Published online first in *Noûs* doi: 10.1111/nous.12441. If you'd like a pdf of the published version, send me an email, and I'll be glad to send you one: david.christensen@brown.edu.

Epistemic Akrasia: No Apology Required[1]

1. Self-Aware Believers and Akrasia

We think about the world; and we ourselves are parts of the world we think about. In some ways, our thinking about ourselves is much like our thinking about other people, or non-human animals, or inanimate things. I have beliefs about my age, the age of my partner, the age of my son's dog, and the age of my house. And when an epistemologist theorizes about what makes these various beliefs rational or irrational, it does not seem that the belief about my own age is peculiar, or that it plays any particularly interesting role.

But sometimes, self-awareness is more epistemologically interesting. One way this can happen is when an agent is able to reflect normatively on her own epistemic situation. Suppose that Anya, a medical resident, has views not only about, for example, whether dosage D of a drug is the appropriate dosage for her patient, but also about whether high confidence that dosage D is appropriate is *rational* in her situation.[2] Once she can do this, it raises questions for the epistemologist about the relationship between two things: (A) what the agent is rational to believe about the world in general; and (B) what the agent is rational to believe about what she's rational to believe. It's natural to think that rationality requires that these two things cohere in a particular way: for example, it's natural to think that it can't be rational for Anya to be highly confident that dosage D is the appropriate dosage, while also being highly confident that it's not rational for her to be highly confident that D is appropriate. An agent who has such beliefs can be thought of as exhibiting epistemic akrasia, and epistemic akrasia is often taken to be irrational—even paradigmatically irrational.[3]

In many cases, akratic beliefs do seem clearly irrational. For example, suppose we fill in the case in one natural way: Suppose that Anya's belief about dosage D is the product of moderately complex figuring. And suppose that she is told that she's been awake for 30 hours, and she knows she has a long history of miscalculating dosages when fatigued. This makes it rational for her to doubt the rationality of her high confidence that D is appropriate for her current patient. But it also seems that Anya should lose confidence in the appropriateness of dosage D (and this seems true, even if the information on her patient's chart actually does support D as the appropriate dosage). Let us suppose

[1] Many thanks to Nomy Arpaly, Zach Barnett, Jessica Brown, Adam Elga, Sophie Horowitz, Ram Neta, Josh Schechter, Jonathan Vogel, Alex Worsnip, Orfeas Zormpalas, two anonymous referees, and the students in my seminar at Brown for helpful discussion and/or comments on earlier drafts. Versions of this paper were given at the University of Rochester, the University of St. Andrews, and the University of North Carolina at Chapel Hill; many thanks to the audiences at all three, and to Julia Staffel, my commentator at the last of these talks.

[2] Here, and throughout, I'll use "rational" to refer to epistemic, rather than practical, rationality.

[3] See, for example, Wedgewood (2002), Adler (2002), Bergmann (2005, 423), Gibbons (2006, 32), Christensen (2010), Smithies (2012), or Littlejohn (2018).

that she does not do this—instead, she continues to be highly confident in the appropriateness of D, while acknowledging that her high confidence in D's appropriateness is likely irrational. In this case, her beliefs will be paradigmatically irrational.[4]

So epistemologists have come up with a variety of "enkratic" principles: principles forbidding akratic beliefs. To fix ideas, and to connect most directly with some principles in the literature, I'll formulate a sample (very rough) enkratic principle which focuses on epistemic rationality, and on credences (or degrees of confidence).[5]

> **Sample Enkratic Principle**: There is no situation in which the maximally rational doxastic response includes both (1) high credence that P, and also (2) high credence that high credence that P is not rational in one's present situation.

This principle is obviously not fully general or precise. It is also intentionally weak: I think that most of those who would impose enkratic requirements would find it congenial, even if they thought it didn't forbid enough.[6]

Enkratic principles derive significant motivation from cases such as Anya's. But there are also cases of apparently rational akrasia; these seem to provide strong arguments for rejecting anything like our Sample Enkratic Principle. Section 2 reviews a couple of these cases. Section 3 examines some general arguments for imposing enkratic principles; it argues that our cases can help us understand why these arguments go wrong. Section 4 examines some ways of formulating moderate enkratic principles—ones that allow certain intuitively akratic beliefs but disallow others. It argues that while these principles avoid certain implausible commitments of strong enkratic principles, the lessons from the previous sections actually show why these moderate enkratic principles fail as well. Section 5 discusses an attempt to allow some akratic beliefs to count as rational, while insisting that they must manifest a different kind of epistemic failure in their subjects. It concludes that no such criticism is generally warranted.

2. Cases of Intuitively Rational Akrasia

---

[4] I take it that this is a natural intuitive judgment; see Horowitz (2014) for detailed arguments that akratic beliefs in this sort of case are irrational.

[5] Other formulations in the literature focus on different epistemic notions (such as justification, evidential support, or epistemic "oughts"), and some focus on categorical beliefs rather than credences. I do not think that the differences will affect the substance of the points made below.

[6] For example, this principle does not forbid what David Alexander, in a recent APA talk, called "omissive akrasia": that is, failing to have an attitude you think is rationally required. Some enkratic principles would forbid omissive akrasia as well. Alexander himself takes omissive akrasia to be permissible, but on grounds that would not also justify violating our Sample Enkratic Principle. I will concentrate on the "commissive" sort of akrasia that would violate our principle.

Several examples of intuitively rational akratic beliefs have been offered in the literature.[7] Here I'll look at two of them in detail.

a. Berto: Evidence of Anti-reliability[8]

One sort of example falls out of a certain class of theories of rational belief for agents who get evidence of their unreliability in a certain area. We can illustrate with a dramatic toy example involving very strong evidence of unreliability.

Suppose Berto, a philosophy graduate student, is taking part in testing a powerful new recreational designer drug. The drug consistently causes people to misevaluate the validity of certain sorts of moderately complex symbolic logic arguments when they think try to think through them directly. In fact, people on the drug are highly anti-reliable at this task: they almost always take the valid arguments to be invalid, and vice-versa. But the drug also leaves them feeling perfectly clear-headed. Berto was an unknowing subject in earlier trials of the drug, and has cringed while watching videos of himself confidently asserting incorrect answers to this sort of problem, over and over. This time, he's given the pills in advance, and asked to do this same sort of problem while his blood pressure is being monitored. Berto looks at Argument 1, which is, as a matter of fact, valid.

What is the maximally rational doxastic response to Berto's situation? Notice first that the fact that the experimenters gave him pills does not mean that he's actually rationally impaired—in fact, we may stipulate that the pills were a placebo this time, though Berto has no evidence suggesting that. So supposing that Berto is actually maximally rational, what will he believe about Argument 1?

Of course, when Berto looks at the argument and thinks through the relationship between the premises and the conclusion, he will reason correctly from the premises to the conclusion: his ability to reason through symbolic logic arguments directly is functioning perfectly. But given what he knows about the drug, it seems that it would be highly irrational for him to trust this reasoning. In fact, given his history of consistently getting this sort of problem wrong while under the influence of the drug, it seems that the maximally rational view for him to end up with is that the argument is very likely *invalid*. So maximally rational Berto will have high credence that Argument 1 is invalid.

I take it that this is a plausible verdict: that Berto is rational to have high credence that Argument 1 is invalid. And a general theory of rationality that yields this result would flow from a natural thought: that the most rational response to evidential situations like Berto's would involve both (1) correct direct assessment of the symbolic argument, and (2) rational accommodation of the anti-reliability evidence the agent has about herself. (The idea is that an agent's direct assessment of an argument's

[7] See, e.g., Coates (2012), Horowitz (2014), Sliwa & Horowitz (2015), Christensen (2016, 2021a), Weatherson (2019), Kappel (2019), Lasonen-Aarnio (2020), Barnett (2021), Hawthorne, et. al, (2021).

[8] The example in this section is adapted from Christensen (2016, 2021a). Examples with a similar upshot are in Sliwa and Horowitz (2015).

validity is a rationally assessable part of their response to their total evidence. This avoids seeing the agent's direct assessment as mere evidence the agent has—which would make what the argument actually says irrelevant to the rationality of the agent's belief about its validity.[9]) So let us suppose that the Correct Theory of Rationality yields this sort of result in general, for both valid and invalid arguments.

Now let us ask what Berto should think about the rationality of his own high credence that Argument 1 is invalid. Of course, this will depend on what general theory of rationality Berto is rational to believe. So let us stipulate that he's been well-educated in epistemology, and rationally believes the Correct Theory of Rationality. On this theory, as we have seen, when an agent has strong evidence of anti-reliability in assessing validity of certain sorts of arguments, and is looking at an invalid argument of that sort, they will be rational to have high confidence that it's *valid*, not *invalid*.

Now, what happens when Berto applies this theory of rationality to his current situation? Berto is rationally confident that he is looking at an invalid argument. So applying the theory of rationality he rationally believes, he'll be confident that only a high credence in *validity* is rational in his circumstances. But in that case, he will be highly confident that his own high credence in *invalidity* is irrational. And this confidence, too, will be rational for Berto.

