almeida 2007 and Sorensen 1988, Ch. 1 for related discussion.

[20] E.g., Christensen 2004 and Way forthcoming.

Now one possible response is to reject SINGLE PREMISE CLOSURE altogether, and deny any direct relationship between rationality and logic.[21] But to many, this seems like overkill. Perhaps SINGLE PREMISE CLOSURE should be somehow restricted, or replaced by something else connecting rationality and logic in a different way.[22] Or perhaps ideal rationality is extremely demanding after all.[23] Proponents of SELF-KNOWLEDGE must choose from a similar menu of responses, but the very idea of a self-knowledge requirement doesn't hang in the balance.[24]

The same goes for opposition to SELF-KNOWLEDGE on the grounds that self-knowledge is defeasible, as stressed for example by David Christensen and Brie Gertler.[25] Suppose I believe that my mother loves me, but there are psychologists lining up to tell me that I don't really believe this. It could be claimed that if so, I cannot know that I hold this belief, arguably in contrast to SELF-KNOWLEDGE.

It is debatable whether retreating to HIGHER-ORDER BELIEFS or even NO HIGHER-ORDER ERRORS helps with defeasibility worries. Following Christensen, we might say I rationally should withhold belief on whether I believe that my mother loves me, in contrast with HIGHER-ORDER BELIEFS.[26] Perhaps I even should outright believe that I do not believe my mother loves me. If so, then retreat to NO HIGHER-ORDER ERRORS might not go far enough. Perhaps a further retreat is warranted, to something with a 'no defeaters' condition.

But again, defeasibility worries are not idiosyncratic to self-knowledge requirements. They can be raised for many alleged requirements, including SINGLE PREMISE CLOSURE. If I rationally believe p, and have logicians lining up to tell me (falsely) that p doesn't entail q, then intuitively I shouldn't believe q.[27] Again, we might claim that SINGLE PREMISE CLOSURE needs restriction to cases where one lacks defeaters. Or claim that one can be faced with conflicting requirements.[28] Or bite the bullet, and claim that ideal rationality does require believing q.[29] Or something else. Again, this is a problem for rational requirements in general, not a distinctive problem about self-knowledge.

[21] Harman 1986.

[22] One common restriction, also available for self-knowledge requirements, is to cases where one considers the relevant proposition.

[23] Christensen 2004 and Smithies 2019.

[24] For some attempts, see Shoemaker 1996 and Smithies 2012b, 2016, and 2019, Ch. 4.

[25] Christensen 2007a, Sec. 4 and Gertler 2011, pg. 62. As Gertler emphasizes, the defeasibility of self-knowledge is nowadays commonly acknowledged. See, e.g., Byrne 2018, Ch. 7 and Cassam 2014, pp. 42-43, and for criticism Smithies 2016 and 2019 and Sorensen 1988, pp. 390-392.

[26] 2007a, Sec. 4.

[27] Cf. Lasonen-Aarnio 2008 and Schechter 2013.

[28] Cf. Lasonen-Aarnio 2014.

[29] Smithies 2019.

A final objection to SELF-KNOWLEDGE adapts Williamson's argument that no nontrivial conditions are luminous. If sound, this argument would show that one can believe that p without being able to know that one believes that p.[30] This might seem to spell trouble for self-knowledge requirements. On a controversial but plausible ought-implies-can principle, you cannot be rationally required to know something you are unable to know.[31] So if you cannot know you have a belief, you cannot be required to know it, contradicting SELF-KNOWLEDGE.

While I oppose self-knowledge requirements, I think their defenders have responses to the anti-luminosity argument that are plausible by their lights.[32] But we do not need to rehash this debate here, or even get into the substance of Williamson's argument. For even if the argument gives us reason to reject SELF-KNOWLEDGE, it still leaves open the broader question of special requirements involving higher-order beliefs. For example, since you can believe something without knowing it, the anti-luminosity argument does not impugn HIGHER-ORDER BELIEFS via an ought-implies-can principle.

Perhaps HIGHER-ORDER BELIEFS could be impugned in another way, for example by an objectivist premise that you are never required to believe what you cannot know. But this premise, like NO ERRORS above, is more controversial than anything the luminosity argument is usually advertised as assuming. Suppose you have strong but misleading evidence. You are plausibly rationally required to believe what your evidence supports, but your belief will fail to be knowledge because it is false. And if we think there is a reliability condition on knowledge, this yields other plausible cases of requirements to believe what you are in no position to know. Suppose you see a real barn while unknowingly traveling through fake barn county. Given your perceptual evidence you might be rationally required to believe that it is a barn, but your belief arguably is not knowledge, because of unreliability. Cases of this latter kind are especially germane to Williamson's anti-luminosity argument. The failures of luminosity that he alleges involve higher-order beliefs that, while true, are not knowledge because they violate a reliability condition. Even if such cases are possible, they need not be cases where true higher-order beliefs are not rationally required. They might be counterexamples to SELF-KNOWLEDGE, but not to HIGHER-ORDER BELIEFS.

Despite my opposition to self-knowledge requirements, I doubt that these or any other quick and easy refutation of them succeeds. Any refutation will involve complicated engagement with very broad debates about the nature of rationality and self-knowledge. In contrast, some supporters of self-knowledge requirements claim a simple and decisive argument in their favor. The argument is supposed to stem from Moore's paradox.

[30] Williamson 2000, Ch. 4. See also Byrne 2019, pg. 38 and Silins 2012.

[31] As an anonymous referee points out, the Williamsonian might have trouble accepting this ought-implies-can principle. If nothing is luminous, then we cannot always know what rationality requires of us. So the Williamsonian can accept ought-implies-can only on the assumption that we can do what rationality requires without knowing it. I myself accept this assumption, and think the Williamsonian should, too. But see Barnett forthcoming for some related objections to Williamson's broader views.

[32] See especially Smithies 2019, Ch. 10.

Moore's paradox is often presented as a motivation for rationalism about introspection, in some discussions the primary motivation.[33] It is also a common motivation for views that take critical reflection to be central to rationality.[34] Sometimes Moore's paradox is cited in defense of broadly consequentialist thinking about rationality,[35] and a tension between vaguely Moorean views and evidentialism is widely acknowledged.[36] 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.[37]

So Moore's paradox is arguably the leading quick and easy motivation for self-knowledge requirements if anything is. Even for theorists who simply find these requirements intuitively attractive, I wonder if Moore's paradox might go some way towards bringing their attraction into sharper focus. It is often said that even though Moorean propositions are logically consistent, there is some subtler kind of conflict present in one's asserting them. Proponents of self-knowledge requirements might similarly think that, even though an agent who violates them can remain consistent in her beliefs, there is some subtler sense in which her doxastic states are in conflict. So I hope that my discussion of Moore's Paradox might still somehow connect with their thinking, albeit indirectly.

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*, such that:

> (NO MOOREAN ASSERTIONS) Rationality requires that one not assert propositions of the form <*p*, but I don't believe that *p*>.[38]

Like other requirements we have considered, NO MOOREAN ASSERTIONS is defeasible. 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 is also questionable as a fully general requirement. One arguably can rationally assert a Moorean proposition if ordered to at gunpoint, after all. What is less open to question is that asserting a Moorean proposition is

[33] See, e.g., Fernández 2005 and 2013; Moran 2001, pp. 69-77; Shoemaker 1996; Silins 2012 and 2013; Smithies 2012b, 2016, and 2019; and Zimmerman 2008, Sec. III.

