David James Barnett
DRAFT: October 2019

Cogito and Moore

Abstract. One cannot falsely judge that one exists, but that is not because the proposition *I exist* is a necessary truth. Rather, it is because its truth is somehow guaranteed by one's affirming it. Relatedly, the Moorean proposition *It will rain, but I don't believe it will rain* is not a necessary falsehood. But it still is arguably one that cannot be correctly judged to be true, since one's doing so would in some sense guarantee its falsity. These facts are sometimes supposed to account for the rationality of *cogito*-like judgments, and the irrationality of Moorean ones. But if so, we are left with a puzzle. The necessary truth that *cogito*-like judgments are self-verifying can hardly entail the contingent truth that one exists, or even provide non-deductive evidential support for it. Likewise, that Moorean judgments are self-falsifying neither contradicts nor provides evidence against the possibility that it will rain despite one's failing to believe it. So it is hard to see why these facts about self-verification and self-falsification should matter to the rationality of the corresponding judgments, so long as we think judgments are subject to conventional evidentialist epistemic norms. This puzzles can be resolved, I argue, by supposing judgments to be subject to norms of *practical* reason. But at what cost?

1. Introduction

In the Second Meditation, Descartes's Meditator judges that he exists. The reasoning preceding this judgment is elementary enough for beginning students to grasp, but it has proven surprisingly difficult for interpreters to reconstruct. Notably, the Meditator gives no argument for the conclusion that he exists; the famous "*cogito, ergo sum*" appearing only in other work. Instead, we find an argument for the distinct conclusion that the proposition *I exist* is **self-verifying**, in roughly the sense that a thinker's judging the proposition to be true guarantees that it is.[1]

It might seem obvious that recognizing that *I exist* is self-verifying justifies the Meditator in affirming it (i.e., judging it to be true). But it is not obvious how. The necessary truth that *I exist* is self-verifying does not entail the contingent truth that someone exists, let alone that any particular person does. In fact, it is hard to see how it could probabilistically support the Meditator's or anyone else's existence. So it is no wonder this passage has been a source of so much disagreement, apparently even in Descartes's own writings. When pressed, he often seems to concede that *I exist* is simply inferred from the introspectively known premise *I think*. Yet as I'll discuss, this does not do justice to the idea that self-verification is relevant to the judgment's justification.

Recent discussions of self-knowledge and epistemic paradox have emphasized a related phenomenon. Loosely inspired by G. E. Moore, many philosophers claim that propositions of the form *p, but I don't believe that p* are **self-falsifying**, in the sense that one's affirming

[1] Suppose Al affirms *I exist*, Betty affirms *I exist*, and Charlie affirms *Al exists*. Throughout I will assume it is Al and Betty, not Al and Charlie, who affirm the same proposition. But this is just a terminological convenience. The important thing is just that Al's and Betty's judgments can be classified together *somehow*, and that a rational agent can deliberate about whether to do whatever it is they both do.

them guarantees their falsity. The idea is this. Judging guarantees believing, and believing a conjunction guarantees believing each conjunct. Thus in affirming a Moorean conjunction, one believes its first conjunct, and thus guarantees its second conjunct is false.

Many philosophers have thought that on this account, Moorean judgments must be irrational.[2] But if this is so, it is not because they cannot be supported by one's evidence. Here is one example adapted from Declan Smithies and Ralph Wedgwood:[3]

> **Stubborn Stella:** Stella has conclusive meteorological evidence supporting that it will rain. But Stella stubbornly withholds belief that it will rain, and she she can tell by introspection that she withholds belief.

Stella knows that she does not believe it will rain. But her meteorological evidence supports that it will rain. So her total evidence supports the Moorean conjunction *It will rain, but I do not believe that it will rain*. Even so, Smithies and Wedgwood think Stella is in no position to rationally affirm this conjunction.

Here I will offer an account of why self-verifying judgments might be rational (or justified) even when evidentially unsupported, and self-falsifying ones irrational even when supported. The central conceit of the account is that judgments are governed by norms of *practical reason*. I use 'practical reason' in a broad sense, to include any deliberation whose object is the question what to do. It thus includes deliberation about what to believe or judge, even if undertaken in light of purely alethic aims to affirm truths but not falsehoods.

If I am right, some common intuitions about *cogito*-like and Moorean judgments will follow if we view judgment as governed by widely accepted norms of practical rationality. *Should* we view judgment that way? I will remain officially neutral, though I will voice some concerns in Section 4. If you like the common intuitions, you can take the account I develop here as explaining their truth. If you don't, you can take it as diagnosing a confusion underlying common but mistaken intuitions. But either way you take it, the account will have an important upshot for theories of self-knowledge. The phenomena of *cogito*-like and Moorean judgments are often considered not to be mere idle curiosities, but rather illustrative of central features of the nature of self-knowledge. My account will cast some doubt on this common view.

2. The Cogito

On my reading, Descartes had at least two distinct ideas about how judging *I exist* is justified, though I see no evidence that he saw them as distinct. Some commentators think they can reconcile the apparent inconsistencies in his various remarks, but I'm less optimistic.[4] My aim is not so much a faithful interpretation of Descartes's overall view as a reconstruction of one strand of his thinking that has proven especially influential.

---

[2] Shoemaker 1996, pg. 76; Smithies 2016 and forthcoming; 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.

[3] Smithies 2016 and Wedgwood 2017, pg. 45.

[4] E.g., Markie 1992.

Start with the idea that seems to me dominant in Descartes's own thinking, though it won't be my focus. It is that one begins with knowledge of particular thoughts, doubts, sensory perceptions, and the like, and from this infers one's existence. Since this this account has one's justification for judging *I exist* ultimately trace back to receptive knowledge of one's particular mental states, I call it the *introspective* account.[5]

The introspective account is suggested by several passages in the *Meditations*, and is directly endorsed in several other places. For example, when discussing the piece of wax late in the Second Meditation, the Meditator argues that any basis for affirming the wax's existence serves as a stronger basis for affirming his own existence. For instance, the fact that he seems to see the wax might provide some evidence of the wax's existence, but it "entails much more evidently" the existence of himself, the one to whom it seems this way.[6] So at least by this stage of the *Meditations*, the Meditator knows of his particular sensory perceptions, and can infer from them that he exists.

The introspective account is also supported by many passages outside the *Meditations*, beginning with correspondence preceding their publication, and continuing in the Fifth Replies and later the *Principles*.[7] It also fits the famous slogan "*Cogito, ergo sum*," which suggests knowledge of *I exist* proceeds by inference from an antecedently known premise about one's thinking.[8] While the slogan is absent from the *Mediations*, it appears in earlier and later writings, and in the replies to the *Meditations*.

