The Rationality of Epistemic Akrasia

Lasonen-Aarnio, Maria

Wiley Blackwell
2021

Lasonen-Aarnio, M, Hawthorne, J & Isaacs, Y 2021, 'The Rationality of Epistemic Akrasia', Philosophical Perspectives, vol. 2021, no. 35.1, https://doi.org/10.1111/phpe.12144, pp. 206-228. https://doi.org/10.1111/phpe.12144

http://hdl.handle.net/10138/585864
10.1111/phpe.12144

unspecified
acceptedVersion

The Rationality of Epistemic Akrasia

by John Hawthorne, Yoaav Isaacs, and Maria Lasonen-Aarnio

Introduction

To be epistemically akratic is either

(1) to believe *p* and also believe that believing *p* is rationally forbidden

or

(2) to not believe *p* and also believe that not believing p is rationally forbidden[1]

The epistemic akratic either possesses a belief which she believes it is rationally forbidden to possess, or lacks a belief which she believes it is rationally forbidden to lack.

If rational akrasia is possible then rational false belief must be possible, since rational akrasia must involve a false belief. If the rationality of a belief entails its truth then the impossibility of rational akrasia follows automatically. But a general factivity constraint on rational belief is not in the spirit of the anti-akrasia literature, as it maintains that akrasia cannot be rational for a different reason. The driving idea is that the epistemic akratic somehow fails by her own lights, and that there is some distinctive sort of incoherence involved in failing by one's own lights. Indeed, requirements positing the irrationality of akrasia are often seen as one instance of requirements prohibiting a specific kind of *structural irrationality*.[2]

First, an initial bit of ground clearing. Insofar as a proposition can be believed under multiple guises, then epistemic akrasia is clearly unproblematic in a variety of cases where multiple guises are in play.[3] Consider:

---

[1] Standard definitions of epistemic akrasia all have the same basic structure. There are two elements:

(1) a belief / lack of belief in some proposition

(2) a belief that such a belief / lack of belief in that proposition is bad in some specific way

The sort of badness we've employed is being rationally forbidden. Other sorts have also been employed, and most of what we have to say can be adapted to these other notions.

[2] See, for instance, Worsnip (2018).

[3] Cf. Lasonen-Aarnio (2020: 609-610).

Through a Glass Darkly: You are touring a castle. You look out through a clear window and you see a ship. You are struck by the ship's beauty, and so you believe that it is beautiful. Moving elsewhere, you look out through an occluded window. You see an indistinct shape. You can tell that it's a ship, but you can't tell much else about it. You thus form the belief of the ship that you ought not to believe that it is beautiful. But unbeknownst to you (and surprisingly) the two windows look out onto the same ship. So there is a ship *x* such that you both believe that *x* is beautiful and believe that you ought not believe that *x* is beautiful.

There is clearly nothing shameful about the akratic combination in this scenario. Anti-akrasia epistemology must either deploy a sufficiently fine-grained conception of propositions so that cases of this sort cannot arise, or else refine the anti-akrasia principle in a way that makes explicit some constant guise assumption. We shall charitably assume that guise worries of the sort raised by the above example can be adequately controlled for.

Even assuming that the guise issue can be controlled for, the thesis that all epistemically akratic states are irrational is subject to counterexamples. We shall, in part one below, present various families of counterexamples.[4] Some proponents of anti-akrasia principles concede the existence of isolated counterexamples that they hope to circumscribe, thereby preserving the irrationality of epistemic akrasia in all but a certain special class of cases.[5] Against this background, the point we want to make is that counterexamples are pervasive, and have various distinct sources. In part two, we look at two positive lines of argument for anti-akrasia principles. In part three, we look at a strategy for keeping the anti-akratic sensibility alive in the light of the examples presented in section one, a strategy that appeals to idealization. All told, the case against anti-akratic principles is surprisingly strong and the case for anti-akratic principles is surprisingly weak.

Part I: The counterexamples

I.1 *Lack of Access to Beliefs*

One might have a belief and yet fail to realize that one has it. One might lack a belief and yet fail to realize that one lacks it. These access failures may have various sources. They may simply be due to a failure of introspective acuity, analogous to a failure of perceptual acuity. Or they may be more deeply rooted in a false theory about the nature of belief and its relation to other states like knowledge. Either way, imperfect access to one's belief states can yield cases where epistemic akrasia seems rationally unproblematic.[6]

First, a note on the relevance of anti-luminosity considerations for the broader debate on structural incoherence. It is symptomatic, we think, that attempts to state what the incoherence at issue is, as well as to draw out its distinct kind of badness, often end up making luminosity assumptions. It might be tempting to think that certain combinations of states – those are incoherent

[4] These counterexamples can easily be adapted to apply to practical akrasia, though we will leave the case of practical akrasia aside.
[5] See, for instance, Horowitz (2014) and Titelbaum (2015).
[6] Cf. Lasonen-Aarnio (2020: 609) and Titelbaum's (2015) remarks on what he calls failures of "state luminosity".