[34] See, e.g., Gibbons 2013, Smithies 2012a and 2019, and Shoemaker 1996.

[35] Wedgwood 2017, pp. 44-46.

[36] E.g., Barnett 2016 and MSb, and 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'.

[37] E.g., Douven 2009, pp. 367-68; Egan and Elga 2005, pg. 83; Huemer 2011; van Fraassen 1984, pg. 247 and 1995, pg. 19.

[38] Cf. Chan 2008 and Fernández 2005, pg. 534. Note also that NO MOOREAN ASSERTIONS concerns an "omissive" Moorean conjunction, in contrast with the "commissive" <*p*, but I believe that not-*p*>. The distinction is highlighted in many discussions (e.g., Williams 1979 and Sorensen 1988, Ch. 1), but omissive conjunctions will do most of the work here.

irrational in **normal circumstances**, given the knowledge, beliefs, and aims held by speakers in such circumstances.

Explaining NO MOOREAN ASSERTIONS is not so easy as explaining why, say, it is irrational to assert an obvious contradiction. As is commonly observed, Moorean conjunctions are logically consistent, and often true. But despite the initial difficulty, a remarkably diverse range of explanations has 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", Gricean accounts citing conversational maxims or communicative intentions associated with assertion,[39] 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.[40] (These accounts are usually cast as explaining the *impropriety* of Moorean assertions, but they can be adapted to explain their *irrationality* by supposing that agents in normal circumstances aim to avoid impropriety and know it when they see it.)

How is NO MOOREAN ASSERTIONS supposed to get us to self-knowledge requirements? Probably the most detailed argument, from Sydney Shoemaker, is examined below in Section 6. But first, I want to consider a rough motivation that seems to me far more prevalent. Instead of proceeding directly from NO MOOREAN ASSERTIONS, it appeals to a corresponding requirement involving belief:

> (NO MOOREAN BELIEFS) Rationality requires that one not believe propositions of the form <*p*, but I don't believe that *p*>.[41]

It is widely supposed that NO MOOREAN BELIEFS gets us close to self-knowledge requirements, or to closely aligned rationalist theories of introspection, which deny the possibility of self-blindness.[42] But there is less consensus on finer-grained questions about exactly which self-knowledge requirements are supported, or how.

One natural idea, proposed for example by Jordi Fernández, is that NO MOOREAN BELIEFS directly supports NO HIGHER-ORDER ERRORS.[43] While we will see some wrinkles in Section 4, this suggestion has obvious appeal. In effect, NO HIGHER-ORDER ERRORS merely prohibits jointly believing the conjuncts of a Moorean conjunction. And jointly believing the conjuncts might seem obviously irrational if believing the whole conjunction is.

[39] For accounts drawing on Grice's work in different ways, see Martinich 1980 and Shoemaker 1996, pp. 38-40 and 75. As Martinich observes, Grice himself (1989, pg. 42) rejected Martinich's version.

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

[41] See, e.g., Chan 2010; de Almeida 2001 and 2007; Fernández 2005 and 2013, pg. 112; Gibbons 2013, pp. 3 and 231; Heal 1994, pg. 6; Kriegel 2004; Moran 2001, pg. 70; Shoemaker 1996, Chs. 2, 4, and 11; Silins 2013, pg. 297; Smithies 2012b, 2016, and 2019; Sorensen 1988, Ch. 1; Wedgwood 2017, pg. 45; Williams 2006 and 2007; and Zimmerman 2008, pg. 331.

[42] E.g., Fernández 2013; Moran 2001, pp. 69-77; Shoemaker 1996; Smithies 2016 and 2019; and Zimmerman 2008.

[43] Fernández 2013, pg. 15. See also Douven 2009.

Now it might seem that this direct motivation at best supports NO HIGHER-ORDER ERRORS, the weakest of the self-knowledge requirements. In my book this alone would be of interest, since I think NO HIGHER-ORDER ERRORS makes higher-order errors out to be different from errors about other matters of fact. But there might also be a way of directly supporting the stronger requirement HIGHER-ORDER BELIEFS. Following Declan Smithies, we might think that if conventional Moorean beliefs are irrational, then so are beliefs in conjunctions like <It will rain, but *maybe* I don't believe it will rain>.[44] Smithies plausibly takes the rationality of believing that maybe something is the case to go with that of lacking a belief that it is not the case. If so, HIGHER-ORDER BELIEFS also might be directly supported by broadly Moorean considerations.

A different idea is that self-knowledge requirements are supported indirectly, by an inference to the best explanation. This too is suggested by Fernández and Smithies, among others.[45] If Moorean beliefs are irrational, it might be hard to explain why without invoking a rationalist theory of introspection that vindicates SELF-KNOWLEDGE. Indeed, I think the inference to the best explanation is especially forceful if we accept the above direct motivation for NO HIGHER-ORDER ERRORS. If we thought it possible for an ideally rational agent to lack a capacity for introspective self-knowledge, it is hard to see what could prevent this self-blind agent from sometimes rationally holding false higher-order beliefs, for example when confronted with misleading behavioral evidence about her first-order beliefs.

In any case, we don't need to settle on the best path from NO MOOREAN BELIEFS to self-knowledge requirements, because I will claim in the next two sections that NO MOOREAN BELIEFS itself is unmotivated. In Section 4, I will consider the common view that Moorean beliefs must be irrational because they are in some sense guaranteed to be false. Then in Section 5, I will turn to the suggestion that the irrationality of Moorean beliefs is what explains the irrationality of Moorean assertions. I think both motivations for NO MOOREAN BELIEFS fail, and that the former would undermine the broader case for self-knowledge requirements even if it worked. After these sections opposing NO MOOREAN BELIEFS, I will get back to Shoemaker's argument in Section 6.

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

Many treat NO MOOREAN BELIEFS as an obvious datum, which should serve as a starting point for any plausible account of Moore's paradox.[46] I don't buy it. At best, what's an obvious datum is that Moorean beliefs are guaranteed to be in error, as claimed for example by Sydney Shoemaker, Declan Smithies, Roy Sorensen, Ralph Wedgwood, Aaron

[44] Smithies 2012a. Note Smithies' examples are different, because his topic is knowledge of justification rather than belief.

[45] Fernández 2013, pp. 138-140 and Smithies 2016, pg. 408 and 2019, Ch. 4. Though less explicit, an inference to the best explanation seems present in Moran (2001, pp. 69-77), Shoemaker (1996, pg. 77), and Zimmerman (2008, Sec. 3).

[46] E.g., Chan 2010, pg. 212; de Almeida 2001, pg. 33; Fernández 2005, pg. 534 and 2013, pg. 112; Gibbons 2013, pg. 213; Heal 1994, pg. 6; Kriegel 2004, pg. 100; Moran 2001, pg. 70; Shoemaker 1996, pp. 75-76; Smithies 2016, pg. 397; Wedgwood 2017, pg. 45; Zimmerman 2008, pg. 331.