If this is right, then the maximally rational attitudes to take in certain situations involve a clear violation of our Sample Enkratic Principle: Berto's attitudes are sharply akratic. It might seem surprising that it could be rational for Berto to be highly confident that the argument is invalid, while also being confident that this confidence is irrational. How is his case different from Anya's? Why would it not be more rational for Berto to change his credence in invalidity to a lower one that he could see as rational? I'll leave discussion of detailed arguments for the general irrationality of akrasia for later sections, but for now, it's worth noting a key difference between Berto's case and Anya's.

The difference is one whose importance was first noticed by Sophie Horowitz, in contrasting a different example of intuitively rational akrasia with an example of intuitively irrational akrasia.[10] The theory of rationality that Berto rationally believes entails that rational credence and accurate credence should be expected to come sharply apart in certain cases. For agents with Berto's sort of anti-reliability evidence, the theory entails that they'd be rational to have high confidence in invalidity when looking at valid arguments, and high confidence in validity when looking at invalid arguments. Since Berto knows he has the relevant sort of anti-reliability evidence about himself, he would not

---

[9] For criticisms of this latter sort of view, see Sliwa and Horowitz (2015), Christensen (2016).

[10] See Horowitz (2014). Horowitz's case of intuitive akrasia derives from examples used by Timothy Williamson to argue that a proposition can be known even if it's improbable, on the agent's evidence, that it is known. I won't use Horowitz's exact formulation for dividing rational from irrational akrasia (see Weatherson (2019) and Hawthorne et al. (2021) for criticisms of Horowitz's formulation). But my discussion will, I think, reflect the fundamental insight behind her formulation. Sliwa and Horowitz (2015) develop this line of thought, as does Christensen (2016, 2021a). And Kappel (2019) develops a related position.

(rationally) take the irrationality of his current credence as indicating inaccuracy—in fact, he would take it as indicating accuracy.[11] This is a marked contrast to the case of Anya. When she got evidence that her high credence in the appropriateness of dosage D was irrational, that evidence also cast doubt on the accuracy of her credence.

b. Chitra: Evidence for a False Theory of Rationality[12]

Chitra is an undergraduate philosophy student. She has taken a few epistemology courses taught by the college's faculty, who happen to be strong advocates of a theory of rationality they derive from Hume. Their account incorporates the following principle:

> **Deductive Purism:** Inductive reasoning is not a rational way of supporting beliefs. So, for example, if you wonder whether the sun will rise tomorrow or not, or whether the next bread you eat will nourish or poison you, it's not rational to think either alternative more likely just because of what's happened in the past. Inductively-supported beliefs are indeed *accurate* in general. But rationality is not just about accuracy—it's about support by the right kind of reasons. And only deductive reasoning can render rational support.

Now Deductive Purism is false, of course. But given Chitra's evidence, it seems that it could be rational for her to give it a lot of credence. After all, Chitra's professors argue skillfully for their view; they deftly slice and dice the inductivist readings they assign to the students; and they have confident responses ready whenever Chitra and her classmates raise objections to Deductive Purism. And most importantly, Chitra knows that her professors are professionally certified experts on epistemic rationality! So I'll suppose that Chitra, responding to her evidence as rationally as anyone could, gives high credence to a theory of rationality that incorporates Deductive Purism.[13]

---

[11] This is not to say, of course, that Berto needs to consciously think explicitly about the accuracy of his belief about Argument 1. It's just to say that, given the theory of rationality he believes, the irrationality of this belief would not indicate inaccuracy, so his confidence that his belief is irrational does not defeat his belief by suggesting that it is inaccurate.

[12] This example is taken from Christensen (2021a), which adapted it from Barnett (2021). Lasonen-Aarnio (2020, 613) gives a schematic example where an agent's evidence favors belief in a false epistemological theory, in order to argue that there are cases where evidence will support an akratic pair of beliefs. Weatherson (2019, 170ff) offers a concrete example in this vein, though it differs in a respect I'll discuss below. Hawthorne, et al. (2021) also offer examples of this form. Kappel (2019) uses a similar example to argue for the rationality of what he calls "rule-akrasia." And Feldman (2005, 109) floats the possibility that mistakes about the general nature of evidential support could provide counterexamples to his enkratic requirement to "respect the evidence."

[13] I take this to be the intuitively plausible verdict. One might worry that students are not typically rational to accept the views promoted by their philosophy teachers, especially if they realize that the views are controversial. But we can imagine that Chitra's professors have given her the impression that opposition to Deductive Purism is outdated, and that she has no idea how controversial it really is. Thanks to Zach Barnett for raising this point.

There are also those who, on theoretical grounds, would reject the idea that it could be rational for Chitra to give any credence to Deductive Purism under any circumstances; we'll look at their arguments below.

Now suppose it's lunchtime, and Chitra is thinking about the sandwich she just bought at the food truck—will the bread nourish her, or poison her? She raises the question to her favorite professor, only to receive a dismissive response: "Don't be silly!" he smiles, chewing on his own sandwich. "We never said that inductively-supported beliefs weren't generally accurate, right? Of course they are! It's just that they're not *rational*, okay? I mean, unless—heh-heh—you've found yourself some non-circular way of justifying induction!"

In this situation, it seems clear that Chitra is rational to be highly confident that her sandwich will nourish her, just as all her previous sandwiches have. But it also seems that it's rational for her to be confident that this very high credence is irrational. In other words, the maximally rational response to Chitra's evidential situation is a sharply akratic one.

Again, it's worth briefly noting a key difference between Chitra's attitudes and Anya's. Chitra, like Berto, is rational to see her situation as one in which rational belief and accurate belief come apart. The evidence that makes it rational for Chitra to doubt the rationality of her confidence that her bread will nourish her does not thereby make it rational for her to doubt that her bread will nourish her. So evidence for certain false theories of rationality can have the same result as evidence for one's own unreliability.

3. Considerations against Rational Akrasia

People sometimes dismiss the possibility of rational akrasia by considering cases like Anya's, and emphasizing (sometimes with amusing dialogues) how silly their akratic agents sound.[14] But this alone should not be convincing: It's easy to find examples of irrational akratic beliefs; but the fact that, say, Anya's akratic beliefs are clearly irrational cuts little ice when we consider cases like Berto's or Chitra's.

There have, however, been some general considerations offered that would support denying the rationality of akratic beliefs. So let us turn to consider two prominent ideas of this sort. And in particular, let us examine how they would apply to our examples of intuitively rational akrasia.

a. "No rational mistakes about rationality"

Any rational akratic pair of beliefs has to involve the agent being rationally misled about rational requirements. For suppose that an agent is highly confident of both "P" and "Believing P is irrational in my situation." If their first belief is rational, their second belief is inaccurate—so they've been misled about a rational requirement. And if that second belief is also rational, they've been rationally misled. So there's a natural connection between the idea that akrasia can be rational and the idea that one can be rationally misled about rational requirements. Michael Titelbaum (2015, 2019) argues from the irrationality of akrasia to the irrationality of mistakes about rational requirements. And Clayton Littlejohn (2018) argues for the irrationality of mistakes about rationality in order to explain how an

[14] See, for example, Elga (2013), Horowitz (2014), Littlejohn (2018), and Silva (2018).

enkratic principle could be maintained in conjunction with other claims he takes to be plausible. So the idea that akrasia is irrational, and the idea that mistaken beliefs about rational requirements are irrational, make a mutually-supportive pair of views.

But why think that being rationally misled about rational requirements is impossible? [15] One reason one might have flows from thinking about the rational requirements in these cases as determinable *a priori*. In Berto's case, it's plausibly *a priori* that Argument 1 is valid; but Berto's confidence in the irrationality of his belief about Argument 1 is based on thinking that Argument 1 is invalid. In Chitra's case, it might be held that Deductive Purism is *a priori* false; but Chitra's confidence that her sandwich belief is irrational is based on thinking that Deductive Purism is true. So if one held that the *a priori* support for such beliefs was indefeasible, one might conclude that it would be most rational for Berto to believe—perhaps to be absolutely certain!—that Argument 1 was valid (despite all of his experience with the drug). And one might conclude that Chitra would be most rational to reject Deductive Purism—perhaps to be absolutely certain that it was wrong! A position along these lines is put forward by Titelbaum; it is intended to apply both to logical truths, and to *a priori* truths about rational requirements: "every agent possesses a priori, propositional justification for true beliefs about the requirements of rationality in her current situation. An agent can reflect on her situation and come to recognize facts about what that situation rationally requires. Not only does this reflection provide her with justification to believe those facts; that justification is ultimately empirically indefeasible."[16]

Now since Titelbaum argues from the irrationality of akrasia to the irrationality of mistakes about rational requirements, it might seem odd to discuss his position here: after all, he is not arguing from the irrationality of mistakes about rationality to the irrationality of akrasia.[17] But Titelbaum is here offering an *explanation* for why akrasia is irrational. Insofar as there's reason to doubt this explanation, there is reason to think that maybe akrasia can be rational after all. Moreover, as we've seen, rational akrasia requires the possibility of making rational mistakes about rational requirements. So if there was independent reason to think that rational mistakes about rational requirements were impossible, that would be reason to reject the possibility of rational akrasia as well. A defender of akrasia, then, should take seriously reasons that are offered for thinking that rational mistakes about rationality are impossible.

