Final version published in the 2021 *Proceedings and Addresses of the APA.* I'll be happy to email you a pdf of the published version if you'd like: david.christensen@brown.edu.

Rationality for the Self-Aware[1]

David Christensen
Brown University

Introduction

Giving a lecture named for Ernest Sosa is a great honor—and an intimidating one. I'm also challenged by the description of the lecture I'm supposed to give, which emphasizes that it should be accessible to a broad philosophical audience. So I will try here to suppress my tendency to get sucked into technicalities that would only interest a small subset of epistemologists, and I'll take a somewhat big-picture view.

One of Sosa's major contributions to epistemology—a cornerstone of his overall view— is his distinction between *animal knowledge* and *reflective knowledge*. Animal knowledge results from our using reliable faculties, like perception, to form beliefs. Reflective knowledge only comes into play when agents can take a *perspective* on their own believing. In particular, it comes into play when agents can consider their own reliability in forming beliefs. Sosa argues at length that reflective knowledge is especially valuable, and that it deserves special attention in our epistemic theorizing. And Sosa's work clearly shows us the theoretical richness that flows from focusing on reflective knowledge.[2]

Now, I myself have never given serious thought to knowledge; I've always been drawn to just thinking about rational belief. But I have become increasingly convinced that it makes a huge difference, when we think about rational belief, whether or not we take into account the agent's ability to reflect on herself as a believer—and, in particular, her ability to reflect on her own reliability as a believer. I'm also convinced that this sort of self-awareness deserves special attention in our epistemic theorizing, and I will try to make a case for that here. I will draw on work that I and others have been doing, in order to illustrate some of the theoretical richness that emerges from thinking about self-aware agents. I believe that taking self-awareness into account yields a picture of rational belief that is surprising, in a number of different, but interconnected, ways.

The richness I'll focus on emerges most clearly in cases that involve so-called "higher-order evidence." As I'll use that term, it will refer to evidence an agent may get that bears directly on the reliability of some part of her own thinking. So it's the kind of evidence that becomes important only when an agent can take a perspective on her own beliefs. Some of the cases that prove most interesting involve agents getting evidence *against* the reliability of some part of their own thinking.

This can happen, for example, in cases of disagreement—in particular, in cases where I find out that I disagree with someone who has (at least roughly) the same evidence that I have, and whose thinking I have good reason to respect. Their disagreement would generally seem to be evidence against the reliability of my own thinking about the disputed matter. So disagreement provides one instance of higher-order evidence.

---

[1] I would like to thank the donors who made the Ernest Sosa Lecture Prize possible, and the APA Committee on Lectures, Publications, and Research for asking me to give this talk. Thanks also to Zach Barnett, for valuable discussion and comments on an earlier draft. And finally, many thanks to Ernest Sosa, for his many contributions to philosophy.

[2] See, for example, Sosa (1985, 1997).

And there are plenty of other ways I can get evidence against the reliability of some part of my own thinking. I might get psychological evidence that I'm sexist, and so I'm likely to under-rate the CVs of female job applicants. Or, I might get evidence that I'm biased toward my own children, and so I'm likely to over-rate their accomplishments. Or I could get evidence that I'm sleepy, or drunk, or otherwise drugged, in ways that would mess up my thinking about complicated matters.

When a person gets this sort of evidence about some of their thinking, it raises this question: What is the rational way of accommodating that evidence? That's the question that brings out the interesting complexities. And the reason that answering that question gets tricky derives from the element of self-reflection we have in these cases. The agent has to act, in effect, as judge in her own case. She has to accommodate evidence that some of her thinking is unreliable. But she obviously has to do that by using her own thinking. And the necessity of the agent's playing this double role—both judge and judged—leads to some surprising consequences.

1. Taking Higher-Order Evidence Seriously

It will be useful to start with a concrete example; I'll use a phenomenon that's been discussed quite a bit in the literature. *Hypoxia*, or oxygen-deprivation, often occurs at high altitudes, e.g., in mountain-climbers or people in unpressurized airplanes. As people become hypoxic, their judgment deteriorates—their thinking becomes unreliable. And it's particularly dangerous (writers often call it "insidious") because its victims don't feel impaired. Often, in fact, they feel particularly good. And naturally, this has led to a number of unfortunate occurrences on mountains and in airplanes.

With that in mind, consider the following case:

> **Fuel Levels**: Amelia is flying an unpressurized plane, and wondering if she has enough fuel to reach Sitka, which is a bit further away than she'd planned to fly. She knows the relevant formula for figuring out whether she has enough fuel, on the basis of her instrument readings. (We can even imagine that she has the formula written down in front of her.) Amelia looks at her instrument-readings, which are clearly visible, and she does a little mental calculation. On the basis of this figuring, she comes to a confident belief that the instrument readings indicate she has enough fuel to reach Sitka. And on the basis of that, she becomes confident that she does have enough fuel. Then, Amelia notices that her altimeter shows that she's at 11,000 feet, and she knows that at 11,000 feet, there's a high risk of hypoxia.

In this case, Amelia has strong higher-order evidence suggesting that her thinking about fuel-levels is unreliable. And to make the case most interesting, let us add one more detail: Amelia's initial thinking about fuel-levels is, in fact, perfect. Amelia has not actually been affected by hypoxia. So her instrument-readings in fact do indicate that she has enough fuel to reach Sitka.[3]

There has been a fair amount of debate about cases like this. I, and others, have argued that it would not be rational for Amelia to remain fully confident that she has enough fuel—even though her calculations are, in fact, correct. Others would hold that it would be rational for Amelia to remain fully confident, despite her evidence about hypoxia. If you take that position, maybe you'd even want to say that it would be rational for Amelia to act on her confidence, and start flying off toward

[3] My use of hypoxia, and the pilot example, are based on Elga (ms).