Zimmerman, and Mitchell Green and John Williams.[47] The idea is that believing a conjunction of the form <*p*, but I don't believe that *p*> suffices for believing the first conjunct. And believing the first conjunct suffices for the second conjunct being false.[48] 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 one not believe that *p* if it is necessary that if one believes that *p*, then not-*p*.

But NO GUARANTEED ERRORS has unappealing implications. It arguably contradicts the common intuition that Lois Lane can rationally believe that Clark Kent is not Superman. And supposing a functioning cortex is necessary for having beliefs, it means a scientifically uninformed agent could not rationally believe a doctor's testimony that her cortex has ceased functioning.[49]

However, there is another way to defend NO MOOREAN BELIEFS, advanced by Declan Smithies, and by Mitchell Green and John Williams.[50] Rather than just assuming the obvious datum that Moorean beliefs are guaranteed errors, it assumes that this is an obvious datum. If it is obvious, then a rational agent with the relevant concepts is in a position to know it. So NO MOOREAN BELIEFS will follow given:

> (NO FORESEEABLE ERRORS) Rationality requires that one not believe that *p* if one can know that if one believes that *p*, then not-*p*.

To many, NO FORESEEABLE ERRORS might itself seem obvious. For comparison, it presumably is normally irrational to deliberately make *assertions* that are foreseeably erroneous, at least if one aims to speak the truth. For example, if I usually keep my meteorological opinions to myself, my evidence might support the quasi-Moorean conjunction <It will rain, but I will not assert that it will rain>. But it still would be irrational for me to deliberately assert this, since I can know that it must be false if I do.

My suspicion is that something like this analogy between assertion and belief (or judgment) is often behind the claim that NO MOOREAN BELIEFS is an obvious datum.

[47] Shoemaker 1996, pg. 76; Smithies 2016 and 2019; Sorensen 1988, Ch. 1 and pg. 388; Wedgwood 2017; and Williams 1994, pg. 165; Zimmerman 2008, pg. 329, and Green and Williams 2011, pp. 249-250. See also Briggs 2009, pg. 79, and see Chan 2010 for criticism.

[48] It is unclear beliefs in comissive conjunctions of the form <*p*, but I believe that not-*p*> are guaranteed errors. But they plausibly are still *likely* errors. This weaker claim is arguably all my opponent here needs to adapt her view to comissive cases, if she accepts NO FORESEEABLE ERRORS below.

[49] For discussion, see Barnett MSb. Maybe NO GUARANTEED ERRORS still could be upheld, for example by objectivists who accept NO ERRORS, which forbids false beliefs of any kind. But as I discussed in Section 2, this would deflate the interest of a weak self-knowledge requirement against higher-order errors. I think it also would undermine an indirect Moorean motivation for stronger self-knowledge requirements, since NO ERRORS would explain the irrationality of Moorean beliefs without them.

[50] Green and Williams 2011, pp. 249-250 and Smithies 2016 and 2019.

Shoemaker stresses such an analogy when introducing this now common claim.[51] And the analogy makes appearances in recent discussions from Alan Hájek, Richard Moran, Antonia Peacocke, Nico Silins, Declan Smithies, Timothy Williamson, and Mitchell Green and John Williams.[52]

But if this is the rationale behind NO FORESEEABLE ERRORS, it is a problematic one. The problem is not with its assumption that foreseeably erroneous assertions are irrational; they plausibly are, as discussed in Section 5 below. It is rather with the analogy between assertion and belief. Beliefs and assertions are importantly different, and the requirements for one might not apply to the other. Most obviously, beliefs, unlike assertions, are involuntary. And this plausibly has normative significance.[53] If beliefs were directly voluntary, you might settle whether to believe a proposition *p* by deliberating directly about the question whether to believe it. But beliefs typically are not settled this way, and perhaps cannot be. Instead, in ordinary doxastic deliberation, the object of your deliberation is the question whether *p*. This is a different question, and it might be rational to answer it differently.[54]

Indeed, even if you could indirectly produce beliefs voluntarily, for example by pressing belief-causing buttons, the buttons it would be rational to press might not always match the beliefs it is rational to hold. It is a familiar point that it can be rational to cause yourself to be irrational.[55] This goes for beliefs, too. Suppose you are offered a cash prize for believing some tree in the park has exactly 224,618 leaves. It plausibly would be irrational to believe this, but it surely is rational to press a button that will cause you to believe it.[56] The potential for mismatch between what it is rational to believe and what it is rational to cause yourself to believe is most obvious in cases like this, where your causing the belief is motivated by monetary aims, rather than purely truth-directed ones. But potential mismatch is not limited to these cases. Even if your aims are truth-directed, there can be mismatch in familiar cases of epistemic tradeoffs.[57] If I promise to tutor you in algebra if you believe I'm the coolest kid in school, that will not make it rational to believe I'm the coolest. But it might make it rational to cause yourself to believe it, if one measly false belief about me is outweighed by all the true beliefs you would get about algebra.

I raise all this because I think Moorean beliefs and some others under the purview of NO FORESEEABLE ERRORS are further cases of mismatch.[58] Now I don't mean the great many beliefs that are foreseeably erroneous just because one knows the relevant proposition

[51] 1996, pp. 78-79.

[52] Hájek 2007, pg. 219; Moran 2001, pg. 70; Peacocke 2017; Shoemaker 1996, pg. 78-79; Silins 2012; Smithies 2016 and 2019; Williamson 2000, pp. 255-256; and Green and Williams 2007, pg. 3.

[53] See, e.g., Hieronymi 2009.

[54] Cf. Hieronymi 2005 and Shah and Velleman 2005. Note that related points plausibly hold for other attitudes.

[55] E.g., Parfit 1986, pp. 12-13.

[56] Cf. Kelly 2002, pg. 171 and Rinard 2019, Sec. 3

[57] See, e.g., Berker 2013.

[58] See also Barnett MSb.

is false regardless of whether one believes it. It is of course irrational to believe propositions you know to be false. I have in mind Moorean conjunctions, and other unusual propositions that arguably must be false if believed, but which otherwise might be true. For example, the proposition that you doubt everything must be false if you believe it, and that seems like a good reason not to deliberately cause yourself to do so. Even so, a Pyrrhonian skeptic might have strong evidence supporting that she doubts everything, just as she can have strong evidence that she has hands. It is at least arguable that her doubts about both these propositions are equally irrational, because they violate a requirement to believe what one's evidence supports. For another example even closer to Moorean conjunctions, 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.

Sylvie rationally believes premises that entail the conjunction <Today's prediction is that I won't now believe any new conjunctions, and today's prediction is true>. But she can tell that if she were to believe it, then it would be false. That plausibly is enough to make it irrational for Sylvie to press a button causing herself to believe the conjunction. If she aims to have only true beliefs, it would be irrational to deliberately cause a foreseeably erroneous one. But does that mean it is irrational to believe the conjunction?