But while the introspective account is surely an important part of Descartes's thinking, it doesn't exhaust it. Just consider the central discussion of the *cogito* in the Second Meditation:

> [H]ow do I know that there is not something else which does not allow even the slightest occasion for doubt? Is there not a God, or whatever I may call him, who puts into me the thoughts I am now having? But why do I think this, since I myself may perhaps be the author of these thoughts? In that case am not I, at least, something? But I have just said that I have no senses and no body. This is the sticking point: what follows from this? Am I not so bound up with a body and with senses that I cannot exist without them? But I have convinced myself that there is absolutely nothing in the world, no sky, no earth, no minds, no bodies. Does it follow that I too do not exist? No: if I convinced myself of something then I certainly existed. But there is a deceiver of supreme power and cunning who is deliberately and constantly deceiving me. In that case I too undoubtedly exist, if he is deceiving me; and let him deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something. So after considering everything very thoroughly, I must finally conclude that this

---

[5] I count as introspective both interpretations that say the justification of *I exist* must be inferential, and some that allow it to be intuitive, such as Markie's (1992). As I see things, the common distinction between inferential and intuitive accounts does not quite carve at the joints.

[6] CSM II 22.

[7] CSM III 98, CSM II 244, and CSM I 195. See also CSM II 409-410.

[8] See Hintikka 1962 for an attempt to distance the slogan from the introspective account.

> proposition, *I am*, *I exist*, is necessarily true whenever it is put forward by me or conceived in my mind.[9]

To be sure, it is possible to read parts of this as supporting the introspective account. The Meditator first considers his existence as a candidate for certainty when discussing the origins of some particular thoughts he is aware of. And he takes a more definitive stance on his existence after asserting:

(i) I have convinced myself that there is nothing in the world.

While I read this assertion as merely tracking the dialectic, it could be interpreted as playing a more substantive role. For it is after asserting (i) that the Meditator first observes:

(ii) If I convinced myself of something, then I exist.

And if we take the Meditator here as substantively committing himself to (i), then his assertion of (ii) can be read as part of a *modus ponens* argument that he exists.

But this reading struggles with the remainder of the passage. Rather than immediately concluding that he exists, the Meditator continues with what seems intended as an elaboration of the same point. And here his intention clearly is not to argue *modus ponens* as follows:

(iii) There is a deceiver constantly deceiving me.

(iv) If there is a deceiver constantly deceiving me, then I exist. ("In that case I too undoubtedly exist, if he is deceiving me.")

While the Meditator asserts (iv), surely he should not be understood as literally asserting (iii). Descartes would not have the Meditator judge that he exists based on the false premise that there is a deceiver.

Things get even worse for the introspective account with the final sentence, which states the upshot of the Meditator's reasoning. It is not that he exists, but rather that the proposition *I exist* is necessarily true whenever it is put forward by him or conceived in his mind.[10] Here too it seems implausible that the intention is to infer his existence from the premises:

(v) I am conceiving the proposition *I exist*.

(vi) If I am conceiving the proposition *I exist*, then I exist. ("*I am*, *I exist* is true whenever it is…conceived in my mind.")

Perhaps the Meditator is at this stage in a position to know (v) introspectively, and thus, once he appreciates (vi), to infer *I exist*. But if that were the intended reasoning, (v) and (vi) could just as well be replaced with:

---

[9] CSM II 16-17.

[10] Cf. Frankfurt 1970, Ch. 10.

(vii) I seem to see a piece of wax.

(viii) If I seem to see a piece of wax, then I exist.

Presumably these premises are no less certain than (v) and (vi). The only apparent difference is that (vii) is true when the Meditator looks at the wax, while (v) is true when he considers his existence. So the argument from (v) and (vi) is distinguished only by the fact that on those occasions when his existence is what happens to be on his mind, (v) will be among the available introspective premises.

This reading seems to leave something out.[11] In concluding with (vi) rather than (viii), the Meditator highlights something distinctive about the judgment that one exists. Since one cannot judge a proposition true without conceiving it, (vi) entails that a judgment that one exists cannot be in error. It would seem strange for this distinctive feature of the judgment *I exist* then to do no work in its justification.

This point is further reinforced by the Third Meditation, where the Meditator briefly rehearses the *cogito* while making another point. There he says "let whoever can do so deceive me, he will never bring it about that I am nothing, so long as I continue to think I am something."[12] This seems to reinforce the significance of one's thinking *I exist* as guaranteeing its truth.

So I think the better reading of the central *cogito* passage from the Second Meditation is that somehow (vi), and perhaps even (ii) and (vi), are supposed to directly justify judging *I exist*, without supplemental introspective premises like (i), (iii), and (v). How are they supposed to do that? Here again I think Descartes was not entirely consistent. Some of his writings, especially the unfinished *The Search for Truth*, seem to emphasize (iv) rather than (vi).[13] That is, they emphasize that one's existence is presupposed by the skeptical hypotheses that are required to make doubt reasonable, or maybe by the act of doubting itself. Perhaps the idea is that one cannot doubt that one exists because any grounds for doubting anything must involve skeptical hypotheses that presuppose one's existence. On some interpretations, this is the dominant idea in the *Meditations* as well.[14] While much of what I will say could be adapted to suit this interpretation, here I will stick to a more straightforward reading that emphasizes (vi). On this reading, what matters is not that *I exist* is presupposed by skeptical doubts, but just that it must be true whenever judged to be true —i.e., that it is self-verifying. I choose this focus in part because I am not sure the presuppositions of doubt really have the significance in the *Meditations* that they are claimed to, but also just for simplicity, and to better highlight an aspect of Descartes's thinking with a particular contemporary relevance.

In stressing that *I exist* is self-verifying, I join what are sometimes called **performative** accounts, a name I'll stick with despite finding it mildly inapt. While I have minor

---

[11] Cf Frankfurt 1970, Ch. 10, Kenny 1968, pp. 55-56.

[12] CSM II 25.

[13] See, e.g., CSM I 127 and 183-184, and especially CSM II 409-410 and 415-417—though I think some parts of the latter source plainly favor an introspective reading.

[14] Broughton 2002 Ch. 7 and Curley, 1978, Ch. 4.

disagreements with both the technical details and the textual evidence offered by other performativists like Jaakko Hintikka, I agree with their broader emphasis of the self-verifying character of *I exist*.[15] Rather than nit-picking disagreements with Hintikka's definition of (what I call) self-verification, my focus here will be answering a question that so far as I can see Hintikka and others do not even consider, perhaps because they assumed the answer was obvious: If *I exist* is self-verifying, how does this *justify* one in affirming it?