Sitka without a further care! That's a battle I won't try to fight here. Instead, I want to explore what seem to me to be some interesting and surprising upshots of holding that Amelia is rationally required to become less confident that she has enough fuel—and, more generally, that in many cases where agents have strong higher-order evidence suggesting that their thinking about a particular matter is unreliable, rationality will demand that they revise their confidence about the matter in question. I'll refer to accounts taking stance as requiring agents to "take higher-order evidence seriously."

2. First Upshot: The Structure of Evidential Support

The first upshot begins with a puzzle about evidence—in particular, how different parts of an agent's evidence interact. I'd like to concentrate on Amelia's attitude toward a particular claim—the claim that *these instrument-readings indicate that she has enough fuel*. I'll suppose that, given the formula Amelia has, it's a fairly simple mathematical fact that the instrument-readings indicate that she has enough fuel. So with this in mind, we might ask: how can the inductive evidence about hypoxia even be relevant? After all, the instrument-readings' indication that she has enough fuel is not mediated by some claim about Amelia being clear-headed. If Amelia's ordinary evidence were simply "weighed against" her higher-order evidence, it would seem that she could reason as follows:

"Well, as the formula shows, my instrument-readings indicate that I have enough fuel. And that's exactly the result that I calculated. So, although I am at 11,000 feet, hypoxia has not messed up my accuracy on this matter, at least!"

That train of reasoning would support Amelia remaining extremely confident in having enough fuel. But it strikes me as blatantly irrational. It relies crucially the kind of mathematical reasoning that Amelia's higher-order evidence makes her likely to get wrong. So when we model the way these two parts of Amelia's evidence interact, we need to do it in a way that doesn't allow the incredibly strong first-order support to swamp or defeat the higher-order evidence.

I'm inclined here toward what's become known as an Independence requirement. In Amelia's case, it would come to something roughly like this: When Amelia takes a perspective on her own thinking, in order to judge how reliable it is, she should do this in a way that does not depend on a train of reasoning that's targeted by the higher-order reasons for doubt. The independence requirement disallows the irrational train of reasoning we were just thinking about. This explains why Amelia is rational to doubt that that she reasons reliably about fuel-levels. And given that doubt, we can see why she should not be confident that she has enough fuel.[4]

This is where the first surprising implication of self-awareness comes in. We're not denying that Amelia has the evidence of the instrument-readings. And she also has the formula that they feed into. And Amelia can even see clearly how the math goes! So in some sense, it seems like she has the strongest possible support for confidence that her instrument-readings indicate she has enough fuel. But our model says she's not rational to be confident in that, after all!

Notice that this surprising result is closely tied to Amelia's taking a perspective on her own thinking. Suppose we consider a different agent—one who has no reason to doubt their own thinking. And suppose we give them all of Amelia's evidence: we show them the instrument-readings and the formula, and we give them the evidence about Amelia's possible hypoxia. Surely they would be perfectly rational to be highly confident that the instrument-readings indicate that Amelia has enough fuel! So it's Amelia's self-awareness that puts her in this predicament: She's

[4] The need for an Independence principle is defended in detail in Christensen (2018). An attempt at formulating such a principle much more carefully than is done here can be found in Christensen (2019).

acting as judge in her own case. And to avoid begging important questions about her own reliability, she has to, in some sense, not give some of her evidence its ordinary due.[5]

3. Second Upshot: Epistemic Dilemmas

The second upshot I'd like to discuss is closely related to the first; it involves the possibility of so-called *epistemic dilemmas*. It's natural to think that rationality puts requirements on agents—and here, I'm thinking about ideal, epistemic rationality. And it's natural to think that all the rational requirements must be simultaneously satisfiable. If they're not—if rational requirements conflict—then even the most rational conceivable agent, would sometimes be doomed to violate some rational requirement. In other words, there would be dilemmas for epistemic rationality.

Now before going further, I should make one thing clear about terminology. For some people, conflicting requirements do not suffice for a dilemma; they hold that, in a dilemma, there must be no best option for the agent. I won't use the term that way. I'd like to separate the question of whether requirements can conflict from the question of whether there are situations with no best option. And so I won't assume that the best option automatically satisfies all requirements. And I'll allow "dilemma" to cover cases where requirements conflict, even if there is a best option. To put it another way, I'll leave room for what Rosalind Hursthouse, in discussing moral dilemmas, called "resolvable dilemmas". Those are cases where there is a best option for the agent, but where that best option involves violating a requirement. As Hursthouse puts it, even the best option can sometimes involve a "remainder" or "residue."[6]

On this understanding, it's plausible that self-awareness can put agents in epistemic dilemmas. To see how, we can go back to Amelia in her airplane, and look at a slight variation of our example. Suppose that, instead of doing calculations about fuel, she's doing a little truth-functional deductive reasoning.

> **Logic in Flight:** Before Amelia took off, her reliable friend handed her a note with two facts about some people they know:
>
> **A:** Karla was born in May if and only if Kayla wasn't.
>
> **B:** Either Kayla, or Layla and Lola, the Lumpkin twins, were born in May.
>
> During the flight, Amelia looks at the note, and rationally becomes highly confident in A and B. Then, to amuse herself for a while, Amelia ponders these bits of information. After a bit of logical reflection, she realizes that they entail:
>
> **C:** Karla was not born in May unless Layla Lumpkin was.
>
> So Amelia comes to have high confidence in C, as well. Then, as before, she notices the altimeter, and realizes that she's likely to be hypoxic. And she knows that hypoxia is likely to degrade the sort of complex truth-functional thinking she just did in deriving C from A and B—even if the thinker feels perfectly lucid.