NO FORESEEABLE ERRORS says it does, at least without further qualification. For it prohibits holding foreseeably erroneous beliefs, not just deliberately causing them. Thus in Unbelievable Consequences, it implies Sylvie is required not to believe the conjunction. Now I do not take this implication to be obviously false. It is a tricky case, after all. But I do think that on close examination it is better to reject it, and to say Sylvie is required to believe the conjunction. Let me tell you why, before discussing a fallback position for those who are unconvinced.

Sylvie is not in the position of deliberating about whether to produce in herself a belief in the conjunction. She is in the position of deliberating about whether the conjunction is true. Since Sylvie rationally believes premises that entail that it is true, I say it is rational to believe it.

Indeed, Sylvie might have no other permissible options available. For we can suppose that the beliefs in her premises are themselves required. Sylvie surely knows, and cannot permissibly doubt, the premise that the Oracle predicts she will believe no new conjunctions. And she can have arbitrarily strong inductive evidence for the premise that this prediction is true.[59] To be sure, if Sylvie goes on to believe the conjunction, and has typical powers of introspection, she will be able to recognize that she has done so, despite the Oracle's prediction. At that point, Sylvie will have reason to change her mind about the premise that

[59] Depending on how liberal we are with attributing beliefs, it might seem unrealistic to suppose any agent might go even a moment without forming beliefs in many conjunctions. If so, we might suppose Sadie considers herself unusually skeptical, or else that the prediction includes some more idiosyncratic condition that the belief in question happens to meet. Thanks to anonymous editors for pressing this.

the Oracle's prediction is true, and so may no longer be required to believe her conjunction. If fact, if you accept (over my protests) that self-blindness is impossible, then you will think all this is bound to happen.[60] But this shows only that Sylvie's requirement to believe the conjunction is short-lived. Once she believes what is required, she will learn something new, and her situation will be different. This does not mean she is not now required to believe the conjunction, given her current situation.

So on my preferred view, NO FORESEEABLE ERRORS faces a counterexample, and should be rejected. It can be rational, and even required, for Sylvie to adopt a belief that is foreseeably erroneous. And if so, it is hard to see why the same could not go for Moorean beliefs, unless we have some *other* reason to think self-knowledge is rationally required. For if there could be an ideally rational but self-blind agent who believes it will rain while lacking introspective knowledge of her belief, she might find herself in a situation like Sylvie's. For she might falsely believe that she does not believe it will rain, and indeed might even be required to, if she has strong misleading behavioral evidence that she lacks the belief. On the view I am pushing, it then will be rational to believe the Moorean conjunction that it will rain, but that she does not believe it will. To be sure, even a self-blind agent ought to foresee that a belief in a Moorean conjunction would be erroneous. And that might mean it is irrational even for self-blind agents to deliberately cause themselves to believe Moorean conjunctions. But that does not mean it is irrational to believe them.

While that is what I think we should say about Moorean and other peculiar self-falsifying beliefs, you do not need to agree me in order to see some bigger problems with relying on them to support self-knowledge requirements. Supporters of NO FORESEEABLE ERRORS have two plausible avenues for resisting my claim that it is rational for Sylvie to believe her conjunction, and for a hypothetical self-blind agent to believe a Moorean one. But both come at great expense. My fallback position is that even if these succeed as local defenses of the irrationality of Moorean beliefs, they end up undermining the broader motivation for self-knowledge requirements in other ways.

The first response is to say that Sylvie can permissibly withhold belief from her conjunction, even while she believes each conjunct separately. This means rejecting:

> (RESTRICTED MULTI-PREMISE CLOSURE) If one considers the matter and one knows that *p* and *q* jointly entail *r*, then rationality requires that if one believes both that *p* and that *q* with sufficient confidence, then one believes that *r*.

But RESTRICTED MULTI-PREMISE CLOSURE is hard to give up. We have seen that unqualified closure requirements like SINGLE PREMISE CLOSURE face overdemandingness and defeasibility objections. Maybe this means they should be rejected, or qualified. Or that rationality's requirements are in fact demanding and indefeasible. Whatever we say about these issues, RESTRICTED MULTI-PREMISE CLOSURE sidesteps them by applying only when one considers a conclusion and knows that it is entailed by one's premises. Multi-premise closure principles also are known to face distinctive problems involving the accumulation of

[60] Cf. Smithies' (2019) discussion of finkish propositional justification. See also fn. 66 below for more.

risk over multiple premises, for example in preface cases.[61] RESTRICTED MULTI-PREMISE CLOSURE avoids those, too, by applying (stipulatively) only to cases where one believes two premises with sufficient confidence to prevent the accumulation of risk from being a factor. (We can suppose Sylvie believes the relevant premises with arbitrarily high confidence, and considers whether their conjunction is true.)

For all that, maybe RESTRICTED MULTI-PREMISE CLOSURE could be rejected by the proponent of NO FORESEEABLE ERRORS. She might want to further restrict the closure requirement so as to not require foreseeably erroneous beliefs, or even say that rationality has no closure requirement at all. But even if this response is plausible taken on its own, it is not available to the Moorean supporter of self-knowledge requirements. For it saves NO FORESEEABLE ERRORS at the expense of undermining its broader relevance to self-knowledge. Remember, the point of prohibiting foreseeable errors was to motivate NO MOOREAN BELIEFS, which prohibits beliefs in Moorean conjunctions. And that was supposed to motivate, if not a requirement to know one's beliefs, then at least a prohibition against having erroneous higher-order beliefs about them. But this crucially assumes that if it is irrational to believe the Moorean conjunction <*p*, but I don't believe that *p*>, then it must be irrational to jointly believe that *p* and believe that one does not believe that *p*. It is hard to see why we should assume this if we reject even a weak closure requirement like RESTRICTED MULTI-PREMISE CLOSURE. We surely should not assume it if rationality has no closure requirement at all. And we still should not assume it if rationality has a closure requirement that is restricted so as to not require foreseeably erroneous beliefs, since Moorean beliefs are foreseeably erroneous, too.

Thus the first response undermines a crucial assumption behind what in Section 2 I called the 'direct' motivation for self-knowledge requirements. In doing so, I think it also manages to undermine the 'indirect' motivation, which appeals to an inference to the best explanation. For suppose Sylvie can, if she is ideally rational, avoid a foreseeably erroneous belief in her conjunction even while believing both its conjuncts. If so, then a hypothetical self-blind agent should be able to avoid Moorean beliefs even without any capacity for self-knowledge. For even if the self-blind agent is misled into erroneous higher-order beliefs, he still will avoid beliefs in the relevant foreseeably erroneous Moorean conjunction, just as Sylvie allegedly will. The upshot is that if we reject RESTRICTED MULTI-PREMISE CLOSURE, there is no apparent way to motivate self-knowledge requirements on Moorean grounds.

The second response takes another tack. It grants that Sylvie is required to believe the conjunction, as RESTRICTED MULTI-PREMISE CLOSURE says. But it says she also can be required not to believe it, in keeping with NO FORESEEABLE ERRORS. The response thus deems Unbelievable Consequences a rational dilemma, in which rationality's requirements are not mutually satisfiable.