I think that there is surely something attractive in the thought that the *a priori* justification we have for certain necessary truths—such as truths about logic, or about rational requirements—would be empirically indefeasible. These truths do not depend on any empirical contingencies, and the *a priori*

---

[15] Lasonen-Aarnio (2020) presses this question nicely.

[16] Titelbaum (2015). Titelbaum (2019) develops this position, and Ram Neta (2018) defends a related position. Smithies (2019, ch. 10, also (forthcoming)) argues agents always have propositional justification for absolute certainty in the correct view about what beliefs are rationally required in their situations, and that ideal rationality requires being certain of those facts. He does recognize a separate set of "non-ideal" rational standards that allow for uncertainty about rational requirements.

[17] Thanks to an anonymous referee for prompting me to address this point.

justification we have for believing them is similarly independent of empirical matters. So how could empirical information possibly defeat that justification? If one conceives of an ideally rational agent as simply reasoning about logic, rationality, and the world surrounding her, I think that it is hard to see how any empirical information she might come across could undermine or oppose the justification she had for beliefs about *a priori* rational requirements.

But I think that it would be a mistake to conceive of ideally rational agents as reasoning only in this "outer-directed" sort of way. Agents can think about themselves as well. And this sort of reflection can complicate our theory of rationality. It is important to keep in mind that even an ideally rational agent can get misleading evidence about contingent matters. One contingent matter she may get evidence about is the reliability of her own thinking processes. This is a separate matter from the truth or falsity of some rational requirement, or the normative question of whether a certain belief is rational: it is about whether the agent's own cognitive machinery is malfunctioning or not. And evidence may bear on the general reliability of certain of one's cognitive processes, irrespective of whether those processes happen to involve *a posteriori* reasoning about contingent matters, or *a priori* reasoning about logic or rational requirements. After all, the psychological processes that constitute *a priori* reasoning are no more immune to glitches than the ones constituting *a posteriori* reasoning are. Of course, if an agent is actually an ideal reasoner, they will in fact always reason perfectly. So there will be some causal or nomological sense in which all their reasoning is immune to cognitive errors. But this is a fact about the world (that portion of the world that includes the agent's cognitive equipment). And, like other causal or nomological facts, it is not something one can know *a priori*. So even an ideally rational agent cannot rationally be certain that her reasoning faculties are not malfunctioning.

Now if an agent can get evidence rationalizing doubts about the reliability of her own thinking on some matter, it seems that rationality will require her to take these doubts seriously in the following way: it will not be rational for her to completely trust or rely on the results of her thinking about the matter in question. This phenomenon may well be different from standard sorts of defeasibility. It is clearly not a case of rebutting defeat. And it even seems to differ from the usual examples of undermining defeat—as when the justification that one's visual experience provides for one's belief that a table is red is undermined by the information that the lighting in the room is deceptive. But it does seem to be a kind of defeat, at least in the following sense: it affects the level of confidence an agent may rationally have in certain beliefs—the beliefs that are products of the targeted thinking. And this sort of defeat can occur whether or not the belief in question was a logical theorem, or a truth about rational requirements.[18]

---

[18] For a detailed development of essentially this line of argument, see Christensen (2007).

One might concede that an ideal agent would be ideally rational to doubt their own general reliability in thinking about *a priori* matters, yet to be certain, about any particular *a priori* matter P, that they got that matter right. Smithies (2019, Ch. 10.4) suggests this sort of response, suggesting that the agent should remain absolutely certain of P while doubting that their belief that P was properly based, since proper basing would require reliable thinking. I do not see how it could be ideally rational to be confident in P while thinking that one's reasoning to one's belief in P was likely unreliable. See Skipper (2021) for a similar point.

So there seem to be reasons to doubt that the *a priori* justification we have for beliefs about rational requirements really is indefeasible. Without some independent reason to think that it is indefeasible, we have here no independently plausible explanation for the irrationality of akrasia. So while having such an explanation would help defend enkratic principles against intuitive counterexamples, the claim that we have indefeasible justification for true beliefs about rational requirements does not seem to provide that.

A different argument for the irrationality of false beliefs about rational requirements is offered by Littlejohn (2018, 270):

> If your first-order attitudes violate rational requirements … you'll manifest the kind of incompetence at handling reasons that merits the charge of irrationality. If instead you judge that you should form beliefs that happen to violate these requirements, this judgement reflects the same incompetence. … This is why mistaken beliefs about what rationality requires of you are themselves irrational beliefs.

This thought, too, has something plausible about it—at least if we focus on certain sorts of agents. So if an agent irrationally embraces some conspiracy theory, and also thinks themselves rational to do so, both beliefs will likely spring from the same rational flaw. But things look different when we think about examples of intuitively rational akrasia.

Consider Chitra's mistaken belief about what rationality requires of her: She believes that rationality would require her to have low confidence that her sandwich will nourish her. Does this belief manifest the same incompetence that would be manifested were Chitra to actually have low confidence that her sandwich would nourish her? It seems highly unlikely that it would.[19] If Chitra were to have low confidence that her sandwich will nourish her, it would be a result of not believing in accordance with overwhelming inductive evidence. That, of course, would likely manifest an incompetence. But nothing like that is involved in Chitra's false belief about rationality. Her belief about rationality is based on her taking seriously the testimony of acknowledged rationality-experts, not on flouting induction. In fact, it's not clear to me that Chitra's belief about rationality manifests any incompetence at all.

Similarly, consider Berto's false belief that rationality requires low credence that Argument 1 is invalid. That belief is based on Berto's *correctly* appreciating the validity of the argument, and then concluding, because of the strong (but misleading) evidence of his anti-reliability, that the argument is invalid. Berto's basic appreciation of logical relations is spot-on. And his incorrect belief that rationality requires low credence that Argument 1 is invalid is based on his being rationally misled, by strong,

[19] Lasonen-Aarnio (2020, fn. 44) also registers skepticism about Littlejohn's claim.

clear evidence into believing that his own ability to directly assess the validity of this sort of argument has been compromised. Again, I see no incompetence here.

So the strategy of denying the possibility of rational false beliefs about rational requirements does not seem to me a promising one for defending enkratic requirements. But there is one caveat worth bringing out. The examples we've looked at seem to provide very strong reason to believe that, in certain situations, the most rational doxastic response possible is an akratic one. However, this is consistent with thinking that even these responses involve some element of rational imperfection. After all, Berto's belief about Argument 1 does violate logic—and logic plausibly furnishes a rational constraint on beliefs. And if theories of rationality are *a priori*, Chitra's belief about Deductive Purism might be argued to run afoul of a similar constraint. If we took this line, then Berto and Chitra would end up violating some rational requirement no matter what they ended up believing. But these would still be cases where an akratic doxastic state was more rational than any non-akratic one.[20] (In the discussion that follows, I'll usually avoid cumbersome formulations by using "rational" to refer to the beliefs that are the most rational possible in the agent's evidential circumstances.)

There is also one final point worth emphasizing about the position that akratic agents exhibit rational imperfections in virtue of having false beliefs about rational requirements. Akrasia is a relational phenomenon: it involves the relation between a belief about what's rational, and an ordinary belief. But the rational imperfection we're now talking about would lie only in the agent's false belief about rationality.[21] We can see this if we imagine Chitra, for example, giving up her belief that the sandwich would nourish her, while retaining her false belief about rationality. In such a case, Chitra would no longer be akratic. But the rational imperfection—if such there be—would be there still. Conversely, if Chitra began by just having the false belief about rationality, and then formed the belief that her sandwich would nourish her, she would at that later point become akratic—but she would exhibit no new rational imperfection.

The upshot of this is that even if we granted that akrasia could only occur in agents who exhibited a certain rational imperfection, that would not show that akrasia, *per se*, was problematic. (Compare: fitting apologies can only be given by agents who have done something wrong. But that doesn't show that there is something wrong with fitting apologies.) Nothing in the sort of argument we've been

---

[20] Titelbaum (2015) notes, but does not endorse, the view that certain situations involve unavoidable violations of rational requirements. (For a defense of this view of rational requirements, see Christensen (2021b)). Neta (2018, 322), which argues against the rationality of akrasia on grounds that mistakes about rationality are themselves irrational, acknowledges that akratic beliefs may be the most rational ones certain agents can form, given their "limited time and energy". I'd be inclined to say something stronger—that the most rational doxastic state possible for agents in situations like Berto's and Chitra's is akratic, even if the agents have unlimited time and energy. It's the evidence they have, not any cognitive limitations, that rationally requires akrasia.