Now, it seems to me that Amelia is rationally required to lose considerable confidence in C. If she remained fully confident in C, she would not be taking seriously the evidence that her complex truth-functional reasoning was likely impaired. But I also think it's plausible that logic puts rational constraints on Amelia's degrees of confidence—especially if we're asking what levels of confidence would be ideally rational in Amelia's situation. Probabilistic coherence is a plausible requirement on ideally rational confidence. But probabilistic coherence would require Amelia to be at least as

---

[5] This phenomenon is discussed in more detail in Christensen (2010).

[6] See Hursthouse (1999, ch. 2).

confident in C as she is in the conjunction of A and B. And this won't happen, if she loses confidence in C due to doubts about her complex truth-functional thinking. Those doubts don't give her any reason to lose confidence in A or B. If she remains confident in A and B, but loses confidence in C, she'll violate probabilistic coherence.

So this is the (apparent) epistemic dilemma: If Amelia remains fully confident in C, she'll violate a requirement to take her higher-order evidence seriously. But if she loses confidence in C, she'll violate probabilistic coherence. Now, I think that the second option is pretty clearly the more rational one. But I think we should also acknowledge, that it's a resolution that leaves a residue. There's something rationally imperfect about being confident in A and B, and not confident in C.[7]

Some philosophers find this result unpalatable. And people have made various proposals that are motivated, at least in part, by the desire to avoid dilemmas. One proposal is to deny that rational beliefs are sensitive to higher-order evidence.[8] Another is to deny that epistemic rationality is constrained by evidence at all.[9] Another is to distinguish between two senses of rationality, and hold that probabilistic coherence is part of one, and respect for higher-order evidence is part of the other.[10] And still another is to hold that rationality goes indeterminate in situations such as Amelia's.[11] I won't take space here to complain about all of these proposals. But I would like to lobby for a different sort of reaction. And that is, that epistemic dilemmas be seen as an interesting discovery, not as a problem to be avoided.

The first thing to notice is that rationality-evaluations don't need be tied to any kind of blameworthiness. I think that's clear from examples having nothing to do with dilemmas. For example, I take it that there's no sense in which someone living with delusional schizophrenia is blameworthy for their irrational beliefs. So when we say that Amelia violates a requirement of rationality by losing confidence in C, we're not implying that she's somehow blameworthy for believing as she does.

The second thing I'd like to suggest is that epistemic dilemmas are actually a natural outgrowth of self-awareness. Once an agent can reflect on her own thinking, she can get empirical evidence about its reliability. And rationality will require her to take that evidence seriously. But empirical evidence will sometimes be misleading. Even if some part of an agent's thinking is perfectly good, they might get evidence which suggests that it isn't. That's what happened to Amelia: She got evidence that her truth-functional reasoning was likely defective, even though it actually wasn't. But rationality will require that agents respect their evidence, whether or not it's actually misleading. After all, misleading evidence just is evidence that rationally speaks in favor of believing things that are in fact false.

So, rationality constrains the beliefs of self-aware agents in two different ways: In a first-order way, it requires agents to reason correctly, for example, by respecting the logical relations among propositions. In Amelia's case, it requires her to be confident in C if she's confident in A and B. But when agents can take a perspective on their own believing, more requirements come into play. In a higher-order way, rationality requires self-aware agents to take account, in their beliefs, of cognitive errors that the evidence suggests they're likely to make. (In Amelia's case, she's required

---

[7] A full discussion of epistemic dilemmas, based largely on this example, is in Christensen (forthcoming).

[8] Positions of this type are put forth in Lasonen-Aarnio (2014) and Titelbaum (2015).

[9] See Worsnip (2018).

[10] See Smithies (forthcoming).

[11] See Leonard (2020).

not to be fully confident in C on the basis of A and B, because she should not trust her own ability to do this sort of derivation.)

It seems natural to expect these two sorts of requirements to come into conflict. When an agent gets evidence that her first-order thinking about some matter has been compromised, then she cannot both believe as her first-order evidence requires, and also refrain from believing in ways that are likely—given her higher-order evidence—to be inaccurate. So, whatever the agent believes, her beliefs will violate some principle that in general helps characterize ideal epistemic functioning. And this suggests that we should see epistemic dilemmas as a natural outgrowth of the richness in normative structure that comes with the territory, when we theorize about agents who can take perspectives on their own thinking.

4. Third Upshot: Rational Epistemic Akrasia

The next upshot I'd like to discuss arises when we theorize about agents who can take a somewhat different kind of perspective on their own thinking—maybe a more sophisticated kind of perspective. So far, we've been looking at agents who can consider the reliability of their own thinking, which is a descriptive matter. But we might also want to think about agents who can take a normative or evaluative perspective on their own thinking. For example, somewhat sophisticated agents might think of their own beliefs as rational, or irrational. I take it that all of us members of the APA are at least this sophisticated. And I suspect that really most humans—or at least most adult humans—are.

Once agents can think about the rationality of their own beliefs, new questions arise when we theorize about them. For instance, what is the relationship between:

(A) what an agent is rational to believe about the world in general; and

(B) what she is rational to believe about what she's rational to believe?

How does the question of whether she's rational to believe P relate to the question of whether she's rational to believe that she's rational to believe P?