Now there may be general objections to allowing the requirements of rationality to pull agents in conflicting directions like this.[62] But even setting general objections aside, allowing them in the present context again undermines the broader Moorean motivation for self-knowledge requirements. If Unbelievable Consequences is admitted as a dilemma, an

[61] See, e.g., Kyburg 1961 and Christensen 2004, and for further connections to Moore's paradox, Sorensen 1988, pp. 23-26.

[62] For further discussion, see Barnett forthcoming and MSc.

opponent of self-knowledge requirements can say the same for an alleged self-blind agent who believes it will rain, but whose behavioral evidence supports that he doesn't believe it. That is, even if the agent is prohibited from believing the relevant Moorean conjunction, he might still rationally hold beliefs that generate a conflicting requirement to believe it.

In response, it could be claimed that if believing the Moorean conjunction is prohibited, it follows by RESTRICTED MULTI-PREMISE CLOSURE that this allegedly self-blind agent is required not to jointly believe both conjuncts. But I am not sure this does follow if rationality's requirements are admitted to be inconsistent. For comparison, if we say Sylvie is in a dilemma with respect to her belief in the conjunction, that does not obviously make her beliefs in the premises irrational. And even if did follow, it would provide only a Pyrrhic victory for for rationalist theories of introspection that uphold self-knowledge requirements. For it would at best show that if the self-blind agent rationally believes it will rain, he is *required* to avoid the erroneous higher-order belief that he does not believe it. It will not mean he is *permitted* to avoid this erroneous higher-order belief, much less permitted to adopt the true higher-order belief that he does believe it will rain. Although technically consistent with some self-knowledge requirements, it is hard to square this result with an overall satisfying view connecting self-knowledge and rationality. A ban on foreseeably erroneous beliefs cannot support a *bona fide* self-knowledge requirement if it cannot even guarantee us permissible higher-order belief.

5. Second Route to Self-Knowledge Requirements: Inference to the Best Explanation

There is another way to motivate NO MOOREAN BELIEFS, the rational prohibition on believing Moorean conjunctions. It draws on influential **epistemic accounts** of Moorean assertions, as proposed for example by Claudio de Almeida, Timothy Chan, Uriah Kriegel, Sydney Shoemaker, and John Williams.[63] Instead of invoking troublesome assumptions about the general irrationality of foreseeably erroneous beliefs, it appeals to the widely acknowledged datum that *asserting* Moorean conjunctions is irrational. The idea is that Moorean beliefs being irrational would elegantly explain why Moorean assertions are irrational. And if so, the irrationality of Moorean beliefs could be supported by an inference to the best explanation.

The details of the epistemic account will take some filling in. But the rough 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. So if Moorean beliefs are irrational, Moorean assertions are, too.

To be sure, supporters of epistemic accounts do not always represent themselves as supporting NO MOOREAN BELIEFS by an inference to the best explanation. Sometimes they take the irrationality of Moorean beliefs to be independently obvious, in addition to best explaining the data about Moorean assertions. I explained in Section 4 why I don't buy that NO MOOREAN BELIEFS is a datum like NO MOOREAN ASSERTIONS. In this section, I address the claim that it best explains the uncontroversial data. I will not deny that epistemic accounts also can explain the irrationality of Moorean assertions. Instead, I will propose an alternative explanation, which I think is better because there are further data that only it can explain.

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

Now it is widely acknowledged that epistemic accounts are not the only game in town. As noted in Section 3, there are also **pragmatic accounts**, which include 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.

While my main objective is to oppose epistemic accounts, I do not accept a traditional pragmatic account, either. This is because I agree with supporters of epistemic accounts that the pragmatic aims or norms these accounts invoke are not really needed to explain the irrationality of Moorean assertions. Whatever other aims speakers in normal circumstances have, they surely have alethic aims to assert truths but not falsehoods. And these alethic aims suffice to explain the irrationality of Moorean assertions, as illustrated by:

> **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, her behavioral evidence about her belief notwithstanding. But she has no Gricean aims to cooperate with an audience, or Wittgensteinian aims to express herself. Nor does she have any other relevant non-alethic aims, such as a Williamsonian aim to mark only statements that she knows.[64] (It might for example be rational 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).

Epistemic accounts do not have this problem, since they can appeal only to alethic aims that Sadie shares with normal asserters. That is something I think epistemic accounts get right. Where they go wrong is in is in assuming that this makes the relationship between rational belief and assertion a straightforward one. Let **endorsement** be a general category covering assertion, marking as true on Sadie's exam, or any 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. (I assume this ratio to be high, since otherwise assertion might not be strong evidence of belief—a point I return to in Section 6.) Here is a first stab at the relationship between belief and endorsement:

[64] Cf. Littlejohn 2010 and Williamson 2000, pp. 253-254.

> (ENDORSEMENT→BELIEF) Rationality requires that one not endorse that *p* unless one believes that *p*.[65]

ENDORSEMENT→BELIEF arguably favors epistemic accounts. For it entails that a Moorean endorsement, whether an ordinary assertion or an endorsement on an exam like Sadie's, can be rational only if a Moorean belief can be rational. And so the irrationality of Moorean assertions can be explained by the alleged irrationality of Moorean beliefs.

But I think this explanation leaves something out. There are many statements it might be irrational to endorse because it is irrational to believe them, such as obvious contradictions. But Moorean endorsements seem to have a more distinctive self-undermining character, which I think should feature in a complete account of their irrationality. Despite important differences we will return to, just compare Sadie's Exam to the following:

> **Ned's Exam:** Neutral Ned is taking an exam like Sadie's. When he reaches the final statement, he still has not endorsed any statements, and based on his track record he believes he will not endorse the final statement, either. He then reads the final statement, which says 'I, Ned, do not endorse any statements on this exam.'

Ned should not endorse. Even though he believes the statement is true, he should recognize that it must be false if he endorses it. And plausibly, if Ned is deciding whether to endorse, he should consider not just whether the statement is true, but whether it will or would be true if he does so.

I will argue that something similar goes for Moorean statements like Sadie's. Even supposing that it could be rational to believe a Moorean statement, it still would be irrational to endorse it. For one should recognize that it probably is not true if one does. That of course not show that Moorean beliefs really can be rational. But it does undercut a reason to think they cannot be, namely the epistemic account's inference to the best explanation of Moorean endorsements.

Statements like Ned's complicate the relationship between belief and endorsement. It can be irrational to endorse them even if you believe they are true. From the looks of things, ENDORSEMENT→BELIEF cannot explain why, since it merely prohibits endorsing

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

statements that are not believed.[66] And if so, the right account will need some other requirement to supplement or replace it.[67] The crucial question for us is whether the replacement will end up applying to Moorean conjunctions, too. For despite some similarities, we will see that there are important differences between Ned's statement and a typical Moorean conjunction, differences which closely track those between two potential replacements for ENDORSEMENT→BELIEF:

> (ENDORSEMENT→BELIEF$_C$): Rationality requires that one not endorse that *p* unless one believes that if one were to endorse that *p*, then *p*.
>
> (ENDORSEMENT→BELIEF$_I$): Rationality requires that one not endorse that *p* unless one believes that if one does endorse that *p*, then *p*.