One possible answer says that affirmation is justified simply because it is true that *I exist* is self-verifying, whether the thinker knows it or not. But this view is substantively implausible if we grant the possibility of *a posteriori* necessary conditions for judgment, as Descartes himself probably would not have. For a crude example, suppose it turns out that neurons firing is a necessary condition for one to make a judgment. Then it would be true that if one judges that neurons are firing, then neurons must be firing. Even so, this judgment seems potentially unjustified if made by a scientifically uninformed thinker who is unaware that neurons firing is necessary for judgment. A related problem arises for Descartes, even assuming he denied *a posteriori* necessary conditions for judgment. Consider the proposition *More than 7,995 - (65 × 123) thinkers exist*. This proposition is self-verifying, since 7,995 - (65 × 123) = 0. But presumably Descartes would not consider one justified in affirming it unless one recognizes that it is self-verifying.

These examples suggest that the judgment *I exist* is justified if at all by one's *knowing* that it is self-verifying, or by knowing some closely related truth such as (vi). More generally, it might be thought that for any proposition ϕ,

> (SELF-VERIFICATION) If you know that ϕ is self-verifying, this makes it rational for you to judge that ϕ.

If Descartes indeed accepted SELF-VERIFICATION or something like it, he is in good company.[16] But if true, SELF-VERIFICATION raises serious puzzles. Typically, knowing *q* will justify you in affirming *p* only if you can infer *p* from q. The premise *q* will need to entail *p*, or inductively support *p*, or in some other way amount to evidence from which *p* can be inferred.[17] But the fact that a proposition is self-verifying usually does not entail or even support that it is true. Many false propositions are self-verifying. For every number *n*, the proposition that one is now thinking of *n* is self-verifying, since in judging that one is thinking of *n*, one will think of *n*. Even so, almost all of these propositions are false. Perhaps somehow one still can be justified in judging, say, that one is thinking of the number 36, given one's knowledge that such propositions are self-verifying. But it is not because, in

---

[15] Hintikka 1962. See also, e.g., Ayer 1953 and Williams 1978, pp. 74-77. And for problems with Hintikka's proposal, see Feldman 1973 and Frankfurt 1966 and 1970, Ch. 10.

[16] See, e.g., Burge 2013, Chs. 1-9, and also Pryor 2006 and MS for review and critical discussion.

[17] Cf. Barnett 2016 and Pryor 2005, Sec. 4.

recognizing this, one knows a premise from which it can be justifiably inferred that one is thinking of 36.[18]

Even so, I think there are already widely acknowledged rational norms that support SELF-VERIFICATION. The hitch is that they are norms governing practical, rather than theoretical reasoning. The official account will come later, after introducing related phenomena involving Moore's paradox. But the idea can be illustrated with an analogy to the ordinary voluntary act of assertion. Suppose you lack any evidence that you will just now refer in speech to the number 36. Even if your only aim is to speak the truth, you still can have sufficient practical reason to assert *I am now referring to 36*. Since you know that your asserting this proposition guarantees that it is true, you can simply decide to do so. Importantly, you do not need to assert first and then, once you realize you are making the assertion, for the first time gain justification for it. Sufficient reason for making the assertion is available antecedently, before you know you will make it. That is why the decision to assert can be rational in the first place.

So it is natural to look for an account of the *cogito* that stresses a resemblance between judgment and assertion. Perhaps this is why the resemblance was stressed by Hintikka,[19] not to mention Descartes himself.[20] In Section 4, I will explain what one must assume about the judgment and its resemblance to ordinary assertion to vindicate SELF-VERIFICATION. But first, I want to consider some related issues arising in contemporary discussions of Moore's paradox.

3. Moore's Paradox

G. E. Moore famously observed that it is somehow "absurd" to assert propositions of the form ϕ*, but I don't believe that* ϕ. Many recent philosophers have gone further, and endorsed a related claim for judgment:

(MOORE) It is irrational to affirm ϕ*, but I don't believe that* ϕ.

Indeed, MOORE is often considered an obvious datum, which should serve as a starting point for any plausible account of Moore's paradox.[21] But even if we regard MOORE as obvious, we should still find it puzzling. Moorean propositions are logically consistent, and can even be supported by one's evidence, as with Stubborn Stella. Why would it be irrational to affirm a proposition that one's evidence supports?

---

[18] Perhaps it could be claimed that there is simply a primitively rational transition from knowledge that a proposition is self-verifying to judgment that it is true. But without a more general explanation of why these transitions are rational, this proposal is liable to seem *ad hoc*. Pryor MS considers an explanation that he attributes to Ralph Wedgwood. I agree with Pryor's criticisms, and have argued for related claims in Barnett 2016.

[19] Hintikka 1962, pp. 13 and 18-19

[20] When the Meditator say *I exist* is true whenever "put forward [*profero*]" (CSM II 17), I think he probably means uttering it in speech, in contrast to conceiving it in his mind.

[21] E.g., Chan 2010; de Almeida 2001 and 2007; Fernández 2005 and 2013, Ch. 4, pg. 112; Gibbons 2013, pp. 3 and 231; Heal 1994, pg. 6; Kriegel 2004; Moran 2001, pg. 70; Setiya 2011; Shoemaker 1996, Chs. 2, 4, and 11; Silins 2013, pg. 297; Smithies 2012b, 2016, and forthcoming; and Williams 2006 and 2007.

A popular answer appeals to the claim that Moorean propositions are self-falsifying, in the sense that one's affirming them guarantees that they are false.[22] This claim plausibly follows from two premises; first, that affirming a proposition guarantees believing it, and second, that believing a conjunction guarantees believing each conjunct. For suppose one affirms the conjunction *It will rain, but I don't believe it will rain*. By the first premise, one is guaranteed to believe this conjunction, and then by the second premise guaranteed to believe its first conjunct. But that guarantees that the second conjunct, and hence the whole conjunction, is false.

Are these two premises plausible? This might depend on what exactly we mean by 'guarantees'. In the case of *I exist*, we can help ourselves to a strong reading, where guaranteeing means metaphysically sufficing for. But while one's judging *I exist* plausibly suffices for one's existing, things are murkier with our two premises here. It is at least defensible that affirming a proposition suffices for believing it. But it is not clear that believing a conjunction metaphysically suffices for believing each conjunct.