Finally, this point does not seem to me to support Smithies' view that akrasia is only rational in a "non-ideal" sense, while ideal rationality always precludes akrasia. For agents in the situations we've been considering, even if we see some rational imperfection in their doubting truths about rational requirements, there is, I think, no sense in which any non-akratic set of beliefs would be more rational.

[21] See Hawthorne, et al. (2021, 223) for a similar point.

looking at would implicate the relation between beliefs which constitutes akrasia. In the next section, then, we'll turn to look at a different kind of anti-akratic argument—one that explicitly targets the relation that holds between an akratic pair of beliefs.

b. "Akrasia Can't Make Sense from the Agent's Own Perspective"

A number of writers have located the problem with akratic beliefs in some failure to make sense, from the perspective of the akratic agent. For example, Titelbaum (2019, 227) writes, "[R]ationality involves an agent's attitudes' making sense from her own point of view". He then goes on to argue that making sense means a lack of internal tension, and that akratic beliefs stand in tension with one another. Littlejohn (2018, 267) says this about an agent who is knowingly akratic: "The mindset of this person is opaque. It's hard to see how rationality could sanction such a mindset." And Smithies (2019, § 9.5) argues that justified beliefs must be capable of withstanding justified reflection, and that "there is no basis on which akratic attitudes can withstand justified reflection" (308). He goes on to cite Burge (1996) in arguing that "the *unity* of one's reflective perspective on the world depends on the existence of immediate rational connections between one's first-order beliefs and one's higher-order reflection on those beliefs" (310, my emphasis).

Again, there's something initially plausible about this sort of line—at least, if we concentrate on paradigmatic cases of akrasia such as Anya's. When Anya believes that dosage D is the appropriate dosage for her patient, and also believes that, due to her tendency to make dosage-figuring mistakes when fatigued, her belief in the appropriateness of D is probably not rational, it does seem that Anya's reflection on herself sits ill with her continued confidence that D is appropriate. If we asked Anya how she could be so confident in the appropriateness of D while acknowledging that this belief is likely irrational, it's hard to see how she could offer an intuitively reasonable response. After all, her reason for thinking her dosage belief likely irrational is that she is currently likely to have made the kind of rational error that tends to produce inaccurate beliefs.

However, we should not be so quick to conclude that all akratic beliefs are hard to make sense of from the agent's own perspective.[22] Suppose, for example, we were to put an analogous question to Berto: "How can you be confident that Argument 1 is invalid, while acknowledging that this belief of yours is likely irrational?" It seems that Berto has a perfectly reasonable response to this sort of worry.

Let's first consider Berto's high confidence in the proposition that Argument 1 is invalid. It is true that Argument 1 looks valid to Berto when he considers it directly. But Berto thinks he's been given a drug which reliably causes people to directly misevaluate the validity of symbolic logic arguments like Argument 1. So, in light of the evidence he has about the drug, Berto's direct assessment sits well with

[22] See Lasonen-Aarnio (2020, 618) for doubts about whether the notion of states being "difficult to make sense of from a first-person perspective" is applicable in a clear way to certain cases of akrasia.

believing that the argument is actually invalid. So far, there seems to be nothing mysterious or perplexing about Berto's belief, even from his own point of view.

Now let's consider Berto's attitude toward the proposition that high confidence in the invalidity of Argument 1 is irrational in his situation. His attitude toward this proposition of course depends on the account of rationality he believes. As stipulated above, Berto believes a theory of rationality with the following consequence: Any agent who (a) has Berto's particular sort of anti-reliability evidence, and (b) is looking at an invalid argument, is rational to be highly confident that the argument is valid, not invalid. Given that Berto believes this account of rationality, and believes that he is just such an agent, he naturally concludes that the rational attitude toward Argument 1's validity, in his own current situation, is to have high confidence that it's valid, and that high confidence in invalidity is irrational. This, too, does not seem mysterious or perplexing: Berto is simply applying his account of rationality to his own situation, as he sees it.

So it is not hard to see why Berto believes each of the two propositions. These considerations also, of course, show why he believes that these two beliefs of his are accurate. But perhaps the perplexity comes out in having these two beliefs simultaneously: isn't there some tension or instability in Berto having a belief he thinks is irrational? Shouldn't Berto be bothered by this, feeling some pressure to change one of these beliefs?

In general, if one believes that there is some tight correlation, in one's own situation, between beliefs that are rational and beliefs that are accurate, then there is indeed a sharp tension between believing that P, and simultaneously believing that one's belief that P iss irrational. And I take it that, in the vast majority of cases, for the vast majority of beliefs, it *is* rational to think that rationality and accuracy are tightly correlated. This explains why it would be natural to think that there was some general tension inherent in having akratic pairs of beliefs. But insofar as it can be rational to believe that rationality and accuracy come apart—for certain particular beliefs in certain particular situations—the source of the tension that usually holds between the members of akratic pairs of beliefs will be absent for those beliefs in those situations.

So if we grant that it can be rational for Berto to believe the account of rationality he believes, then it seems that his having the two beliefs together actually makes perfect sense, from Berto's own perspective. There is nothing at all opaque about Berto's mindset, and nothing disunified about his perspective on himself and the world. Berto has reflected rationally on his belief that Argument 1 is invalid, and the belief has withstood this reflection. It has withstood his reflection not because he considers it rational, but because careful reflection reveals that irrationality in his situation would not indicate inaccuracy.

A similar lesson emerges from thinking about Chitra's case. Suppose we ask her how she can believe her sandwich will nourish her, despite thinking that this belief is irrational. She will invoke Deductive Purism in explaining how inductively-based beliefs tend to be accurate, even if they're not rational.

Again, there is nothing disunified, opaque, or nonsensical about her perspective on herself, and the careful reflection that leads her to think that her sandwich belief is irrational does not also support doubts about the nutritive powers of her sandwich.[23]

Both examples depend on the agent expecting that rationality and accuracy will come apart in their situations. One might claim that this expectation itself somehow could not make sense. But I'm not sure why this would be. For one thing, though I think that Deductive Purism is pretty clearly false, theories of the sort that Berto believes have serious defenders.[24] Insofar as these theories have some plausibility, it's not implausible that the correct theory of rationality actually entails that rationality and accuracy diverge in cases involving certain kinds of evidence.

Moreover, it's not even necessary, in order for the agent's akratic beliefs to make sense from her own perspective, that the theory of rationality she believes be true. What matters is that she believe it—or, at most, that she believe it rationally. And barring some reason to think that it can never be rational to give credence to any but the One True Theory of Rationality, it's easy to see how the most rational response to an agent's evidence can involve giving substantial credence to false theories of rationality.

Thus it seems to me that examples such as Berto's and Chitra's do more than just providing facially plausible counterexamples to our Sample Enkratic Principle. When we look at exactly how Berto or Chitra's general beliefs about rationality help make sense, from their own perspectives, of their own akratic beliefs, we also can see why other cases—such as Anya's—really do fit with the idea that akratic beliefs can't make sense from the agent's own perspective. The examples thus help clarify the root of the "can't make sense" intuition. And, in doing so, they help us see both why the idea that akratic beliefs can't make sense from the agent's own perspective is initially plausible, and also why it does not apply in full generality.

4. Might only moderate akrasia be rational?

a. Two moderate enkratic principles

As we've seen, akrasia is tightly related to making mistakes about rational requirements; and denying the possibility of making rational mistakes about rational requirements strains plausibility. But one might hope to allow some rational uncertainty about rational requirements without totally giving up on the idea that there's a rationally mandated connection between beliefs about things in general, and

[23] This is why it makes sense that Chitra would feel no urge to change her sandwich belief. In her rational equanimity about her belief, Chitra differs from Weatherson's akratic Aki (2019, 171 ff). Aki, like Chitra, holds a false theory of rationality, and takes one of her rational beliefs to be irrational. But unlike Chitra, she holds this belief because she "can't bring herself" to believe in the way she thinks rationality requires.

[24] See, e.g., Sliwa and Horowitz (2015) or Christensen (2016).

beliefs about what it's rational to believe. And indeed, some have offered principles, in what might be thought of a tempered enkratic spirit, which aim to do just this.

One example is Elga's (2013) New Rational Reflection principle. If you satisfy NRR, then your credence in p, on the supposition that a certain credence function Pr is rationally ideal, is equal to the credence that that function Pr assigns to p, on the supposition that it is indeed the ideally rational credence function:

NRR: $Cr(p \mid Pr \text{ is ideal}) = Pr(p \mid Pr \text{ is ideal})$.[25]

Elga offers NRR as an improvement on Rational Reflection (Christensen 2010), which would impose a very tight enkratic condition relating credences about rationality to credences in general.[26] NRR, like Rational Reflection, is intended to capture the idea that one should treat the credences that are rational in one's situation as an "expert": as opinions that are likely to be accurate. Elga shows that NRR, unlike Rational Reflection, allows for the possibility of agents being rationally uncertain about what credences it's rational to have in their own situations.[27]

Moreover, Elga shows that NRR is consistent with the rationality of certain instances of intuitively rational akrasia—the kind that figure in the Williamson-style cases discussed in Horowitz (2014). But it is clearly violated in paradigmatically irrational instances of akrasia. So NRR would embody a general, rationally-required connection between beliefs about things in general, and beliefs about what one is rational to believe. Elga argues persuasively that NRR does a better job than Rational Reflection of capturing the intuition that one should treat rational credences as an "expert" about the subject matters of one's beliefs.