It's very attractive to think that the answers to these two questions have to line up in a certain way. In particular, it's attractive to think that it just can't be rational for an agent to believe P, while also believing that P is not rational for her to believe! That would involve what's become known as epistemic akrasia. And epistemic akrasia can seem obviously irrational.

Now, there are lots of different ways of formulating enkratic principles—ones that prohibit akrasia. Some involve beliefs about what beliefs are rational, and others involve beliefs about what beliefs are justified, or what beliefs one's evidence supports, or what beliefs one ought to have. Some treat beliefs as categorical; others apply to degrees of confidence. I'll stick to thinking about judgments of rationality. And I'll talk about degrees of confidence, or credences, rather than categorical beliefs, just because I think that comes closer to carving epistemic nature at its joints. But I think that the main lessons here will carry over to different ways of setting up the issue. In this setting, a very rough enkratic principle might say something like this:

> **Rough Enkratic Principle**: It cannot be rational to have high credence in P, and also have high credence that high credence in P is not rational in one's present situation.

Obviously, this would need to be made much more precise and general, in various ways, but I don't think that will matter for present purposes.

Let's look at why enkratic principles for rational belief seem so attractive. I think that one of the main motivations comes from thinking about certain paradigmatic cases of epistemic akrasia. So here's one, adapted from Sophie Horowitz's paper on the topic: Sam is a detective, working to

identify a jewel thief. He works late into the night, analyzing the evidence closely, and he becomes highly confident that the thief is Lucy. Then his partner Alex tells him that he almost always interprets forensic evidence wrong, when he works late at night. Sam had never realized this—he doesn't keep track. But Alex is very reliable, and Sam has very good reason to trust her testimony. Now suppose Sam reacts to all this evidence in the following way: First, he remains highly confident that Lucy is the thief. But second, he becomes highly confident that high confidence in Lucy's guilt is not rational in his situation: After all, he has probably misinterpreted the forensic evidence, and so the evidence probably doesn't support thinking that Lucy is the thief!

Sam's beliefs here are intuitively highly irrational. And, as Horowitz argues in detail, they would also lead to further irrational beliefs, and to irrational actions. So this paradigm case of epistemic akrasia seems clearly irrational.

Nevertheless: I would argue that epistemic akrasia turns out, in the end, not to be irrational at all! Sam's akratic beliefs are surely irrational—but I would argue that they're not irrational in virtue of being akratic. In other cases, I think it turns out that epistemic akrasia is actually rationally required! There are some complex cases in the literature, including some from Horowitz and a couple from me.[12] But I'd like to focus on a particularly simple case, which I'll adapt from a recent paper by Zach Barnett. It involves someone who gets misleading evidence about what the correct theory of epistemic rationality is.[13]

Suppose that Dara is a college student, studying epistemology. He's taken a bunch of epistemology courses from the professors at his college. And they're all fervent followers of Hume—or, at least, of Hume as they understand him. So they all push a particular Hume-inspired account of rational belief. It's built around a principle I'll call Deductive Purism, which says that inductive reasoning is not a rational way of supporting beliefs. So, for example, if you're wondering whether the sun will rise tomorrow or not, or if you're wondering whether the next bread you eat will nourish you or poison you, Deductive Purism says that it's not rational to think either outcome more likely, just because of what's happened in the past. The view acknowledges that inductively-supported beliefs are generally *accurate*. But: it holds that rationality is not just about accuracy. Rational beliefs must be supported by the right kind of reasons. And only deductive reasoning can render *rational* support.

Now I think that most of us will see Deductive Purism as pretty clearly wrong. I take it that even Popperians will reject it, as I've set it out. But I think that, if we fill in the details right, Dara would be rational to be quite confident that Deductive Purism was correct. After all, his professors have persuasive, sophisticated arguments. They're clearly very intelligent. They express high confidence in Deductive Purism. And, most importantly, they are recognized experts about epistemic rationality! So if we tell the story right, I think that the most rational thing for Dara to believe about Deductive Purism is that it's very likely to be correct.

Now, suppose Dara begins to reflect on his own beliefs—and in particular, on his belief that the sandwich he packed for lunch will nourish him, and not poison him. He realizes that his confidence that it won't poison him is based on inductive evidence. And he gets a bit worried. So he goes to his favorite professor's office hours, and asks, "Can't I even rationally believe that my sandwich won't poison me? I mean, otherwise, shouldn't I play it safe, and just throw it away?"

Dara's professor sits back in his chair, sighs, and smiles condescendingly. "Oh, Dara. Please don't go throwing away your lunch! No one ever told you that inductively supported beliefs are not

---

[12] See Horowitz (2014, 2019) or Christensen (2016).

[13] See Barnett (2020); I make use of this example, along the lines pursued in this section and the next, in Christensen (2020). A somewhat similar example is developed in Weatherson (2019, p. 170 ff.)

accurate, by and large. Of course, they're accurate! It's just that these beliefs are not *rational!* After all, there's no non-circular way to justify induction…right?"

In this case, it seems pretty clear that Dara is rational to remain highly confident that his sandwich will nourish him, and not poison him—just like all the other sandwiches he's eaten over the years. But it also seems that he's rational to be confident that Deductive Purism is the correct account of rationality. And so he's rational to be confident that his high credence that the sandwich won't poison him is not rational. And if that's right, it's rational for Dara to be epistemically akratic. So it turns out—surprisingly enough—that our enkratic principle for epistemic rationality is wrong. And while I've of course tailored the example to our particular Rough Enkratic Principle, I think it's clear that similar examples can be worked out for many other principles forbidding epistemic akrasia.