So here I will employ a more permissive sense of guaranteeing, on which *causal* sufficiency is enough. It is at least more plausible that that believing a conjunction in typical conditions causally suffices for believe each conjunct. And while I will avoid a detour into the topic here, in other work I defend a theory which could make do with a merely *epistemic* sense of guaranteeing, which requires merely that believing a conjunction is sufficient evidence one believes its conjuncts.[23]

These permissive senses of guaranteeing should also be welcome to anyone who upholds a requirement against "commissive" Moorean judgments—i.e., who says it is irrational to affirm *It will rain, but I believe it will not rain*. Even if affirming this conjunction metaphysically suffices for believing it, and believing it suffices for believing its conjuncts, it is harder to maintain that believing its first conjunct suffices for its second to be false. If it is possible to have inconsistent pairs of beliefs, then one's believing that it will rain will not metaphysically suffice for one's failing to believe that it will not rain. It is plausible, however, that believing it will rain typically causally suffices for not believing that it will not rain. And it is even more plausible that one's believing it will rain is typically sufficient evidence one does not believe it will not rain.

Even if it were denied that Moorean conjunctions are self-falsifying, the wider phenomenon of self-falsification should not be denied. If one affirms *I am not thinking of the number 36*, for example, this surely guarantees that the proposition affirmed is false. Other examples of self-falsification can be found in discussions of epistemic paradoxes. For a common fanciful example, suppose it turns out that a brain state S is identical to the judgment that one is not now in S. If so, then the proposition that one is not now in S will be self-falsifying. Judging that one is not now in S will guarantee that one is in S, in which case one's judgment is false. And unlike Moorean conjunctions, the proposition is guaranteed to be true if one does not affirm it. If one does not judge that one is not now in S, then one is not in S. This peculiar feature of the proposition is thought to make it especially paradoxical, since unlike Moorean propositions, one cannot straightforwardly

---

[22] E.g., Shoemaker 1996, pg. 76; Smithies 2016 and forthcoming; Sorensen 1988, Ch. 1 and pg. 388; Wedgwood 2017; and Williams 1994, pg. 165 and Green and Williams 2011, pp. 249-250. See also Briggs 2009, pg. 79.

[23] Barnett MS.

avoid irrationality by deciding to withhold judgment. We will return to this complication later on.

Assuming Moorean propositions are self-falsifying, why would that make Moorean judgments irrational? As with *cogito*-like propositions, it is not plausible that the mere fact that Moorean propositions are self-falsifying explains the irrationality of affirming them. Even if the proposition *No neurons are firing* is self-falsifying, that does not automatically make it irrational for a scientifically uninformed agent to affirm it. This suggests that it is at least being in a position to know that a proposition is self-falsifying that makes affirmation irrational, so that:

> (SELF-FALSIFICATION) If you are in a position to know that ϕ is self-falsifying, this makes it irrational for you to judge that ϕ.

Notice that while SELF-VERIFICATION had knowledge of self-verification make a judgment rational, SELF-FALSIFICATION lets merely being in a position to know of self-falsification make it irrational. This difference is independently motivated. In general, we think one must recognize the reasons in favor of a judgment for it to be (doxastically) rational. But one's being oblivious to reasons one possesses against a judgment does not prevent it from being irrational. There is of course room for disagreement about the finer points. If you have it on good authority that only 7,995 - (65 × 123) thinkers exist, and you have not bothered to do the arithmetic, it might be claimed rational for you to affirm this, even though you are in a position to know it is self-falsifying.[24] If so, then perhaps what matters is not what you are in a position to know, but what you ought to know, or what you ought to at least implicitly recognize. These finer points will not matter in what follows if we grant that a thinker with the relevant concepts ought to implicitly recognize that ϕ*, but I don't believe that* ϕ is self-falsifying.

But how does being in a position to know that a proposition is self-falsifying make it irrational to affirm the proposition? Self-falsifying propositions can be highly probable given one's evidence, even if one can know they are self-falsifying. It seems Stubborn Stella's evidence makes probable the relevant Moorean conjunction, for example. And you can have strong inductive evidence that you are not thinking of the number 36, even though you can know that this proposition is self-falsifying. It seems whatever irrationality is involved in affirming these propositions is not explained by one's current evidence failing to support them.

This point seems widely presupposed in discussions of epistemic paradox. Consider again the self-falsifying proposition that one is not now in brain state S, where S is identical to the judgment one is not now in S. This proposition is self-falsifying, which is supposed to count against the rationality of affirming it. But it also has the unusual trait of being guaranteed to be true if you do not affirm it, which would presumably count against refraining from affirming. What should you do? On Earl Conee's view, you should refrain from affirming this proposition.[25] On David Christensen's, you are in violation of a rational

---

[24] Cf. Sorensen 1988, pp. 28-29.

[25] 1987, pg. 327.

ideal whether you affirm it or not.[26] On Roy Sorensen's you should refuse to believe that a state like S exists, no matter your evidence.[27] Later on I will favor Conee's view, applied to judgment if not belief. For now, consider what all these views have in common. On all of them, whether you should affirm the self-falsifying proposition does not depend on what evidence you happen to possess regarding its truth.

But why should your evidence supporting a proposition not matter for the rationality of affirming it? An initial proposal says a self-falsifying proposition's probability on your current evidence is irrelevant because your evidence will change as soon as you affirm it. That is, if you were to affirm it, you would be able to know introspectively that you did so. In that case, you will be in a position to infer from your new evidence that the proposition is false. So even if you initially have sufficient justification to affirm a self-falsifying proposition, that justification vanishes once you do affirm it, and introspectively come to know that you have.[28]

This proposal has some affinity with introspective accounts of the *cogito*. Like those, it says what you have non-introspective reason to believe or judge is just that if you affirm a given self-falsifying proposition, then it is false. To have reason not to affirm it, you need the further introspectively known premise that you do affirm it.

But this introspective account has implications some philosophers are likely to resist. To many, it will seem that if I am rational, then I will refrain from affirming a self-falsifying judgment in the first place. I won't first affirm it and then immediately regret it upon introspectively realizing that I have done so. If this is right, the introspective proposal seems unable to explain why. For it seems that the introspectionist will only grant me reason not to affirm a self-falsifying proposition once I know by introspection that I did affirm it.