A related principle is offered by Kevin Dorst (2020). Like Elga, Dorst is motivated by thinking that it must be rational to be uncertain about what rationality requires in certain situations, but also that one's beliefs about things in general must be constrained in some general way by what one is rational to believe about what one is rational to believe. And, like Elga, he thinks that certain intuitively akratic doxastic states can be rational (see Dorst 2019). He offers his Simple Trust principle as an

---

[25] By 'ideal', Elga means ideally rational in the agent's situation; I'll follow his usage.

[26] If we use Pr_{ideal} to mean the credence function that's ideally rational in the agent's situation, Rational Reflection reads as follows: **Rational Reflection**: $Cr(p \mid Pr_{ideal}(p) = x) = x$.

[27] Rational Reflection goes wrong, Elga argues, by treating as expert the function that's ideally rational *on the agent's evidence*. To the extent that the agent is supposing a certain function to be ideal for the purposes of treating it as an expert, the agent should ask what credences the function would have, given not only the agent's evidence, but also given *that it is indeed the ideal function*. This is what's represented in the right side of NRR. The move from Rational Reflection to NRR parallels the move from the original form of the Principal Principle (which was designed to take objective chances as expert), to the New Principal Principle (see Lewis (1986, 1994) and Hall (1994). Elga shows how conditionalizing the expert function in this way allows for rational uncertainty about what rationality requires.

improvement on NRR. If you satisfy Simple Trust, then your credence in p, on the condition that the ideally rational credence function in your situation assigns p a credence of at least x, must be at least x:

ST: $Cr\,(p \mid Pr_{ideal}(p) \geq x) \geq x$.[28]

ST looks like Rational Reflection, but in a weakened form, and Dorst shows that it, unlike Rational Reflection, allows for rational doubts about rational requirements. Nevertheless, satisfying ST—like satisfying the other principles we've looked at—also does impose a general, rationally-required connection between beliefs about things in general, and beliefs about what one is rational to believe. In particular, Dorst shows how it precludes certain akratic beliefs, though it would not preclude others. Like Elga, Dorst sees his principle as expressing the idea one should treat rational credences like an "expert" about the subject-matters of one's credence; in fact, his informal slogan for Simple Trust is "take expert judgments on trust" (2020, 593).

With these principles on the table, let's look at how they would apply to our cases of intuitively rational akrasia. As it turns out, NRR seems to allow for the rationality of Berto's beliefs. But it clearly rules Chitra's akratic beliefs irrational.[29] Chitra is highly confident that her sandwich will nourish her, but also highly confident that the credences that would be rational in her situation are those that Deductive Purism would sanction. But that means that her credence in her sandwich nourishing her, conditional on the rational credences being those Deductive Purism would sanction, must also be high. To connect this with NRR, let us use n to stand for the proposition that Chitra's sandwich will nourish her, and DP to denote the general account of rationality Chitra has been taught, which incorporates Deductive Purism. So, for Chitra,

$Cr(n \mid$ DP-sanctioned credences are ideal$)$

must be high.

---

[28] ST differs a bit from Dorst's formulation of Simple Trust, for continuity with the other formulations discussed here. Dorst formulates his principle just in terms of the ideally rational credence function:

Simple Trust: $Pr_{ideal}\,(p \mid Pr_{ideal}\,(p) \geq x) \geq x$.

But the import is the same: Simple Trust says that it's rationally required that an agent's credences satisfy the formulation I've used.

In motivating the principle, Dorst notes that it encodes thought that learning that an expert is *at least t confident* in p—that is, learning a lower bound on the expert credence—cannot provide evidence *against* p. Dorst also shows that requiring Simple Trust is equivalent to requiring the same principle, but with the inequalities reversed, so both occurrences of "≥" are replaced by "≤". Thus on his view, learning that the expert is *at most* t confident in p cannot be evidence *for* p. Interestingly, Dorst shows that this does not make Simple Trust equivalent to Rational Reflection. And this allows one to satisfy Simple Trust while being uncertain about what rationality requires of one.

I should also note that Dorst generalizes his principle to one he calls Trust, which entails Simple Trust. But these complications will not affect the discussion.

[29] Christensen (2016) makes the first point, and Christensen (2021a) argues for the second point along the lines set out here.

However, the credence that DP would assign to her sandwich nourishing her, on the supposition that the DP-sanctioned credences were rational, would not be high: After all, according to DP, inductive evidence is irrelevant to rational belief. So the credence it would assign to n would presumably be middling at best.[30] So, in Chitra's situation,

DP(n | DP-sanctioned credences are ideal)

will be middling at best. The two quantities can't match, so the most rational doxastic reaction to Chitra's situation involves violating NRR.

A parallel point holds for ST. This is easiest to see if we use an equivalent formulation of ST with the inequalities reversed; I'll also instantiate the principle to our example:

ST for Chitra's situation: $Cr\,(n \mid Pr_{ideal}(n) \leq x) \leq x$.

It's clear that Chitra's credences violate ST. Suppose we take .7 as x. Chitra is rationally much more than .7 confident that her sandwich will nourish her. But she is also highly confident that DP is the correct account of rational credence, and would assign to n a much lower credence as rational. So she will also be more than .7 confident in n, even on the supposition that the rational credence in n in her situation is .7 or lower. So her credences will violate ST.[31]

It looks, then, as though our principles—ones that are designed to allow for moderate akrasia while maintaining a fairly strong relationship between beliefs in general and beliefs about what beliefs are rational—do not furnish acceptable constraints on rational belief, at least if Chitra's reactions to her situation are indeed rational.

One could of course push back here, saying that Chitra's beliefs are not rational. One way to do this would be to hold that she would be rational to lose confidence in the claim that her sandwich will nourish her (and, presumably, in all other inductively-supported claims). But that seems incredibly implausible. Presumably, the better line would be to hold that Chitra would be rational to be sure that Deductive Purism was false. But this, too, seems implausible, in the present context.

It's important to keep in mind that what the example shows is not that NRR and ST are inconsistent with the truth of Deductive Purism—that would be no problem at all. What the example shows is that the principles are inconsistent with the claim that Chitra could rationally lend a good amount of credence to DP. But if we allow—with Elga and Dorst—that one can be rationally uncertain about what rationality requires, then it is hard to see how denying this claim could be motivated.

---

[30] This could be made precise by formulating DP explicitly, but since it's obvious that some ways of doing this would have this result, there's no reason to fiddle with the details.

[31] The argument in the text is quick and informal; for a worked example with numbers, see the Appendix. Similar reasoning will show that ST will also rule Berto's credences irrational. It would require that:

$Cr(Valid \mid Pr_{ideal}(Valid) \geq .4) \geq .4$. But since Berto is highly confident both that Argument 1 is *not* valid, and also that the rational credence for validity is $\geq .4$, his credences will violate ST.

Chitra's lending credence to DP poses a problem for NRR and ST because DP allows divergences between factors that make beliefs rational and accuracy-conducive factors. But while this may strike some of us as suspect, it's not as if a defender of DP would simply be changing the subject, as would someone who claimed that "rational" beliefs were those that could be expressed in Portuguese palindromes. We can disagree over the principles of rationality, just as we can disagree over the principles of morality. In fact, there are many theories of rational belief actually defended by reputable epistemologists which take rationality of belief to depend, in a transparent way, on non-accuracy-conducive factors.[32] Insofar as Chitra may be rationally uncertain of what rationality requires, surely the fact that a theory of rationality is defended by epistemological experts can rationalize her lending that theory credence. And that is what it takes to open up a gap between the beliefs she rationally expects to be accurate, and the ones she rationally expects to be rational.

So it seems that once we look at how one might try to resist the apparent counterexamples to NRR and ST, we can see why the examples are not just oddities whose intuitive appeal can be easily dismissed. Thinking about what is responsible for generating this sort of example shows why we should expect the principles to fail. The reason is this: As we saw, there are two root motivations behind these principles. One is that we must allow for people being rationally misled about what beliefs are rational in their situations. The other is that people are rationally required to take rational beliefs as "experts"—that is, as likely to be accurate about their subject matters. But unless we had some argument showing how, although one could rationally lend credence to various false claims about rationality, one could never lend credence to theories such as the ones we've been discussing, we should admit the possibility of rationally believing a theory on which rationality should not be treated as an "expert". So the two fundamental motivations for our moderate enkratic principles turn out to be in serious tension with one another.

b. Can it be rational to expect rationality to be an anti-expert?