Now, I don't want to just leave it at that. There is something strange, at least at first, about akratic beliefs. It might seems like akratic beliefs just couldn't make sense from the agent's own perspective.[14] After all, if the agent thinks that their own degree of confidence in some claim is irrationally high, well, then, why don't they become less confident? That certainly seems to be what would be rational in detective Sam's situation: after hearing from Alex, it does seem that the most rational response for Sam has to involve becoming way less confident that Lucy is the thief.

But I actually think that, in cases like Dara's, akratic beliefs can make sense, even from the agent's own perspective. The reason for this was first brought out by Horowitz, in her discussion of a more complicated case where akrasia seems to be required. The reason is this: in certain cases, it can be rational for agents to expect that rationality and accuracy will diverge in their present situations. In Dara's case, it's the nature of Deductive Purism that delivers this result. On Deductive Purism, inductively-supported credences tend to be accurate, but irrational. And Dara gives high credence to Deductive Purism. So it makes perfectly good sense, from his perspective, to be confident that his sandwich will nourish him, even though he thinks that this confidence is irrational. Irrationality, from this perspective, is not, in this case, an indication of inaccuracy. And that's the key difference between Dara's case and Sam's. In Sam's case, his sleepiness would be expected to lead to irrationality and inaccuracy together. So for him, irrationality *would* be an indication of inaccuracy.[15]

Now, one line of objection to all of this would question whether the most rational attitude for Dara to take really would be high confidence in Deductive Purism. The argument for akrasia depends on the claim that the most rational response to Dara's situation is to give high credence to a certain *false* theory of rationality. And it might be thought that there was something suspicious about this.

One worry that one might have along these lines is that Deductive Purism is just too wacky a theory for anyone to rationally believe. What might seem suspect, I suppose, is the sharp split it

---

[14] In Weatherson's example, for instance, the agent "can't bring herself" to have the attitude she thinks would be rational.

[15] As David Alexander pointed out in the Q&A following this talk, the point made above about Dara's attitude toward his sandwich would plausibly also apply to Dara's attitude toward Deductive Purism itself. Suppose that Dara takes his teachers' testimony as providing non-deductive support for Deductive Purism. Given his confidence that only deductive reasoning can provide rational support, he should presumably conclude that his high confidence in Deductive Purism is probably irrational. Is this particularly problematic?

I would argue that if we keep in mind the points above about how Deductive Purism separates accuracy from rationality, it is not particularly problematic. Dara should think that the irrationality of his expert-testimony-based confidence in Deductive Purism does not mean that Deductive Purism is less likely to be true, just as the irrationality of his inductively-based belief about his sandwich does not mean that the sandwich is less likely to nourish him.

embraces between rationality and accuracy. But it's worth pointing out that quite a few theories of rational belief that are defended by skillful, intelligent philosophers incorporate this sort of split. Some give importance to the practical advantages of holding a belief.[16] Some claim that categorical beliefs can't be justified by purely statistical evidence—no matter how probable it makes the proposition in question.[17] Some claim that moral factors—in addition to truth-relevant ones—are important determinants of epistemic rationality.[18] And related claims are made about cases where beliefs figure in personal relationships[19], or our ability to make sincere promises[20]. So even though Deductive Purism is a bit of a cartoon, I don't think it's dramatically different in kind from views that are currently being defended by very respectable philosophers.

Still, there is a more interesting way of putting pressure on the claim that the most rational response to Dara's situation is to have high confidence in Deductive Purism.[21] It begins by supposing that questions about the rules of rationality are *a priori* questions, just like questions about logic or math. So Dara would presumably have *a priori* rational support for rejecting Deductive Purism, in favor of—well, whatever the true theory of rational belief turns out to be. We might even suppose that if Dara is perfectly rational, Deductive Purism will seem false to him, when he considers it directly.

Suppose we grant a picture along these lines. Would that mean that empirical evidence—and, in particular, higher-order evidence—could not undermine this *a priori* support? I don't think so. At least not if we're on board with thinking that evidence about hypoxia can undermine the *a priori* support Amelia has for thinking that A and B entail C. The upshot of taking higher-order evidence seriously—in a way that involves incorporating some sort of Independence principle into our account of rational support—is precisely that higher-order evidence does not simply get weighed up against the agent's other evidence. And, in particular, higher-order evidence allows for inductively-supported doubts about an agent's cognitive reliability to undermine even deductive support relations between an agent's evidence and particular claims that the evidence in fact entails. Once we take higher-order evidence seriously, we should acknowledge that even if it is *a priori* that some particular theory of rationality is correct, still, people may be rationally misled about the matter.[22]

Now I should take note of one caveat here: If Dara has *a priori* support for rejecting Deductive Purism, then one might well argue that there is still be some rational imperfection in Dara's confidence that Deductive Purism is correct—just as there some rational imperfection in Amelia's losing confidence in C. We saw earlier that it's plausible that higher-order evidence can have just this sort of effect. So I am not claiming here that epistemic akrasia can occur without there

---

[16] See Nozick (1993).

[17] See Nelkin (2000), Buchak (2014).

[18] See Moss (2018), Schroeder (2018), Basu (2019). See Gardiner (2018) for critical discussion.

[19] Sarah Stroud (2006) argues that we should take this sort of view seriously.

[20] Berislav Marušić (2015) argues that the rationality of certain beliefs depends on non-evidential factors. He does not call this epistemic rationality, but takes it to be the only sort of rationality that apples to the relevant beliefs. See Brinkerhoff (forthcoming) for critical discussion.