A comparison with assertion might again help. Suppose you have sufficient evidence that you will make no assertions right now. Even if your aim is to speak the truth, you still have reason not to assert this. For you can know that if you were to assert that you make no assertions, you would thereby make a false assertion. Importantly, you do not need to assert first and then immediately regret it, once you realize you have done so. Your reasons against asserting are available antecedently.[29]

If judgments are like assertions, how would this support SELF-FALSIFICATION? Here is a rough suggestion. Just as assertions are actions governed by the aim of truthful assertion, judgments are mental acts governed by the aim of truthful judgment. So, it is irrational for an agent to affirm a proposition which she recognizes must be false if she affirms it. For in general, it is irrational to adopt an action that one knows will fail to meet whatever aims motivate it. Following performative accounts of the *cogito*, we can call this general strategy **performativism**. It will take some filling in as we go. But the idea is to explain SELF-

---

[26] 2010, Sec. 6.

[27] 1988, Ch. 11.

[28] E.g., Shoemaker 1996, Kriegel 2004, Silins 2012 and 2013, and esp. Smithies 2016 and forthcoming.

[29] Cf. Hintikka 1962, pp. 18-19.

VERIFICATION and SELF-FALSIFICATION by appealing to some way in which judgments, like ordinary voluntary action of assertion, are performances of the agent.

Perhaps thinking along these lines is what leads many philosophers discussing Moore's paradox to emphasize a commonality between belief and assertion.[30] But in what way must we take judgments to resemble to voluntary actions like assertions for our account to work? And *should* we take them to resemble assertions or other actions in these ways?

4. Judgments as Mental Performances

On Descartes's **voluntarist** view, judgment is an act of the will, as free and voluntary as any action. An alternative **involuntarist** position takes judgments not to be under the agent's direct control. There might also be room for intermediate positions, which take judgments to be "up to us" in some sense that does not involve ordinary voluntariness.[31]

The performative account might seem to commit us to voluntarism about judgment. But this is not so straightforward. We will see that what it requires is instead that our judgments be subject to the norms of practical reason. And it could be held that judgments are evaluable by practical norms even if involuntary. It might for example be that indirect voluntary influence is enough, or that some other form of control is, or even that judgments can be evaluated as practically rational or irrational regardless of whether we control them.[32]

Still, I agree that matters are most straightforward for the performativist on a voluntarist conception of judgment like Descartes's. This voluntarism might be resisted on the grounds that, first, it commits us to voluntarism about belief, and second, that voluntarism about belief is implausible. But both these assumptions are questionable.

Whether voluntarism about judgment entails voluntarism about belief depends on the relationship between judgment and belief. A **constitutivist** view takes judgments to be metaphysically sufficient for belief. Most naturally, it says that judgments are a kind of belief, or are the conscious manifestations of beliefs, or something like that. An alternative **causalist** view takes judgments to reliably but perhaps fallibly cause beliefs. A final, **effectist** view take judgments to be mere effects of preexisting beliefs. While constitutivism and causalism arguably entail that beliefs are voluntary if judgments are, this is not so for effectism. After all, assertions can be voluntary without the beliefs that cause them being voluntary.

While I am pitching these views as offering competing accounts of a single phenomenon, it may be that there are simply distinct phenomena that are well-suited to being called 'judgments'. For example, an effectist voluntarist might take paradigmatic examples of judgment to include reminding oneself in inner speech that one shouldn't interrupt. An involuntarist constitutivist or causalist might take a paradigmatic judgment to be spontaneously realizing while attempting to lock the front door that one has left one's

---

[30] E.g., Green and Williams 2007, pg. 3; Hájek 2007, pg. 219; Moran 2001, pg. 70; Peacocke 2017; Shoemaker 1996, pg. 78-79; Silins 2012; Smithies 2016 and forthcoming; Williamson 2000, pp. 255-6.

[31] E.g., Moran 2001 and Hieronymi 2006 and 2008.

[32] Cf. Feldman 2000 and Rinard 2017 and 2019.

keys inside. It is not obvious that these are instances of a single mental phenomenon that should be given a unified account. So it may be better to take what follows as giving an account of how certain states that could be called 'judgments' are justified.

Assuming constitutivism or causalism, the voluntarist about judgment plausibly must be a voluntarist about belief. Is that a tenable view? It is often thought not, on the grounds that we cannot believe for just any practical reason.[33] If offered a cash prize for believing the capital of Russia is St. Petersburg, for example, it seems you will not be able to do it. But even if we grant that you cannot believe for financial reasons, this does not obviously show beliefs to be involuntary. For comparison, consider the notion of sincere assertion. One cannot sincerely assert for financial reasons against one's evidence, but that does not automatically show sincere assertions are involuntary. Perhaps they count as voluntary because assertions are voluntary, and sincere assertions are simply a type of assertion individuated by their motivation. Correspondingly, the constitutivist or causalist might claim that judgments are motivation-individuated members of some class of voluntary mental states, such as acts of inner speech. The constitutivist for example might claim that an act of inner speech only counts as a belief when it is motivated in the appropriate way. And the causalist might claim, plausibly enough to casual introspection, that our inner speech can reliably produce belief when properly motivated, such as in explicit reasoning in natural language.

It is not clear to me that these maneuvers ultimately are enough to sustain a voluntarist view of judgment, at least if one is a causalist or constitutivist about judgment's connection to belief. Even if assertion is voluntary, that does not obviously show that the motivation-individuated act of sincere assertion is itself voluntary. And the same might go for judgment on the view that it is motivation-individuated. For comparison, suppose a god has decreed that anyone who worships him out of piety are will be rewarded, but anyone who worships out of greed will be punished. You are not sure whether you are more likely to be motivated by piety or greed, should you worship. Can you simply decide to worship out of piety? Maybe you can, or maybe you instead must decide simply whether to worship, and hedge your bets against the possibility that you will be motivated by greed if you do.

While I have these and other reservations about the constitutivist's metaphysical and the causalist's psychological claims, for today I will let them pass. The constitutivist and causalist disagreement with effectism will make a difference for our discussion of Moorean judgments, but in other cases it will matter less. And it would not matter even for Moorean judgments under some theories of practical rationality, including both evidential decision theory and one I propose elsewhere—though I won't push that point here.[34] For now, we will press on with the provisional result that performativism has contestable but still *prima facie* plausible commitments about the sense in which judgments are performances of the agent.

5. Evidence and Reasons for Judgment

If judgments are in some important sense performances, then perhaps they can be subject to the norms of practical reason. As I will now explain, this gives us a principled

---

[33] Feldman 2000, Hieronymi 2006 and 2008, Kelly 2002 and 2003, and Rinard 2017 and 2019.