One way of resisting the train of thought above would be to insist that it can never be rational to expect rationality and accuracy to diverge dramatically, to the point where rationality is—at least in some local domain—an anti-expert. This would be a strong prohibition, as it constrains what a person could be rational to believe about rationality, no matter what their evidence was. What could underlie such a prohibition?

Dorst (2020) develops this general strategy by providing several arguments intended to highlight troubling consequences of any account that says, for example, that Berto is rational to expect high rational credence in validity to correlate inversely with validity. He first objects that any such account makes it impossible to answer the question "Why conform to my evidence?" But if "conforming to my evidence" just means "adopting the beliefs my evidence makes rational," any theory that allows rational akrasia will reject the claim that an agent should *try* to conform to their evidence. Of course, this is not to say that such theories are committed to denying that agents are required to conform to

[32] In addition to the papers mentioned in fn. 8, see, for example: Nozick (1993), Nelkin (2000), Stroud (2006), Buchak (2014), Marušić (2015), Moss (2018), Schroeder (2018), or Basu (2019).

their evidence! When we allow akrasia, we see that conforming to one's evidence can sometimes require adopting beliefs that one *thinks* do not conform to one's evidence.[33] Insisting that the beliefs that agents are rational to form must be the ones they're rational to *believe rational* is simply to reject akrasia from the start.

A similar point applies when we consider how belief feeds into practical reason. Dorst points out that an agent with beliefs like Berto's will expect that betting rationally would lose money.[34] Berto believes that he'd be rational to believe in Argument 1 being valid, so if he takes rational bets to be those based on rational beliefs, he'll believe it would be rational for him to bet on Argument 1's validity. But Berto also believes that Argument 1 is invalid, so he'd expect that betting on Argument 1's validity would result in a loss. Dorst takes the upshot of this phenomenon to be that theories that require agents to expect accuracy to be anti-correlated with accuracy "allow rational requirements to lead to a sure loss," and he points out that many, such as those who offer Dutch Book arguments, "take it as a premise that rational requirements can do no such thing" (2020, 602).

But again, it's important to notice that on the account of rational belief defended above, it's not true that the bets that are *actually* rational for Berto to take are ones that Berto should expect to lose money on. On this account, Berto rationally believes Argument 1 is invalid. So he would be rational to bet on invalidity, and if he did, he'd rationally expect to win! Dutch Book arguments assume that betting on *the beliefs that are actually rational for agents to have* should not lead to expectable losses. But that's not what's going on here at all. Berto would be rational to expect loss if he were to bet on the belief *he incorrectly thinks would be rational* for him to have. But he would not be rational to bet that way. So on the account defended above, it's simply not true that conforming to rational requirements will lead Berto to expect monetary loss.[35]

One more line of intuitive argument is worth addressing here. As Dorst points out, I.J. Good famously argued that it's always rational to accept free evidence, because it helps us make good choices. But if one expects beliefs that the evidence makes rational to anti-correlate with truth, one

---

[33] Compare Arpaly (2002) on moral akrasia: sometimes, doing what's morally best can require doing something one thinks is not morally best. More generally, a theory of rationality need not be held hostage to any requirement that it result in a cognitive-self-help manual that agents can apply self-consciously to guide their believing. See Arpaly (2000) for a parallel point concerning practical rationality.

[34] See Dorst (2020, 602). Dorst's particular argument is formulated in terms of a formal condition which involves expecting losses with *certainty*, which seems to assume that agents are certain of the relevant claims about rationality. But in adapting his argument to our example, which involves high confidence rather than absolute certainty, I think I've been true to the intuition behind the argument. Since Dorst's arguments concentrate on anti-correlation cases, I'll apply them to Berto's case rather than Chitra's. But I think that the root intuition behind Dorst's arguments would apply to Chitra's case as well, as would the responses offered below to these arguments.

[35] A similar point applies to Dorst's argument (2020, 602-3) that the theories he's targeting would have rational agents expect one doxastic option to be rational, but expect an alternative doxastic option to be more accurate than that one. This would be a problem, of course, if the theories required the agents to *have* the credence they expected to be rational. But they don't. And the credences they require the agents to have are not ones that the agents will rationally expect to be less accurate than some particular alternative credence.

might worry that free evidence would not be such a great bargain after all. Surely it would be silly to go about gathering evidence, if one thought that the beliefs made rational by that evidence would skew inaccurate—at least, if one's plan involved then acting on the beliefs one took the new evidence to rationalize.

But does permitting akrasia really make evidence expectedly worthless? Let us think again about Berto's situation. Suppose he's confident that he's been given the anti-reliability drug, but let's vary the example so that he hasn't been shown Argument 1 yet. And suppose he'll have the opportunity to receive a prize if he answers a question about Argument 1's validity correctly. Would he be more rational to just answer without looking, or to look at the argument before answering? Given his long experience with the drug, it's clear that he should plan to look at the argument—and then answer "valid" if it looks invalid to him, and vice-versa. Of course, when he answers, he'll think of himself as believing irrationally but accurately, but he'll still see the evidence as valuable because he expects it to help him make a successful decision. Guessing without looking would only be rational if Berto was bound to answer in the way he thought was rational, rather than the way he thought was correct. But given that he expects rationality and accuracy to diverge, he would not be rational to do that. So the basic insight—that evidence is valuable because it helps make good decisions—is fully consistent with recognizing cases where it's rational to expect accuracy and rationality to anti-correlate.

Now one might rightly point out that the arguments of this section purport to show only that certain particular arguments against the possible rationality of ever expecting rationality and accuracy to anti-correlate do not succeed. And this clearly does not amount to a direct positive argument showing that the relevant sorts of expectations are rational in certain situations. So it is worth making the dialectic clear.[36] The upshot of the previous section was that there's at least a strong *prima facie* reason to think this the relevant sort of expectation can, in certain cases, be rational: For such expectations to be rational, all that's required is that the agent rationally give considerable credence to certain sorts of theories of rationality. And the principles under discussion presuppose that at least some uncertainty about rationality can be rational. So, absent good reason to think that there's some special *barrier* to people getting evidence for, and then rationally lending credence to, any of the particular accounts of rationality that would support the relevant expectations, it would seem that we have good reason to think that the expectations can indeed be rational. And if that's right, then to the extent that initially attractive arguments for such a special barrier do not succeed, we have more reason to think that it can in some cases be rational to expect rationality and accuracy to diverge in the way that permits rational akrasia.

Moreover, thinking through the arguments discussed in this section while paying careful attention to the distinction between believing rationally, and having beliefs one believes to be rational, does something else. It helps show how the expectations in question may seem at first to be problematic, and yet turn out, on closer inspection, not to lead to problems after all. In doing this, it helps explain

[36] Thanks to an anonymous referee for prompting me to do this.

away a certain strand of intuitive resistance to the idea that it can be rational to expect rationality and accuracy to diverge in the ways we've been looking at.

Of course, it's still true that, in the vast majority of cases, we should not expect rationality and accuracy to come dramatically apart. And in many cases, it will indeed be irrational to be akratic. But we should not be tempted to overgeneralize. And once we see exactly how it makes sense to be akratic—even sharply akratic—in certain sorts of situations, we should also resist the temptation to think that there must be some general, abstract, formal relationship—even a moderate one—between what's rational to believe in general, and what's rational to believe about what's rational to believe.

5. But aren't akratic subjects still doing *something* wrong?

Suppose one is convinced that akratic beliefs can sometimes be rational. One might still find this puzzling, given the evident appeal of enkratic principles, at least in the abstract. Maria Lasonen-Aarnio defends the rationality of some akratic beliefs, but considers enkratic requirements so plausible that arguments for violating them create a paradox (2020, 601 ff). Her solution is to argue that there is, in fact, something wrong with akratic beliefs—even if they are the most rational ones possible in their situations. It's just that the agent's failure is not one of having irrational beliefs, but rather one of manifesting epistemically bad *dispositions*.

Lasonen-Aarnio proposes that we may evaluate doxastic states epistemically from two perspectives. One involves epistemic successes—beliefs that are rational, that fit the agent's evidence or reasons for belief. This perspective may, as Lasonen-Aarnio argues, evaluate akratic beliefs positively. But the second perspective involves evaluating doxastic states "by evaluating the dispositions manifested in forming and maintaining them."[37] Good dispositions are ones whose manifestations tend to be epistemically successful. And Lasonen-Aarnio argues that akratic beliefs manifest bad dispositions.

The particular dispositions in question involve how agents respond to the contents of certain higher-order beliefs—for example the belief that "I'm not rational to believe P". Lasonen-Aarnio argues that when such a belief is true, the agent has a conclusive and conspicuous reason not to believe P. Of course, in a rationally akratic agent, the relevant belief is false—that's why she may be rational in believing P as well. But an agent who believes P despite having the higher-order belief that it's irrational for her to believe P will be disposed to persist in believing other things akratically, when she *correctly* believes that they're irrational to believe. Such a disposition will tend overall to manifest in epistemic failures, not successes. So from the perspective that focuses on the dispositions agents manifest in forming and maintaining beliefs, akratic beliefs come up short:

> [E]ven if an akratic subject manages to be in doxastic states that track her evidence, we are baffled, as in failing to respond to the contents of her beliefs—contents that, if true, constitute conclusive and conspicuous reasons to be in, or not be in, certain doxastic states—she

[37] (2020, 620), see also her (2021).