[21] See Titelbaum (2015) for an argument along roughly these lines.

[22] I'm using the term "undermine" non-technically here. I don't mean to imply that higher-order evidence works in the same way as what are standardly termed "undermining defeaters".

being some remainder, or residue, of rational imperfection. What I do want to argue, though, is that the most rational possible doxastic response, in certain evidential situations, involves epistemic akrasia.

5. Fourth Upshot: Self-Undermining Theories of Rationality

The next upshot I'd like to bring up is also related to the phenomenon of agents being misled about what the correct theory of rationality is. This is a topic that has come up mostly in the literature on disagreement. Almost all accounts of disagreement require some degree of *conciliation*, at least in many cases of disagreement when you have independent reason to think that the other people have a good chance of getting the disputed matter right. By 'conciliate' I mean, moving one's belief in the direction of those who disagree. So, for example, on almost any account of disagreement, the disagreement of large numbers of experts can require one to lose significant confidence in one's initial belief.

Now of course, this applies even if the topic of disagreement is one's theory of rationality. So, as various people have pointed out, Conciliatory theories of disagreement will self-undermine, in one, very particular, sense: In certain cases—for example, when a Conciliationist philosopher meets disagreement from non-Conciliationist philosophers she respects—Conciliationism will say that she's not rational to maintain full confidence in Conciliationism. And this self-undermining phenomenon extends to other accounts of rationality that take higher-order evidence seriously.

Many people have found this kind of self-undermining to be a mark against theories that allow it. There are a couple of different ways of pressing this sort of worry. The one I think is most interesting argues that if a theory of rationality is self-undermining in any circumstance, then it's inconsistent, and thus incorrect.

The inconsistency worry was developed really nicely by Adam Elga.[23] He was an early advocate of a Conciliatory view of disagreement which said that when you disagree with someone, and you have good, independent reason to think they're equally likely to get the disputed kind of matter right, then you should give the two opinions equal weight. And if you disagree with someone you rationally think is more likely to get the disputed sort of matter right—or perhaps with a whole bunch of such people—then you should become more confident in their view than in your initial view.

So suppose that Connie has been studying Epistemology, and from her study of the arguments, she has become confident in a couple of positions. First, she has become confident in Conciliationism. Second, she has become confident in Internalism about epistemic justification. But one day, Connie comes across the new study from *PhilPapers*. She starts looking through the results, and is surprised to find out out that a large majority of professional epistemologists reject Internalism, in favor of Externalism. So Connie, as Conciliationism would require, revises her opinion so that she now thinks that Internalism probably isn't true. Then she reads in *PhilPapers* that a large majority of epistemologists reject Conciliationism as well!

Now, since this is a purely fictional example, I'll fill in the details in a somewhat cartoonish way, to make the inconsistency argument as clear as I can. So let's suppose that *PhilPapers* reveals that most epistemologists support a theory of rationality based on autonomy. On the Autonomist view, it's irrational to conciliate with others who share one's evidence. Rationality requires a certain sort of intellectual self-determination, or thinking for oneself. On this view, the rational belief is the one seems right, given one's own most careful consideration of the direct evidence and arguments.

[23] See Elga (2010).

So, Connie conciliates on that issue as well. She becomes more confident in the Autonomist view than in Conciliationism.

But now things have taken a strange turn. And we can see the strangeness, if we think about what happens when Connie reflects on her own beliefs. She now leans toward the Autonomist theory of rationality. So, she should now see her conciliated low credence in Internalism as probably irrational. She'll think that only a higher, non-conciliated, autonomous credence in Internalism would be rational in her situation. So, now it seems that Conciliationism is indirectly (by requiring Connie to embrace the Autonomist view) requiring Connie to raise her credence in Internalism back up! And if that is right, then Conciliationism does have a problem. After all, Conciliationism directly requires Connie to have a low credence in Internalism. So it would seem to issue in inconsistent requirements in this possible situation. And if a principle issues inconsistent requirements in any situation, it cannot be correct.

If this is a fatal flaw in Conciliationism, that's an important result, because the problem would generalize. It would crop up for any theory of rational belief which allows agents to be rationally misled about what theory of rationality is correct. And it seems that it's hard to avoid allowing that to happen, once we begin theorizing about self-aware agents, who can have reason to doubt their own thinking, and to take those doubts seriously. However, I think there's an attractive line of resistance here.

The line of resistance flows naturally from points we were just looking at. It begins by noticing that Conciliationism actually doesn't require Connie to become more confident again in Internalism—even after she loses confidence in Conciliationism, and gains confidence in the Autonomist view. Of course, once she thinks that the Autonomist theory is probably true, she will presumably think that her low, conciliated, credence in Internalism is probably irrational. But that's just to say that she will be epistemically akratic. And epistemic akrasia, *per se*, turns out not to be irrational. So we haven't shown that Conciliationism issues in inconsistent verdicts after all.