Of these, ENDORSEMENT→BELIEF$_C$ will prove to be the requirement favoring epistemic accounts of Moore's paradox. It permits endorsing a statement only when you believe the (non-backtracking) counterfactual that if you were to endorse it, then it would be true. Generally speaking, this will prohibit endorsing a statement which you believe only in cases where you think endorsing might *cause* (or constitute) the statement's being false. In contrast, ENDORSEMENT→BELIEF$_I$ will prove to be the requirement favorable to my account. It permits endorsing only when you believe the indicative conditional that if you do endorse the statement, then it is true. Importantly, this can prohibit endorsing a believed statement even when endorsing is merely evidence that the statement is false already.

Arguably unlike ENDORSEMENT→BELIEF, these requirements both prohibit Ned from endorsing. Ned should recognize that if he will in fact endorse, then the statement is false. And he also should recognize that even if he does not in fact endorse, the statement would have been false if he had endorsed it. But there are other cases where the requirements diverge. Before returning to Moorean statements, consider one more example:

> **Rachael's Exam:** Based on strong evidence, Rachael believes herself to be extremely risk-averse. On her exam, she encounters the statement 'I, Rachael, am extremely risk-averse.' Although Rachael believes the statement, she believes even more strongly that an extremely risk-averse agent would not endorse it.

[66] Could it be that despite appearances, ENDORSEMENT→BELIEF prohibits endorsing? As anonymous editors emphasize, if Ned were to endorse the statement, his beliefs might change. Assuming he knew he was endorsing the statement, he no longer would believe it, and thus would violate ENDORSEMENT→BELIEF.

If that is enough for ENDORSEMENT→BELIEF to deem endorsing irrational, that would be welcome news for me. For the reasons discussed shortly, it would mean ENDORSEMENT→BELIEF can deem Moorean endorsements irrational without Moorean beliefs being irrational. But I'm afraid it would be premature to declare victory. For I suspect that ENDORSEMENT→BELIEF at best explains why endorsing *would have been* irrational for Ned if he had done it. In contrast, it seems to me Ned already has ample reason not to endorse, given his actual beliefs—which after all are the beliefs that must motivate him if he decides to refrain. So rather than look to the beliefs that Ned would have, conditional on his endorsing, I think it is better to do the same job by appealing to Ned's actual conditional beliefs. As we will see, doing so allows more wiggle room for my opponent, since it allows a possible explanation of Ned's Exam, invoking his beliefs in counterfactuals, that would not carry over for Moorean endorsements.

[67] I favor replacement. Suppose the final statement instead had been 'I, Ned, do at some time endorse a statement on this exam.' Ned should endorse, but ENDORSEMENT→BELIEF arguably says otherwise.

Rachael's situation is surely possible. Given strong enough evidence, Rachael could rationally believe herself to be risk-averse, even irrationally so.[68] But even if so, Rachael should not endorse. While she believes the statement is true, she also thinks that if she endorses, then it probably is not true after all. To endorse a statement in such circumstances thus seems irrationally self-undermining.

Rachael's situation is different from Ned's, however, since she should not take endorsing to have the causal power to alter her risk aversion. What she believes is that if she does endorse, then she (already) probably is not risk-averse; not that if she were to endorse, then she would not be risk-averse. For this reason, ENDORSEMENT→BELIEF$_I$ yields the correct verdict that endorsing is irrational, but ENDORSEMENT→BELIEF$_C$ does not.

This shows that ENDORSEMENT→BELIEF$_C$ is too weak. We need to replace or supplement it with something like ENDORSEMENT→BELIEF$_I$.[69] And this matters, because it undermines the motivation for epistemic accounts of Moore's paradox. So long as ENDORSEMENT→BELIEF$_I$ is upheld, we can explain the irrationality of Moorean endorsements without supposing Moorean beliefs to be irrational.

To see why, suppose for sake of argument that it is possible for an agent like Sadie to rationally believe her Moorean statement, 'It will rain, but I, Sadie, do not believe it will rain'. This might happen, for example, if Sadie could be self-blind, and have misleading behavioral evidence about her meteorological beliefs. Even so, ENDORSEMENT→BELIEF$_I$ will in normal circumstances prohibit Sadie from endorsing her statement, just as it did for Rachael. This is because Sadie's endorsing the statement would be evidence Sadie believes each conjunct, so long as Sadie considers herself more likely to endorse the statement if she does. And by providing evidence that Sadie believes the first conjunct, her endorsing also provides evidence that the second conjunct is false. So even if Sadie believes the conjunction is true, she should recognize that if she endorses, then it might not be true after all (and indeed, probably isn't). Thus ENDORSEMENT→BELIEF$_I$ says endorsing a Moorean conjunctions is normally irrational even if we suppose it can be rational for an agent to believe it.

Even so, if a Moorean conjunction like Sadie's is true, her endorsing it would do nothing to change that. If Sadie can rationally believe the conjunction is true, she also should believe that if she had endorsed it, then it still would have been true. So like endorsing Rachael's statement, endorsing Sadie's Moorean conjunction does not violate ENDORSEMENT→BELIEF$_C$.

Moorean endorsements are thus self-undermining in a peculiar way. If Sadie believes the conjunction, she should think endorsing would contribute to her score on the exam. But she should not think it would on the assumption that she actually does endorse. In terminology popularized by Richard Jeffrey, this means Moorean endorsements, along with other endorsements violating ENDORSEMENT→BELIEF$_I$ but not ENDORSEMENT→BELIEF$_C$, are **unratifiable**.[70]

[68] Christensen 2007b.

[69] Again, I favor replacement. Suppose Rachael instead encountered 'I am *not* irrationally risk averse'. ENDORSEMENT→BELIEF$_C$ arguably says endorsing is irrational, but I disagree.

[70] 1983, pp. 15-20.

Roughly speaking, an action is (causally) unratifiable if on the assumption one will adopt it, another of one's options has higher (causally) expected utility.[71] Actions of this general kind have been widely discussed since at least Gibbard and Harper's Death in Damascus case.[72] While controversial, it is often thought that unratifiability somehow counts against an action's rationality.[73] Here is a recent example, which Andy Egan credits to David Braddon-Mitchell:[74]

> **Psychopath Button:** Paul stands before a button marked "KILL ALL PSYCHOPATHS". He would like to rid the world of psychopaths, but not at the expense of killing himself. Paul believes with sufficient confidence that he is not a psychopath to make pressing rational, if not for one final detail: He is certain that only a psychopath would press the button.

Should Paul press? He believes he is not a psychopath, and thus that if he were to press, he would kill all the psychopaths without killing himself. But he also is certain that only a psychopath would press. This makes pressing unratifiable, like endorsing in Sadie's Exam and Rachael's Exam. For while Paul thinks he will not press, and thus that pressing would have the results he prefers, he should at the same time recognize that if in fact he will press, then he must be a psychopath after all, in which case pressing will kill him. Like Egan, I think this makes pressing irrational. If you agree, that is further reason to accept my claims regarding unratifiable endorsements, including endorsement of a Moorean conjunction in Sadie's Exam.