> manifests bad dispositions, dispositions that involve ignoring her reasons or evidence across a wide range of cases. We have here a case where the only way of conforming to an evidential norm, or of correctly responding to one's reasons, is by doing something that is vicious from the perspective of conforming to such a norm across a range of other cases. One-off success is achieved by manifesting a bad disposition. That's why akrasia looks bad across the board. (2020, 629)

But why should we think that our agent's believing P while believing "I'm not rational to believe P" would manifest a general disposition to ignore evidence of her own irrationality in a way that will result in maintaining irrational beliefs? Lasonen-Aarnio argues that it's implausible to think that a human being could discriminate between cases where "it's irrational for me to believe ___" is false, and cases where it's true. So the disposition that manifests in our agent akratically believing P will also manifest in akratic beliefs in Q, R, S, and so on, where Q, R, S and so on *really are* irrational to believe.

There is something appealing about this thought, at least in the abstract. And even if we think more concretely about the dispositions of an akratic agent like Anya, it's not implausible that she would have some disposition toward epistemic recklessness, a disposition that would, for example, manifest in her maintaining confidence in erroneous dosage-calculations done while sleep-deprived. But as we saw above, we should be careful about generalizing from cases of clearly irrational akrasia to global conclusions about akrasia. So let's examine our concrete examples of rational akrasia.

Chitra, as we saw, is disposed to reason inductively, even though she believes, on expert say-so, that inductively-supported beliefs aren't rational. This led to her akratically believing "My sandwich will nourish me" while also believing "It's not rational for me to believe that my sandwich will nourish me." Is there any reason to think that Chitra has some disposition that would tend to manifest in irrational beliefs? She's presumably disposed to form other akratic beliefs in other cases involving inductive evidence—but those will be like her sandwich-belief: not irrational at all. What about different cases where she gets reason to believe that a belief of hers is irrational? Suppose, for example, she goes on to medical school, calculates the dosage for a patient while sleep-deprived, and gets the kind of evidence that Anya got about such dosage-calculations. I see no reason to think that she'd react recklessly, like Anya did. After all, her maintaining her sandwich belief was in response to rationally expecting accuracy and rationality to come apart in that case. She would have no reason to suppose that they came apart in cases of sleep-deprivation—so since she'd naturally believe that irrationality in such cases indicated inaccuracy, she'd revise her confidence in her dosage calculation.

What about Berto? Let us grant, as is psychologically plausible, that he is disposed to react to future logic-problem anti-reliability drug evidence in the same way as he did in our case: he will end up having beliefs about the logic problem that he considers irrational. Now it's worth noting that evidence that one has been drugged in this way should not be expected to be generally misleading (it just happened to be misleading in Berto's case, because he was slipped a placebo). Typically, agents who get this sort of drug evidence will actually be mis-assessing the arguments when they consider

them directly. So Berto's disposition to respect the evidence of his own impairment, by correcting for the likely mis-assessments, is not a bad disposition: it will generally lead to more accurate beliefs about logic problems than would the disposition to ignore evidence that he'd been drugged.

Of course, this is not the disposition Lasonen-Aarnio was worried about. She was worried about the disposition to have beliefs in the face of believing them irrational. And in the general run of cases where Berto gets anti-reliability drug evidence, he will end up believing that his corrected beliefs about the logic problems are irrational: that is just a consequence of his applying the theory of rationality he believes. So is there something bad about Berto's disposition to retain beliefs he deems irrational in these cases? Of course, this will depend on the details of Berto's disposition: is it a disposition simply to ignore evidence of irrationality in general, or is it more finely tuned? It seems to me that the latter interpretation is more natural (and we could certainly explicitly stipulate this). In judging his first-order beliefs irrational, Berto is carefully and self-consciously applying a theory of rationality on which—in cases where agents have strong anti-reliability evidence of the sort Berto has—irrationality does not indicate inaccuracy. If he's disposed to believe akratically only in cases where his account of rationality separates rationality and accuracy in this way, it is not a disposition that manifests pig-headedness or recklessness. In fact, giving up or changing the first-order beliefs in these sorts of cases should generally be expected to result in less accurate first-order beliefs.

Of course, cases resembling the drug case are highly atypical of cases where agents get reason to think that a belief of theirs is irrational. And if Berto was disposed to brush off evidence of his own irrationality in general, this would be a problem, for exactly the reason that Lasonen-Aarnio suggests. But there is no reason at all to think that Berto's disposition to believe akratically in certain special cases would lead in general to Berto's ignoring evidence of his own irrationality. There of course may well be people who are generally cavalier about the possibility of believing irrationally. But nothing about Berto's case would suggest that he is generally cavalier, since we've only stipulated that he maintains beliefs he considers irrational when his theory of rationality divorces rationality from accuracy. Suppose we consider a more standard kind of evidence of irrationality, where this is not the case. Suppose, for example, that Berto has just evaluated some job candidate files, and has formed the belief that Darren is a bit better than Dora. We present him with strong evidence that he is likely to irrationally underrate women's files compared with men's. Is there any reason, based on his akratic response to the logic-drug situation, to suppose that he'll ignore this evidence, saying, "Yeah, my evaluations are probably irrational, but I'm sticking to 'em!"? I don't see any. In fact, Berto's reaction to the logic-drug evidence suggests, if anything, quite the opposite: that he's disposed to take evidence of his own unreliability seriously, and adjust his credences accordingly.

So while it's undoubtedly true that akratic beliefs often manifest bad dispositions, I see no reason to think that this is true of akratic beliefs in general. In fact, thinking about our concrete examples of intuitively rational akratic beliefs supports just the opposite conclusion. For all we've seen, Berto and Chitra look to be paragons of epistemic virtue. And in seeing why our akratic agents need not have bad dispositions, we can also see why it's unnecessary for those who admit rational akrasia to then go

about searching for a way of explaining what's wrong with akratic beliefs, or of finding fault with the agents who have them. Because the answer is, sometimes, there's nothing to find! Not only is akrasia rationally required in certain situations, but the agent's being akratic need not manifest any flaw in her at all. An agent who was ideal in their dispositions to respond to evidence would be akratic in these situations.[38]

6. Conclusion

Epistemic akrasia can seem clearly irrational, and, in many cases, akratic beliefs are indeed irrational. But reflection on certain sorts of cases suggests that akrasia can sometimes be rational after all. This can happen when agents have evidence which supports expecting a divergence between rationality and accuracy in their own situations.

General arguments have been offered against the possibility of rational akrasia. But examining how those arguments would apply to our cases of apparently rational akrasia helps show why we should not find these arguments persuasive. We should allow that, in some cases, agents rationally give credence to views of rationality on which it can be rational to expect rationality and accuracy to diverge in their own case. Once we see that, we can see how akrasia may be the most rational response to certain evidential situations; and we can also see how akratic beliefs may make perfectly good sense, even from the akratic agent's own perspective.

Some have sought to formulate more modest principles of rationality: they would allow some rational uncertainty about rational requirements, and some degree of intuitive akrasia, while still imposing a general formal relationship between an agent's beliefs in general, and her beliefs about which beliefs were rational in her situation. But further reflection on our cases suggests that even this more moderate strain of anti-akrasia is misguided. For the more moderate principles are motivated by taking rationality as a certain kind of 'expert'. And if agents can be rationally misled about rational requirements, they will in some cases come to rationally expect accuracy and rationality to diverge in their situations—in other words, to expect that rationality, in their situation, is not an 'expert' after all.

At this point, then, it looks doubtful that there's any interesting general relation at all that's rationally required to hold between an agent's beliefs in general, and her beliefs about what beliefs are rational in her situation. Moreover, this should not prompt us to go about trying to find some *other* defect in akratic beliefs, or in the agents who have them. The fact that many cases of akrasia do involve irrationality makes it understandable that we're inclined to suspect that akrasia is always somehow

[38] Here I may also be disagreeing with Weatherson (2019, 175), who says that akratic agents, while they may be rational, will say or do things that would not be said or done by ideal agents. Of course, this depends on what one means by 'ideal agent'—akratic agents are not, for example, omniscient, since by definition they must be misled about what beliefs are rational for them. But they may yet be epistemically ideal in believing, and being disposed to believe, in the best possible way, given their evidence. I take it that an ideal agent who has misleading ordinary evidence will be misled, and thus fail to be omniscient. And, as noted above, even if a false belief about requirements of rationality in particular is held to violate a rational requirement, it may yet be the most rational belief possible given the agent's evidence. And there is in any case no further problem created by adding to it another (rational) belief that makes the agent akratic.

problematic. But this suspicion does not survive careful scrutiny. In some cases, an akratic agent may not only have rational beliefs, but she may also have those beliefs as a result of her having impeccable epistemic dispositions. There is simply no call to criticize such an agent. In her akrasia, there is nothing she needs to apologize for.