[34] For details, see Barnett MS.

reason for rejecting an assumption that made SELF-VERIFICATION and SELF-FALSIFICATION seem puzzling. The assumption is the simple evidentialist view that the rationality of affirming that *p* depends on the evidential probability of *p* itself. Informally, you can think of simple evidentialism as taking judgments to issue from deliberation over the question whether it is the case that *p*, and reasons for judgment to be considerations bearing on this question.[35] But an alternative performative conception instead takes the object of deliberation to be the question whether to affirm *p*, and reasons for judgment to be considerations bearing instead on (roughly) whether it would be the case that *p* if one were to affirm *p*. In special cases involving Moorean self-falsifying and *cogito*-like self-verifying propositions, one's evidence regarding these questions can come apart. The evidential probability of a proposition can be low while the evidential probability that it would be true if affirmed is high, and vice versa. And so SELF-VERIFICATION and SELF-FALSIFICATION are vindicated by performativism, even if they are puzzling under evidentialism.

While I hope this general idea is intuitive, I worry that it, and some further claims I will make below, are too impressionistic as they stand. So I want to show how they can be supported via entirely general and independently motivated theories of practical rationality. For concreteness I will adopt **causal decision theory (CDT)**, but most of the main points won't turn on adopting it rather than other prominent competitors.

CDT holds that an action A is rational iff one has no other option exceeding its **causally expected utility**, U(A), which is defined as follows:

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

Here the *K*s are **dependence hypotheses**—i.e., maximal hypotheses about how outcomes depend causally on one's actions that form a partition. The agent's probability function, *Pr*, can be understood as representing the degree to which the agent's evidence supports various propositions she might entertain. The agent's value function, *v*, is a little tricky. It is normally understood to represent the overall degree to which she values various states of affairs, balancing her aims where they conflict. But performativism is a theory of rational judgment, which might be taken to depend solely on the agent's **alethic aims** of affirming truths but not falsehoods. Otherwise, performativism says that it is rational to make a judgment because it will further one's aim to be happy, for example. So we can here take *v* to be the agent's alethic value function, which represents solely her alethic aims.

One further wrinkle concerns the Jamesian distinction between the competing alethic aims of making true judgments and avoiding false ones. Suppose one considers whether to judge that q. Even holding fixed that one cares solely about the truth and falsity of one's judgments, we still need to specify the relative weightings assigned to affirming truths and avoiding affirming falsehoods. Where *T* is that one affirms a truth and *F* is that one affirms a falsehood, when evaluating the causally expected utility of judging that p, the relevant partition of the dependency hypotheses is $\{J(q) \Rightarrow T, J(q) \Rightarrow F\}$. Thus judging that q is rational iff:

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

---

[35] Hieronymi 2005.

Since the expected utility of withholding is a constant, I hereby set it at 0. Thus (2) reduces to:

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

Since J(q) ⇒ T iff not-[J(q) ⇒ F], (3) reduces to:

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

and therefore to:

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

Thus performativism says that (5) is the condition for rationally judging that q. Note that insofar as one's alethic values affect the rationality of a judgment, what matters is the ratio of the disvalue of false belief to the value of true belief. For simplicity, I will assume this is a constant. But one could allow it to vary between agents, if one follows James' apparent permissivism about how "trigger happy" one should be with judgments, or between contexts, if one wants the threshold for belief to vary with practical stakes. (Even so, this ratio ought always to exceed 1, since otherwise it would be rational to make inconsistent affirmations so as to cover all one's bases.) One could even replace an alethic value function with an epistemic value function, which evaluates beliefs not just by their truth, but by their status as knowledge. This modification might be necessary to accommodate the alleged fact that one should not judge that one's lottery ticket will lose. But I think it is an idle wheel in the explanation of Moore's paradox.[36] So I will assume only concern with the truth of one's judgments.

This ends the preliminaries. The real explanatory work regarding Moorean and *cogito*-like judgments depends on how performativism has the rationality of judgment depend on one's probabilities. In particular, it has the rationality of judging q depend not on the probability of q itself, but instead on the probability that if one were to affirm that q, then one would thereby affirm a truth. This would be unimportant if for any q,

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

But (6) is false for instances of q that are self-verifying or self-falsifying.[37] Take for example the Moorean conjunction *It will rain, but I do not believe it will rain*. Assuming causalism about judgment, one's affirming this conjunction will cause it to be false. Assuming constitutivism, it will constitute (or otherwise suffice for) its being false. Thus the

---

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

[37] If evidential decision theory (EDT) is preferred to CDT, this crucial point can be recast accordingly. Whereas performativism developed using CDT has the rationality of judging that q depend on Pr[J(q) ⇒ *T*], using EDT it has it depend on the conditional probability Pr[*T*|J(q)]. This also differs from Pr(q) in cases of self-verification and self-falsification.

probability of the Moorean conjunction can differ from that of the subjunctive conditional that if one were to affirm it, then one would thereby affirm a truth.

Recall stubborn Stella, whose evidence supports that it will rain, but who knowingly refuses to believe it will rain. Stella's evidence assigns a high probability to the Moorean conjunction *It will rain, but I do not believe it will rain*. But the probability that the conjunction would be true if she affirmed it is still low. And under performativism, it is the latter epistemic probability that matters.

This is how performativism yields the desired result that it is irrational to affirm Moorean conjunctions. The same goes for most other self-falsifying judgments. But what about the special case of the proposition that one is not now in brain state S, where S is identical to the judgment that one is not now in S? Affirming this proposition guarantees it is false, but refraining guarantees it is true. So under performativism, it all depends on the ratio of the disvalue of affirming falsehoods to the value of affirming truths. If I am right that this should always exceed 1, this will vindicate Earl Conee's view that you should refrain from affirming that you are not in S, seeing as this as having the highest expected alethic utility.[38]

Similar points hold for *cogito*-like self-verifying judgments. Suppose I realize that for every number *n*, the proposition that I am now thinking of *n* is self-verifying. It might still be improbable on my evidence that I am thinking of the number 36. But it will be probable that if I were to judge that I am thinking of 36, then I would thereby make a true judgment. Thus on performativism it can be rational for me to judge that I am thinking of 36, despite its being improbable given my evidence. To have justification for the judgment, I do not first need to make it and only then gain introspective justification for it.