More generally, I propose that Moorean endorsements are irrational because, by violating ENDORSEMENT→BELIEF$_{\text{I}}$ as discussed, they violate an even more general requirement against unratifiable actions.[75] Call this the **ratificationist account**. It explains the irrationality of Moorean assertions in virtue of what they have in common with other unratifiable actions like pressing in Psychopath Button, rather than by assuming with epistemic accounts that Moorean beliefs must be irrational. In doing so, the ratificationist

[71] For a precise definition, let U(B|A) be the (causally) expected utility of B-ing conditional on the assumption that one As, such that:

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

Here *Cr* is the agent's credence function and *v* her value function, which represents the degree to which she values relevant states of affairs. The *K*s are dependence hypotheses—i.e., maximal hypotheses about how outcomes depend counterfactually on one's actions that form a partition. Given all this, we can say option A is (causally) unratifiable iff one has an option O such that U(O|A) > U(A|A).

[72] Gibbard and Harper, 1978, pp. 156-159.

[73] E.g., Barnett MSc, Egan 2007; Gallow MS; Gustafsson 2011; Harper 1986; pg. 33; and Wedgwood 2011.

[74] Egan 2007.

[75] The details of such a requirement remain controversial, even among those who consider unratifiability to count against an action's rationality. My own proposal, defended in Barnett MSc, holds that ratifiability comes in degrees, and that one should prefer options with higher ratifiability to options with lower. Following the notation of fn. 71, when one's options are A-ing and B-ing, the ratifiability of A-ing is defined as U(A|A) - U(B|A). Importantly, this view says it is rational to adopt an unratifiable action when the alternatives are even less ratifiable. This arguably is needed to allow refraining to be rational in Sadie's Exam, Rachael's Exam, and Psychopath Button.

account appeals to rational requirements that are independently motivated if we think these other unratifiable actions are irrational. This does not directly show that Moorean beliefs can be rational. But without independent reason to think they cannot be, we should not suppose as much simply to explain the irrationality of Moorean assertions.

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

A final Moorean motivation for self-knowledge requirements comes from Sydney Shoemaker's influential discussion of self-blindness, the hypothetical condition of lacking introspective self-knowledge despite being ideally rational. Shoemaker argues from broadly Moorean considerations that it is impossible to suffer from this condition. If he is right, that at least gets us very close to alleged self-knowledge requirements like SELF-KNOWLEDGE, which requires one not to hold a belief without knowing that one does. Notwithstanding all the usual caveats about defeasibility and idealization, Shoemaker's conclusion would apparently mean that an ideally rational agent, who by stipulation satisfies all genuine rational requirements, could not violate the alleged requirement SELF-KNOWLEDGE. Now there might still be some daylight between such claims about the properties of ideally rational agents and claims about what rationality requires.[76] But I won't rest my opposition on these subtleties. Instead, I hope to show that Shoemaker's argument is unsound, considered on its own terms. For all it shows, an agent could be ideally rational despite lacking introspective knowledge of her beliefs.

Shoemaker's argument against the possibility of self-blindness proceeds by *reductio*. Even supposing self-blindness is possible, he claims, any self-blind agent would on account of her rationality avoid Moorean assertions, or any other actions evincing her self-blindness. So, where **self-aware** agents are rational agents with a genuine faculty self-knowledge, Shoemaker claims:

> (BEHAVIORAL INDISTINGUISHABILITY) Self-blind agents would act like self-aware agents.

Shoemaker regards as absurd the conclusion that self-blind agents are possible, but would act just like self-aware agents. So he rejects the supposition that self-blindness is possible at all.

Now BEHAVIORAL INDISTINGUISHABILITY is the crucial lemma that Shoemaker supports via Moore's paradox, and it is the one I focus on below. But other critics might instead just accept the alleged *absurdum* that self-blind agents are possible, but would act like we self-aware agents. As Amy Kind emphasizes, anyone who rejects behaviorism thinks it is possible for two agents to differ mentally while resembling each other behaviorally.[77] Unless we are behaviorists, why not say the same of self-blind and self-aware agents?

But it is not clear to me that the charge of behaviorism sticks. Anticipating it, Shoemaker emphasizes that a faculty for genuine self-knowledge is not just a metaphysical

[76] Shoemaker's conclusion might be weaker than SELF-KNOWLEDGE if necessary conditions on ideal rationality need not be genuine rational requirements. (Maybe there is no rational requirement not to be a round square, despite the impossible of an ideally rational agent who is a round square.) For one way Shoemaker's conclusion might be stronger than SELF-KNOWLEDGE, see Section 7 below.

[77] Kind 2003.

possibility, but is actually present in us, the products of evolution by natural selection. If self-blind agents were possible, but would act just like self-aware agents, then it would be inexplicable why we have such a faculty. For it could not plausibly confer any reproductive advantage without affecting our behavior in some way. 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."[78]

While this response is hardly decisive, I share Shoemaker's sense that there is something mildly fishy, if not entirely absurd, about supposing ourselves to have a psychological faculty for self-knowledge with no behavioral effects. Yet if we grant BEHAVIORAL INDISTINGUISHABILITY, the most obvious way to avoid the fishiness is to follow Shoemaker in denying the possibility of self-blindness altogether. So I say we reject BEHAVIORAL INDISTINGUISHABILITY instead. Our capacity for self-knowledge is not behaviorally inert, and self-blind agents, while possible, would not act like we do.[79]

Why does Shoemaker accept 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.
>
> (CONFORMITY) Self-blind agents would conform to rationality's requirements.
>
> Therefore, (BEHAVIORAL INDISTINGUISHABILITY) Self-blind agents would act like self-aware agents.

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.[80]

[78] 1996, pp. 239-240.

[79] So what are the main advantages conferred by our capacity for introspection? While this is an important question, I will avoid amateur speculation. What can responsibly be done from the armchair, in my view, is to reject Shoemaker's extreme claim that a contingent faculty for self-knowledge would have no behavioral effects at all.

[80] 1996, pg. 237.

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 is what gives them their distinctive self-defeating character. Second, they are assertions. What goes for them might not go for other kinds of actions, even ones with a broadly Moorean character.

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:

> **George's Exam:** 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 believes it will rain based on strong meteorological evidence, but his behavioral evidence misleadingly supports that he does not believe this. (Perhaps he left his umbrella at home.)

A self-aware agent in George's situation could know by introspection that she believes it will rain, despite her behavioral evidence. Being rational, she would then refrain from endorsing the first statement. Will George? His behavioral evidence supports the statement's truth, and we are supposing he lacks introspective knowledge to the contrary. Given this supposition, it would seem the rational thing is for him to believe the statement, and act on his belief by endorsing it. So even if George believes and acts in an ideally rational manner, his lack of self-knowledge will show.

To be sure, George might aim to coordinate first-order and higher-order endorsements, and he 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. Because of his false higher-order belief, when George makes the higher-order endorsement, he will falsely believe that he *is* coordinating.

Now in Section 5, I claimed that rationality requires one not to endorse Moorean conjunctions, even if one believes them. If so, an ideally rational agent like George would refrain from endorsing Moorean conjunctions. Why are things different in George's Exam? It is because one's endorsing a Moorean conjunction is self-defeating in a way that endorsing 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 character, which piecemeal assertions of their conjuncts lack.