Appendix

Suppose, conservatively, that Chitra is rationally .95 confident that her sandwich will nourish her.

1. $Cr\ (n) = .95$.

And suppose she's rationally .8 confident that DP is correct and assigns ≤ .7 credence to n. She'll then be at least .8 confident that the rational credence in n in her situation is ≤ .7. To make things as hospitable as possible to ST, I'll assume that her credence in this is .8, and not greater.

2. $Cr\ (Pr_{ideal}\ (n) \leq .7) = .8$

Chitra's rational credences will presumably obey something like the law of total probability:

3. $Cr(n) = [\ Cr\ (n \mid Pr_{ideal}\ (n) \leq .7) \cdot\ Cr\ (Pr_{ideal}\ (n) \leq .7)\ +\ Cr(n \mid Pr_{ideal}\ (n) > .7) \cdot Cr(Pr_{ideal}\ (n) > .7\ ]$

$= [\quad x \quad \cdot \quad 8 \quad + \quad y \quad \cdot \quad .2 \quad]$

Can Chitra's credences obey ST? ST, applied to the present case, requires that:

4. $Cr\ (n \mid Pr_{ideal}\ (n) \leq 7) \leq 7$,

meaning that x in line 3 could be at most .7. So to make the whole sum as high as possible consistent with ST, let x be .7, and let y be 1. We would then get the following from 3:

5. $Cr(n) = [.56 + .2] = .76$

This would make Chitra's credence in n still way less than .95. So Chitra's credences cannot obey ST.

References

Adler, J. E. (2002). *Belief's Own Ethics*. Cambridge, MA: MIT Press.

Arpaly, N. (2000). On Acting Rationally against One's Best Judgment. *Ethics,* 110, 488–513. doi: 10.1086/233321

---. (2002). Moral Worth. *Journal of Philosophy,* 99, 223 – 245. doi: 10.2307/3655647

Barnett, Z. (2021). Rational Moral Ignorance. *Philosophy and Phenomenological Research,* 102, 645-664. doi: 10.1111/phpr.12684

Basu, R. (2019). Radical Moral Encroachment: The Moral Stakes of Racist Belief. *Philosophical Issues,* 29, 9 – 23. doi: 10.1111/phis.12137

Bergmann, M. (2005). Defeaters and Higher-Level Requirements. *The Philosophical Quarterly,* 55, 419-436. doi: 10.1111/j.0031-8094.2005.00408.x

Buchak, L. (2014). Belief, Credence and Norms. *Philosophical Studies,* 169, 285-311. doi: 10.1007/s11098-013-0182-y

Burge, T. (1996). Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society,* 96, 91-116. doi: 10.1093/aristotelian/96.1.91

Christensen, D. (2007). Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals. *Oxford Studies in Epistemology,* 2, 3-31.

---. (2010). "Rational Reflection. *Philosophical Perspectives*: *Epistemology*, 121-140. doi: 10.1111/j.1520-8583.2010.00187.x

---. (2016). Disagreement, Drugs, etc.: from Accuracy to Akrasia. *Episteme,* 13, 397-422. doi: 10.1017/epi.2016.20

---. (2021a). Akratic (Epistemic) Modesty. *Philosophical Studies,* 178, 2191–2214. doi: 10.1007/s11098-020-01536-6

---. (2021b). Embracing Epistemic Dilemmas. In K. McCain, S. Stapleford, and M. Steup (Eds.), *Epistemic Dilemmas: New Arguments, New Angles* (pp 165-181). New York, Routledge. doi: doi:10.4324/9781003134565

Coates, A. (2012). Rational Epistemic Akrasia," *American Philosophical Quarterly,* 49, 113-124.

Dorst, K. (2019). Higher-Order Uncertainty. In Skipper, M. and A. Steglich-Peterson (Eds.), *Higher-Order Evidence: New Essays*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198829775.003.0002

---. (2020). Evidence: A Guide for the Uncertain. *Philosophy and Phenomenological Research,* 100, 586–632. doi: 10.1111/phpr.12561

Elga, A. (2013). The puzzle of the unmarked clock and the new rational reflection principle. *Philosophical Studies*, 164, 127-139. doi: 10.1007/s11098-013-0091-0

Fantl, J. and M. McGrath (2002). Evidence, Pragmatics and Justification. *Philosophical Review,* 111, 67-94. doi: 10.2307/3182570

Feldman, R. (2005). Respecting the Evidence. *Philosophical Perspectives,* 19, 95-119. doi: 10.1111/j.1520-8583.2005.00055.x

Gibbons, J. (2006). Access Externalism. *Mind*, 115, 19–39. doi: 10.1093/mind/fzl019

Hall, N. (1994). Correcting the Guide to Objective Chance. *Mind*, 103, 505–518. doi: 10.1093/mind/103.412.505

Hawthorne, J., Y. Isaacs, and M. Lasonen-Aarnio (2021). The rationality of epistemic akrasia. *Philosophical Perspectives,* 35, 206–228. doi: 10.1111/phpe.12144

Horowitz, S. (2014). Epistemic Akrasia. *Noûs,* 48, 718-744. doi: 10.1111/nous.12026

Kappel, K. (2019). Escaping the Akratic Trilemma. In M. Skipper. and A. Steglich-Peterson (Eds.), *Higher-Order Evidence: New Essays* (pp. 124–143). Oxford: Oxford University Press. doi:10.1093/oso/9780198829775.001.0001

Lasonen-Aarnio, M. (2020). Enkrasia or evidentialism? Learning to love mismatch. *Philosophical Studies,* 177, 597–632. doi: 10.1007/s11098-018-1196-2

---. (2021). Coherence as Competence. *Episteme,* 18, 453-476. doi: 10.1017/epi.2021.33

Lewis, D. (1986). A Subjectivist's Guide to Objective Chance. In his *Philosophical Papers*, vol. 2 (pp. 83-132). New York, Oxford University Press. doi: 10.1093/0195036468.003.0004

---. (1994). Humean Supervenience Debugged. *Mind,* 103, 473–490. doi: 10.1093/mind/103.412.473

Littlejohn, C. (2018). Stop Making Sense? On a Puzzle about Rationality. *Philosophy and Phenomenological Research,* 96, 257-275. doi: 10.1111/phpr.12271

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

Moss, S. (2018). Moral Encroachment. *Proceedings of the Aristotelian Society,* 118, 177 – 205. doi: 10.1093/arisoc/aoy007

Nelkin, D. (2000). The Lottery Paradox, Knowledge, and Rationality. *Philosophical Review*, 109, 373-409. doi: 0.2307/2693695

Neta. R. (2018). Evidence, Coherence and Epistemic Akrasia. *Episteme,* 15, 313 -328. doi: 10.1017/epi.2018.25

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

Schroeder, M. (2018). When Beliefs Wrong. *Philosophical Topics,* 46, 115 – 127. doi: 10.5840/philtopics20184617

Skipper, M. (2021). Higher-Order Evidence and the Normativity of Logic. In K. McCain, S. Stapleford, and M. Steup (Eds.), *Epistemic Dilemmas: New Arguments, New Angles* (pp. 21-37). New York: Routledge. doi: doi:10.4324/9781003134565

Silva, P. (2018). Explaining enkratic asymmetries: knowledge-first style. *Philosophical Studies,* 175, 2907-2930. doi: 10.1007/sll098-017-0987-l

Sliwa, P and S. Horowitz (2015). Respecting *all* the evidence. *Philosophical Studies,* 172, 2835-2858. doi: 10.1007/s11098-015-0446-9

Smithies, D. (2012). Moore's Paradox and the Accessibility of Justification. *Philosophy and Phenomenological Research,* 85, 273-300. doi: 10.1111/j.1933-1592.2011.00506.x

---. (2019). *The Epistemic Role of Consciousness*. Oxford: Oxford University Press. doi:10.1093/oso/9780199917662.001.0001

---. (forthcoming). The Unity of Evidence and Coherence. In N. Hughes (Ed.), *Epistemic Dilemmas*. Oxford: Oxford University Press.

Stroud, S. (2006). Epistemic Partiality in Friendship. *Ethics,* 116, 498 – 524. doi: 10.1086/500337

Titelbaum, M. (2015). Rationality's Fixed Point (or: In Defense of Right Reason). *Oxford Studies in Epistemology,* 5, 253–294. doi: 10.1093/acprof:oso/9780198722762.003.0009

---. (2019). Return to Reason. In M. Skipper and A. Steglich-Peterson (Eds.), *Higher-Order Evidence: New Essays* (pp. 226–245). Oxford: Oxford University Press. doi:10.1093/oso/9780198829775.003.0011

Weatherson, B. (2019). *Normative Externalism*. Oxford: Oxford University Press. doi:10.1093/oso/9780199696536.001.0001

Wedgewood, R. (2002). The Aim of Belief. *Philosophical Perspectives*, 16, 267-297. doi: 10.1111/1468-0068.36.s16.10