In fact, when we think about the case, it seems plausible that it is rational for Connie to believe akratically in this case. Autonomy is obviously not the same thing as accuracy. Connie's low credence in Internalism is based on her thinking that the majority of professional epistemologists is more likely to reach accurate beliefs about epistemology than she is. But since she now thinks that rational beliefs have to be autonomous, she'll see her own situation as one where accuracy and rationality come apart. And there's nothing mysterious about this, if rationality is tied to autonomy, in the way that the fictional epistemologists in the example believe it is. So, once we see that epistemic akrasia can be rational, we can also see that Conciliationism, and other theories of rationality that take higher-order evidence seriously, can self-undermine without automatically self-destructing.[24]

6. Fifth Upshot: Moderate Skepticism about Certain Controversial Topics

The last upshot of theorizing about self-aware agents that I'd like to discuss has also mainly been discussed in the literature on disagreement. It is this: it will often be irrational to be highly confident of our views on certain controversial topics. In particular, I'm thinking of our views on controversial

[24] As Jennifer Nagel pointed out in the Q&A, Connie will end up akratic not only with respect to her belief in Internalism, but with respect to her belief in Autonomism. And as Nagel noted, there's something ironic about Connie becoming convinced of Autonomism on the basis of a bunch of other people's opinions! I would argue, though, that this should not be seen as problematic. The reasons are parallel to the ones adduced in discussing Dara's belief in Deductive Purism (see fn. 14 above): given her belief in Autonomism, Connie should not take the (as she sees it) irrationality of her belief in Autonomism as an indication that Autonomism is false.

topics where we have good independent reason to think that those who disagree are just as well-positioned to reach accurate views as we are. I take it that philosophy is full of examples of this sort of controversy. On most philosophical topics, there are several mutually incompatible, detailed views. And each one is defended by excellent philosophers. So anyone should acknowledge that most of the controversial detailed views that get defended, even by the most able practitioners of philosophy, are false![25]

We need to face the fact that we're just not very good at philosophy—at least, if being good involves formulating accurate detailed theories on certain issues. And once we understand that about ourselves, it seems to me, it would be irrational for most of us to have high confidence in the controversial conclusions of our own philosophical thinking.[26]

This upshot has been discussed in the literature on disagreement, but I think that at the fundamental level, it's really just about taking a critical perspective on ourselves as thinkers. Disagreement in academic philosophy shows that we, as a group, are not so great at thinking about certain philosophical questions. But one can get evidence that one isn't so great at some particular kind of thinking, in lots of different ways. For example, we have plenty of evidence that people who are hypoxic are bad at complex calculations. And similar kinds of evidence come from psychologists who study bias, or anchoring, or other accuracy-distorting factors that affect us humans. And one can also get evidence from one's own track record—for example, I now know better than to trust my own intuitive thinking about probabilities, or about driving directions. So the theoretical framework that yields this upshot is really a general one. It just has to do with the implications of our ability to take a critical perspective on our own believing.

This upshot for philosophy is obviously somewhat skeptical. I certainly find it sobering—and disappointing, even. But I think it's what emerges from striking a decent balance between skepticism and dogmatism. It's important to notice that the worry is not the radical skeptical one: it's not the worry that we have to we have to somehow globally justify the reliability of our thinking, in ways that don't rely on that thinking. As other people—including Sosa—have argued quite clearly, that's not a reasonable demand of rationality. Instead, the point is just that we have to respect strong positive evidence of our own unreliability. And, unfortunately, in philosophy, that evidence is only too clear.

Now, a philosopher could resist even this moderate degree of skepticism. He could hold seven different highly controversial, detailed views with confidence—or even 17 different views! But I think that the attractiveness of this option fades pretty quickly, when we ask what perspective such a philosopher could take on his own believing. How would such a philosopher explain his having gotten so many controversial issues right, while most philosophers are getting most of them wrong?

I suppose he could think that he had just gotten super-lucky in arriving at the right answers to all these questions. But that doesn't strike me as a rational attitude to take. Or maybe he could think that it wasn't a matter of amazing luck at all. Maybe he could see himself as being consistently quite a bit more intelligent and perspicacious than the vast majority of other philosophers—a very stable genius, perhaps… To my mind, that's not such an attractive view, either.

So I think that, once we engage in a bit of rational reflection on our own believing, a certain degree of skepticism is a natural consequence. It just reflects a reasonable assessment of our powers as epistemic agents. And I think that having a reasonable assessment of our own epistemic

---

[25] I'm assuming that there really is a lot of disagreement out there, and that we're not just talking past each other in ways that just appear to be incompatible. But see Sosa (2011) for discussion of how some apparent disagreement in philosophy may be only apparent.

[26] See Kornblith (2010) for a powerful development of this point.

powers—even if that assessment turns out less rosy than we might have hoped—is something to be grateful for.

Conclusion

I've been trying to illustrate, in a few different ways, the results of developing a strand of thought that's also central to Ernest Sosa's epistemology—though I have to admit, I've been doing it in a framework that's pretty different from his. The strand of thought is that epistemology is much richer when we take care to theorize in a way that doesn't just focus on agents' thinking about the world around them, but also focuses on agents' reflection on their own believing.

This brings with it the possibility of agents getting evidence that bears on the reliability of parts of their thinking. I think that taking this sort of "higher-order" evidence seriously has implications that are pretty profound. We need to structure our principles of rationality to accommodate some sort of Independence principle, which allows higher-order evidence to affect rational belief in a distinctive way. And once we do that, new intricacies emerge.

We encounter so-called "epistemic dilemmas," where even the most rational response to certain evidential situations involves violating some rational requirement. We can see that while paradigmatic cases of epistemic akrasia typically involve irrationality, epistemic akrasia is not intrinsically irrational. And, in some situations, the maximally rational epistemic response is an akratic one. We can see that the correct theory of rationality will, in some cases, not allow us to believe that it is correct! And, more generally, we're led to a somewhat skeptical attitude toward the products of our own philosophical thinking—at least about controversial issues.