The same goes for *I exist*, the most famous example of self-verification. Suppose that in the context of radical doubt, the Meditator lacks any introspectively known evidence or premises from which he can infer that he exists. Even so, if he knows that *I exist* is self-verifying, he can be sure that J(*I exist*) $\Rightarrow$ *T*. (Otherwise, there would have to be a dependency hypothesis with nonzero probability where he judged that he exists but fails to thereby make a true judgment.) And so he can rationally judge that he exists, despite lacking evidence entailing or even making probable his existence.[39]

This account of course involves many details that cannot plausibly be attributed to Descartes himself. But I appeal to CDT largely to emphasize that the basic contours of the account follow from independently motivated general theories of practical rationality. The general idea behind the account does not require adopting CDT, or any competing formal theory. It requires instead only rejecting an evidentialist theory of judgment, which says a judgment is rational only if the proposition judged is probable on one's evidence, in favor of viewing judgment as subject to practical reason.

---

[38] 1987, pg. 326.

[39] This assumes that affirming *I exist* can be among one's option without one knowing that one has it (or anything else) as an option, because one lacks antecedent knowledge that one exists. Brian Hedden (2012) might seem to disagree. But when Hedden says an agent must believe she is able to do something for it to be an option, his real concern is to rule out options that an agent thinks she might fail to perform if she tries. The Cartesian Mediator's predicament instead is uncertainty about whether she even can *try* to affirm her existence.

5. Self-Falsification and Contagion

Suppose we accept the performativist account of Moorean and *cogito*-like judgments offered here. What does it matter?

A common view in discussions of self-knowledge is that *cogito*-like and Moorean judgments are not mere idle curiosities. Instead, these unusual phenomena are just an illustration of a broader way in which our self-knowledge is not receptive or perception-like, but instead grounded in our rationality, or even in the performative character of much of our mental lives.[40]

I think the account offered here casts doubt on some of these more ambitious claims. While I also have the same reservations about the alleged significance of *cogito*-like judgments, here I will focus on Moorean ones, since their alleged significance for self-knowledge has been argued in greater detail. In a nutshell, the idea is that the rational defectiveness of Moorean judgments is in a certain sense **contagious**, spreading their irrationality to failures of self-knowledge in general.

The *locus classicus* for what I am calling contagion comes from Sydney Shoemaker.[41] Shoemaker's targets are theories of self-knowledge that make it out to be too much like ordinary perceptual knowledge. If such views are correct, Shoemaker argued, then just as there can be perceptual impairments like blindness that limit one's perceptual knowledge, there could be introspective impairments like **self-blindness**, the condition of lacking a capacity to know one's beliefs introspectively despite possessing idealized rationality, intelligence, and conceptual sophistication. But according to Shoemaker, self-blindness is impossible. Self-knowledge is unlike knowledge of other contingent matters, in that a rational agent cannot fail to possess it.

This is where the irrationality of Moorean judgments comes in. If Moorean judgments are irrational, Shoemaker and many subsequent authors have argued, that shows self-blindness is impossible.[42] For it shows that:

> (HIGHER-ORDER ERROR) It is irrational to jointly believe that ϕ and that one does not believe that ϕ.

And given HIGHER-ORDER ERROR, we cannot avoid Shoemaker's thesis that self-blindness is impossible. For a rational agent would have no way to avoid higher-order errors if she lacked a capacity for self-knowledge. Or at least, that's the idea.

Now I agree that HIGHER-ORDER ERROR supports the impossibility of self-blindness. What I doubt is that HIGHER-ORDER ERROR follows from the irrationality of Moorean

---

[40] E.g., Burge 2013, Chs. 1-9; Fernández 2013; Moran 2001, pp. 69-77; Setiya 2011; Shoemaker 1996; Smithies 2016 and forthcoming; and Zimmerman 2008.

[41] Shoemaker 1996.

[42] E.g., Fernández 2013; Smithies 2016 and forthcoming; and Zimmerman 2008. See also Barnett MS for critical discussion.

judgments. For I see no way to argue from one to the other without making two assumptions that, while independently defensible, are jointly untenable.

The first assumption concerns the connection between judgment and belief. It says:

> (BELIEF→JUDGMENT) It is rational for one to believe that ϕ only if it is rational for one to judge that ϕ.

Should we accept BELIEF→JUDGMENT? Considered on its own, it seems plausible enough. It does not require that one actually does affirm each proposition that one believes, only that it be rational for one to do so. If one is not in a position to consciously affirm a proposition, it might be thought, one has no business believing it. In fact, a strong constituitivist view might even say that *what it is* to believe a proposition is to be disposed to affirm it.[43] On such views, it is hard to see how the one could be rational without the other.

The second assumption is an instance of a famously contentious principle linking logical entailment and rational belief:

> (MULTI-PREMISE CLOSURE) If ϕ and Ψ straightforwardly entail ω, then if you rationally believe both that ϕ and that Ψ, then it is rational for you to believe that ω.

With these two assumptions, the irrationality of the relevant higher-order errors would follow from the irrationality of Moorean judgments. If it is irrational to affirm the Moorean conjunction *It will rain, but I do not believe it will rain*, then by BELIEF→JUDGMENT it is irrational to believe it. But the conjuncts *It will rain* and *I do not believe it will rain* straightforwardly entail this Moorean conjunction. So by MULTI-PREMISE CLOSURE, it cannot be rational to jointly believe them, as HIGHER-ORDER ERROR holds.

But should we accept MULTI-PREMISE CLOSURE? A familiar theorem of the probability calculus is that if p entails q, then $Pr(p) \leq Pr(q)$. This makes a single-premise cousin to MULTI-PREMISE CLOSURE appealing, at least under the simple evidentialist view that belief in a proposition is rational iff its probability exceeds an invariant threshold. But MULTI-PREMISE CLOSURE famously faces additional difficulties, since, e.g., a conjunction can have a lower probability than each of its conjuncts. Roughly, this is because the conjunction accumulates the error risk of each conjunct. For very long conjunctions, the accumulation of risk can be dramatic, and the probability of the conjunction can be far below that each conjunct. This is plausibly why, if a book contains enough individual claims, you can rationally believe each claim while doubting that *all* the claims are true.[44] But the accumulation is more limited with only two conjuncts, especially if the conjuncts have high probabilities. So anyone who accepts evidentialism plausibly should accept particular instances of MULTI-PREMISE CLOSURE involving only a small number of sufficiently probable conjuncts. Consider Stella, for example, who has ample evidence it will rain, and who knows introspectively that she does not believe it. Where *r* is that it will rain, and *B(r)*

---

[43] E.g., Bach 1984, pg. 48.

[44] Christensen 2004.

that she believes it will rain, nothing prevents us from simply stipulating that Pr[r] ≈ Pr[B(r)] ≈ 1, and therefore that Pr[r & B(r)] ≈ 1.