Now I don't take this objection to be decisive, since it leans on my own ratificationist account of Moorean assertions. 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.

By stipulation, Moorean endorsements must have a certain risk profile; the ratio of the disvalue of false endorsement to the value of true endorsement has to be quite high, to match ordinary assertion. This is why one's endorsing a proposition amounts to strong evidence that one believes it. But many Moorean actions do not have this feature, such as:

> **Umbrella Rental:** George believes that it will rain based on strong meteorological evidence, but his behavioral 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 will rational George avoid a Moorean umbrella rental, like he would avoid asserting the corresponding Moorean conjunction? Not necessarily. Unlike asserting, renting a Moorean umbrella is not strong evidence that George believes it will rain, and so not strong evidence that the Moorean conjunction is false. It would be strong evidence that George lacks the belief that it won't rain, but that is what his behavioral evidence already supports. Given his supposed self-blindness, it is hard to see what might stop him from renting. This stands in contrast to a self-aware agent in George's situation. Since this agent would know introspectively that she believes it will rain, she will know the umbrella is useless, and won't rent it. So contrary to Shoemaker's claims, this introspective self-knowledge will hardly be behaviorally redundant. Once again, the difference between self-aware and self-blind agents will show.

The same goes for Moorean actions that resemble paradigmatic Moorean endorsements by involving an explicit formulation of the relevant proposition, 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 believes it will rain based on strong meteorological evidence, 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 may well mark it as true. For his doing so might be little evidence of belief, and hence little evidence the statement is false. 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 will not mark the statement as true.

The upshot is that we should reject PROXY. Moorean assertions have an 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 necessarily act like self-aware ones in general, even though they resemble them in avoiding Moorean assertions. So BEHAVIORAL INDISTINGUISHABILITY is false, and Shoemaker's argument for the impossibility of self-blindness is unsound.

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 proposal, one might satisfy the familiar requirements without knowing one's beliefs, but still fail to be ideally rational by violating an additional requirement for self-knowledge. On the latter, it is impossible to fully satisfy even these familiar requirements without having self-knowledge.[81]

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.

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

References

Adler, Jonathan and Armour-Garb, Bradley (2007) 'Moore's Paradox and the Transparency of Belief' in *Moore's Paradox: New Essays on Belief, Rationality, and the First-Person*, Mitchell Green and John Williams, eds. Oxford: OUP.

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

——— (forthcoming) 'Internalism, Stored Beliefs, and Forgotten Evidence,' In *Memory and Testimony: New Essays in Epistemology*, Stephen Wright and Sanford Goldberg. Oxford University Press.

——— (MSa) 'Higher-Order Evidence is the Wrong Kind of Reason' available at: www.davidjamesbar.net.

——— (MSb) 'Cogito and Moore' available at: www.davidjamesbar.net.

——— (MSc) 'Graded Ratifiability' available at: www.davidjamesbar.net.

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

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

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

Briggs, Ray (2009) 'Distorted Reflection' *Philosophical Review* 118(1): 59-85.

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.), OUP.

Burge, Tyler. 2013. *Cognition Through Understanding: Philosophical Essays, Vol. 3*. Oxford: OUP.

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

——— (2018) *Transparency and Self-Knowledge*, New York: OUP.

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.

Cassam, Quassim (2014) *Self-Knowledge for Humans*, New York: OUP.

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: OUP.

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

——— (2007b) 'Does Murphy's Law Apply in Epistemology' *Oxford Studies in Epistemology* 2: 3-31.

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: OUP.

Douven, Igor (2009) 'Assertion, Moore, and Bayes' *Philosophical Studies* 144(3): 361-375.

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.

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.

Gallow, Dmitri (MS) 'Manage the Improvement News'

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/>.

Gibbard, Alan and Harper, William (1978) 'Counterfactuals and Two Kinds of Expected Utility', in *Foundations and Applications of Decision Theory, vol. 1*, A. Hooker, J. J. Leach & E. F. McClennen (eds.), Boston: D. Reidel.

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. Oxford: OUP.

——— (2011) 'Moore's Paradox, Truth and Accuracy' *Acta Analytica* 26(3): 243-255.

Grice, H. P. (1989) *Studies in the Way of Words*. Harvard University Press.

Gustafsson, Johan (2011) 'A Note in Defense of Ratificationism' *Erkenntnis* 75: 147-150.

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. Oxford: 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.

Hieronymi, Pamela (2005) 'The Wrong Kind of Reason' *Journal of Philosophy* 102(9): 437-457.

——— (2009) 'Believing at Will' *Canadian Journal of Philosophy* 35(supp.1): 149-187.

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.

Jeffrey, Richard (1983) *The Logic of Decision, 2nd ed.* University of Chicago Press.

Kelly, Thomas (2002) 'The Rationality of Belief and Some Other Propositional Attitudes' *Philosophical Studies* 110: 163-196.

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

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.

Kyburg, Henry, Jr. (1961) *Probability and the Logic of Rational Belief.* Wesleyan University Press.

Lasonen−Aarnio, Maria (2008) 'Single Premise Deduction and Risk' *Philosophical Studies* 141(2): 157-173.

——— (2010) 'Unreasonable Knowledge' *Philosophical Perspectives* 24: 1-21.

——— (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.

——— (forthcoming) 'Objectivism and Subjectivism in Epistemology' in *The Factive Turn in Epistemology*, Veli Mitova (ed). CUP.

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.

Parfit, Derek (1986) *Reasons and Persons*. Oxford: OUP.

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: OUP.

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

Rinard, Susanna (2019) 'Equal Treatment for Belief' *Philosophical Studies* 176(7): 1923-1950.

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

Schechter, Joshua (2013) 'Rational Self-Doubt and the Failure of Closure' *Philosophical Studies* 163(2): 428-452.

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.

Shah, Nishi and Velleman, J. David (2005) 'Doxastic Deliberation' *Philosophical Review* 114(4): 497-534.

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

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

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

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

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

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

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

——— (2019) *The Epistemic Role of Consciousness*. OUP.

Sorensen, Roy (1988) *Blindspots*. Oxford: OUP.

van Fraassen, Bas (1984) 'Belief and the Will' *Journal of Philosophy* 81(5): 235-256.

——— (1995) 'Belief and the Problem of Ulysses and the Sirens' *Philosophical Studies* 77(1): 7-37.

Way, Jonathan (forthcoming) 'Reasons and Rationality' in *The Oxford Handbook of Reasons and Normativity*, Daniel Star, ed. OUP.

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

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

——— (2017) *The Value of Rationality*. Oxford: OUP.

Williams, John (1979) 'Moore's Paradox: One or Two?' *Analysis* 39(3): 141-142.

——— (1994) 'Moorean Assertion and the Intentional "Structure" of Assertion' *Analysis* 54(3): 160-166.

——— (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: OUP.

Williamson, Timothy (2000) *Knowledge and Its Limits*. Oxford: OUP.

——— (2017) 'Ambiguous Rationality' *Episteme* 14(3): 263-274.

——— (forthcoming) 'Justification, Excuses, and Skeptical Scenarios' in *The New Evil Demon Problem*, Julian Dutant and Fabian Dorsch (ed.), OUP.

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

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