This last point may seem more disappointing than exciting. But I think it's a disappointment we should learn to live with. Surely, it's better to embrace a realistic assessment of your epistemic powers than to live in a fool's paradise, where you're forever feeling so clever that you don't see the possibility of your own mistakes.[27]

So to sum up, I would argue that doing epistemology in a way that fully recognizes the implications of self-awareness has two sorts of benefits: First, it reveals surprising and interesting features of the structure of rationality. Second, it helps give us a more accurate picture of what we can accomplish as philosophical enquirers.

Now, to be fair, I should point out that the particular ways I've described the first benefit are currently—well, controversial. What I've been selling as exciting epistemology is currently seen, in certain quarters, as sadly misguided. So, fully appreciating the second benefit, and reflecting on myself as a believer, makes me much less confident that I've described the first benefit correctly.

But in the end, I think that's fine, too. It just means that there's a lot more epistemology left for all of us to do.

[27] Roughly this point was made some time back by Meli'sa Morgan; see her (1986).

References

Barnett, Z. (2020), "Rational Moral Ignorance," *Philosophy and Phenomenological Research*: https://doi.org/10.1111/phpr.12684.

Basu, R. (2019), "Radical Moral Encroachment: The Moral Stakes of Racist Belief," *Philosophical Issues* 29: 9 – 23.

Brinkerhoff, A. (forthcoming), "The Promising Puzzle," *Philosophers' Imprint*.

Buchak, L. (2014), "Belief, Credence and Norms," *Philosophical Studies* 169: 285-311.

Christensen, D. (forthcoming), "Embracing Epistemic Dilemmas," in K. McCain and S. Stapleford, eds., *Epistemic Dilemmas: New Arguments, New Angles* (Routledge).

---. (2020), "Akratic (Epistemic) Modesty," *Philosophical Studies*: (DOI: 10.1007/s11098-020-01536-6).

---. (2019) "Formulating Independence," in Skipper, M. and A. Steglich-Petersen, eds., *Higher-Order Evidence: New Essays* (Oxford University Press).

---. (2018), "On Acting as Judge in One's Own (Epistemic) Case," Marc Sanders Lecture, *Proceedings and Addresses of the American Philosophical Association* 92: 207-235.

---. (2016), "Disagreement, Drugs, etc.: from Accuracy to Akrasia," *Episteme* 13 (4): 397-422.

---. (2010), "Higher-Order Evidence," *Philosophy and Phenomenological Research* 81.1: 185-215.

Elga, A. (ms.), "Lucky to be Rational," presented to the Bellingham Summer Philosophy Conference on June 6, 2008.

---. (2010), "How to Disagree about how to Disagree," in Feldman, R. and T. Warfield, eds., *Disagreement* (Oxford University Press).

Gardiner, G. (2018), "Evidentialism and Moral Encroachment," in K. McCain, ed., *Believing in Accordance with the Evidence: New Essays on Evidentialism* (Springer).

Horowitz, S. (2014), "Epistemic Akrasia," *Noûs* 48:4: 718–744.

---. (2019), "Predictably Misleading Evidence," in Skipper, M. and A. Steglich-Petersen, eds., *Higher-Order Evidence: New Essays* (Oxford University Press).

Hursthouse, R. (1999), *On Virtue Ethics* (Oxford University Press).

Kornblith, H. (2010), "Belief in the Face of Controversy," in Richard Feldman and Ted Warfield (eds.), *Disagreement*, Oxford University Press.

Lasonen-Aarnio, M. (2014), "Higher-Order Evidence and the Limits of Defeat," *Philosophy and Phenomenological Research* 88 (2): 314–345.

Leonard, N. (2020) "Epistemic Dilemmas and Rational Indeterminacy," *Philosophical Studies* 177: 573–596.

Marušić, B. (2015), *Evidence and Agency: Norms of Belief for Promising and Resolving* (Oxford University Press, Oxford).

Morgan, M. (1986): "Fool's Paradise," from *Do Me Baby* (Capitol Records). Song co-written with L. Wilson; recording and lyrics on line at: https://www.lyrics.com/lyric/5939321/Meli%27sa+Morgan/Fool%27s+Paradise, accessed 12/7/2020.

Moss, S. (2018), "Moral Encroachment." *Proceedings of the Aristotelian Society* 118: 177 – 205.

Nelkin, D. (2000), "The Lottery Paradox, Knowledge, and Rationality," *Philosophical Review* 109(3): 373-409.

Nozick, R. (1993), *The Nature of Rationality* (Princeton University Press).

Owens, D. (2002), "Epistemic Akrasia," *The Monist* 85 (3): 381-397.

Schroeder, M. (2018), "When Beliefs Wrong." *Philosophical Topics* 46: 115 – 127.

Smithies, D. (forthcoming), "The Unity of Evidence and Coherence," to appear in N. Hughes, ed., *Epistemic Dilemmas*, (Oxford University Press).

Sosa, E. (1985), "Knowledge and Intellectual Virtue," *The Monist* 68: 224 – 245.

---. (1997), "Reflective Knowledge in the Best Circles," *The Journal of Philosophy* 94 410 – 430.

---. (2011), "Can There Be a Discipline of Philosophy? And Can It Be Founded on Intuitions?" *Mind and Language* 26 (4):453-467.

Stroud, S. (2006), "Epistemic Partiality in Friendship." *Ethics* 116: 498 – 524.

Titelbaum, M. (2015), "Rationality's Fixed Point (or: In Defense of Right Reason)," *Oxford Studies in Epistemology* 5: 253–294.

Weatherson, B. (2019), *Normative Externalism* (Oxford University Press).

Worsnip, A. (2018), "The Conflict of Evidence and Coherence," *Philosophy and Phenomenological Research* 96 (1): 3–44.