This is where this paper's performitivist account of Moorean judgments comes in. In contrast to evidentialism, performativism admits quite dramatic failures of MULTI-PREMISE CLOSURE, at least assuming BELIEF→JUDGMENT. The reason is closely related to how performativism vindicated SELF-FALSIFICATION, by divorcing the rationality of affirming a proposition from that proposition's probability. Even if it is probable that if one judged p one would affirm a truth, and that if one judged q one would affirm a truth, it can still be improbable that if one judged that p and q, then one would affirm a truth. For example in Stella's case, even though

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

and

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

it still is true that

$$(9)\quad \Pr\left[J(r \,\&\, \sim B(r)) \Rightarrow (r \,\&\, \sim B(r))\right] \ll 1.$$

The upshot is that the irrationality of self-falsifying judgments is not contagious. It can be rational to affirm *It will rain* and rational to affirm *I don't believe it will rain*, without its being rational to believe the Moorean conjunction *It will rain, but I don't believe it will rain*. This might mean that we should reject MULTI-PREMISE CLOSURE, and say it can be rational to believe the Moorean conjuncts without its being rational to believe their conjunction. Or it might mean we should reject BELIEF→JUDGMENT, and say it can be rational to believe the Moorean conjunction without its being rational to affirm it. But either way, the inference from MOORE to HIGHER-ORDER ERROR is blocked.

Now one worry here is that too much of this rests on performativism, or my version of it. Perhaps on a better view the irrationality of Moorean judgments will be contagious after all. But I think this is a mistake. It is a *further advantage* of performativism that it offers an explanation of contagion failures. Their existence can be independently motivated by considering examples like:

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

If we accept SELF-FALSIFICATION, we must say that it is irrational for Robin to judge that he is not thinking of Nena. But surely it is rational for Robin to jointly believe each of two premises that straightforwardly entail he is not thinking of Nena. So either it is rational for Robin to believe what he cannot rationally judge, or else irrational to believe the obvious consequence of premises he rationally believes. The upshot is that given SELF-FALSIFICATION, we must reject BELIEF→JUDGMENT or MULTI-PREMISE CLOSURE. Either

way, the case for HIGHER-ORDER ERROR will be undermined, and the irrationality of self-falsifying judgments won't be contagious.

6. Conclusion

Epistemic deliberation as usually understood is concerned with what is going on in the world. A rational deliberating agent considers some proposition *p*, and considers evidence regarding whether it is the case that *p*. But practical deliberation is concerned with how to intervene in the world. What matters instead is evidence regarding what would be the case if this or that intervention were made. In the special case where the intervention is one's affirming the proposition *p*, what matters is evidence regarding not the question whether it is the case that *p*, but the question whether it would be if one were to affirm it. The performative account of *cogito*-like and Moorean judgments exploits this gap between epistemic deliberation as usually understood and practical deliberation about whether to make a judgment. Where self-verifying and self-falsifying propositions are concerned, your evidence regarding whether they are true can point one way, while your evidence concerning whether they would be true if you affirmed them points another way. Subjecting judgments to norms of practical reason thus vindicates common intuitions about the rationality of *cogito*-like judgments and the irrationality of Moorean ones, even as it undermines further claims about their broader significance for self-knowledge.

References

Bach, K. 1984. 'Default Reasoning: Jumping to Conclusions and Knowing when to Think Twice' *Canadian Journal of Philosophy* 65: 37-58.

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

Broughton, Janet. 2002. *Descartes's Method of Doubt.* Princeton: Princeton University Press.

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

Carriero, John. 2009. *Between Two Worlds: A Reading of Descartes's* Meditations. Princeton: Princeton University Press.

Christensen, David. 2010. 'Higher-Order Evidence' *Philosophy and Phenomenological Research* 81(1): 185-215.

Conee, Earl. 1982. 'Evident, but Rationally Unacceptable' *Australasian Journal of Philosophy* 65: 316-326.

Curley, E. M. 1978. *Descartes Against the Skeptics.* Cambridge: Harvard University Press.

Feldman, Fred. 1973. 'On the Performatory Interpretation of the Cogito' *Philosophical Review* 82(3): 345-363.

Feldman, Richard. 2000. 'The Ethics of Belief' *Philosophy and Phenomenological Research* 60(3): 667-695.

Fernández, Jordi. 2013. *Transparent Minds,* OUP.

Frankfurt, Harry G. 1966. 'Descartes's Discussion of His Existence in the Second Meditation' *Philosophical Review* 75(3): 329-356.

Frankfurt, Harry G. 1970. *Demons, Dreamers, and Madmen: The Defense of Reason in Descartes's* Meditations. Reprinted 2008, Princeton: Princeton University Press.

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

Hedden, Brian. 2012. 'Options and the Subjective *Ought*' *Philosophical Studies* 158(2): 343-360.

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

———. 2006. 'Controlling Attitudes' *Pacific Philosophical Quarterly* 87(1): 45-74.

———. 2008 'Responsibility for Believing' *Synthese* 161(3): 357-373.

Hintikka, Jaakko. 1962. '*Cogito, Ergo Sum:* Inference or Performance?' *philosophical Review* 71(1): 3-32.

Kelly, Thomas. 2002. 'The Rationality of Belief and Some Other Propositional Attitudes' *Philosophical Studies* 110(2): 163-196.

———. 2003. 'Epistemic Rationality as Instrumental Rationality: A Critique' *Philosophy and Phenomenological Research* 66(3): 612-640.

Kenny, Anthony. 1968. *Descartes: A Study of His Philosophy*. New York: Random House.

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

Markie, Peter. 1992. 'The Cogito and Its Importance' *The Cambridge Companion to Descartes*, ed. John Cottingham, Cambridge: Cambridge University Press.

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

Pryor, James. 2005. 'There is Immediate Justification' *Contemporary Debates in Epistemology*, Matthias Steup and Ernest Sosa (eds.), Blackwell.

———. 'Hyper-Reliability and Apriority' *Proceedings of the Aristotelian Society* 106(3): 327-344.

———. MS. 'More on Hyper-Reliability and a Priority'

Rinard, Susanna. 2017. 'No Exception for Belief' *Philosophy and Phenomenological Research* 94(1): 121-143.

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

Setiya, Kieran. 2011. 'Knowledge of Intention' *Essays on Anscombe's* Intention, eds. Anton Ford, Jennifer Hornsby, & Frederick Stoutland. Harvard University Press.

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

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

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

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

Williams, Bernard. 1978. *Descartes: The Project of Pure Enquiry*. London: Penguin.

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