in the relevant way – pop out in a way that is difficult to miss. According to this thought, weeding out incoherence should be a simple exercise in mental house cleaning, simpler than being rational in some more demanding sense that requires doing things like taking into account one's evidence in appropriate ways. Niko Kolodny (2007: 242), for example, argues that incoherent subjects are always in a position to know that they are incoherent in a specific way: for instance, if a subject both believes a proposition *p* and believes not-*p*, this fact is 'available' to her, and she is therefore in a position to know that she either believes *p* or believes not-*p* without sufficient reason. This is what, Kolodny thinks, distinguishes coherent but mistaken agents from incoherent ones: coherent agents are not always in a position to know they are mistaken. But if belief and intention are susceptible to anti-luminosity arguments, then such a view is doomed to fail.
Now to the counterexamples. Consider the following case:

Good News: The epistemic oracle reveals some information about the normative status of your belief, or lack thereof, in a proposition *p*. If you believe that *p* then belief in *p* is rationally required, and if you don't believe that *p* then belief in *p* is rationally forbidden. You breathe a sigh of relief—either way it looks like you're in the clear. You believe that you believe that *p*, and thus you believe that belief in *p* is rationally required. But you're not perfectly reliable about what your beliefs are, and in this case you've made a mistake. In fact, you don't believe that *p*. So you do not believe that *p* and also believe that not believing *p* is forbidden. Thus you are epistemically akratic.

Analysis: This doesn't at all feel like a case in which you're failing by your own lights. You believe, after all, that whether or not you believe *p,* you are rational. Given your imperfect introspection about your doxastic state you seem to have behaved quite reasonably. So although you are epistemically akratic, it doesn't seem that you are irrational.

Going Ancient: You're convinced by various texts in ancient philosophy (for example, Plato's Republic) that knowledge is inconsistent with belief. This is a mistake; the contemporary consensus is correct, and in fact knowledge entails belief. You know that your spouse is a good person, and you know that it's important that you know this. Thus you conclude that you ought not to believe that your spouse is a good person, as you believe that such belief would be inconsistent with knowledge. But since knowledge entails belief, you do in fact believe that your spouse is a good person. Thus you both believe that your spouse is a good person and believe that you ought not to believe that your spouse is a good person, and are epistemically akratic.

Analysis: The innocent mistake about the relationship between belief and knowledge satisfyingly explains your mistake about whether or not you ought to believe that your spouse is a good person. You were convinced that your spouse is a good person and believed that you should be convinced that your spouse is a good person—you only made a mistake about what being convinced amounted to vis-a-vis belief. There's nothing remotely incoherent about your epistemic state, and no call to consider you irrational.

Missed It By That Much: You're 79% confident that the Yankees will make it to the World Series next year. You know that you're 79% confident that the Yankees will make it to the World Series next year, and you know that you ought to be exactly that confident. You've thought about the relationship between belief and credence a lot. You believe that a credence of .75 or greater suffices for belief. But while there is a Lockean threshold for belief, its value is actually .8 rather than .75. Thus you believe that you ought to believe that the Yankees will make it to the World Series next year (you know that you ought to have credence .79, and you falsely believe this is above the threshold for belief), but you don't in fact believe that they will.

Analysis: The innocent mistake about where the Lockean threshold[7] is produces an access failure: you fail to believe *p,* but owing to your mistaken views about the threshold for belief, you do not realize that you fail to believe *p* (Similarly, a case where the real threshold is lower than one reasonably thinks it is could produce a failure to believe that you believe *p* when you do.) This mistake satisfyingly explains your erroneous belief about what you ought to believe. You knew how confident you ought to be and knew you were exactly that confident. The fact that, surprisingly, your level of confidence that the Yankees will make it to the World Series next year does not amount to a belief does not make your epistemic state incoherent and does not make you irrational.[8]

I. 2 *False Theories of Rationality*

We have seen that false (but rationally formed) theories of belief can yield intuitively unproblematic cases of epistemic akrasia. The same can happen in cases of false but rationally formed theories about the nature of rationality.[9]

A Weakness: Following Williamson, you believe that you shouldn't believe things you don't know. You know that you don't know that your lottery ticket will lose, though you know you believe that it will. Hence, you believe that you ought not to believe that your lottery ticket will lose, but you do believe that your lottery ticket will lose. In fact Williamson is wrong: you should believe that the lottery ticket will lose.

Analysis: The innocent mistake about the relationship between belief and knowledge satisfyingly explains your mistake about whether or not you ought to believe that your lottery ticket will lose.

[7] See Foley (2009).

[8] Clayton Littlejohn (2018) appeals to a thought common among those who hold akratic states to be irrational. Concerning a rational belief that a certain belief is rationally required, he says "With this belief in place and its blessing from rationality, its hard to see how rationality could then require you to refrain from believing [the proposition in question]". But here, so long as the belief about thresholds is rational, it is obvious enough how one could rationally believe a certain belief to be rationally required and yet it be forbidden. Indeed, to enter into a belief state in this scenario would run afoul of one's knowledge that a credence of .8 or higher was forbidden.

[9] Cf. Lasonen-Aarnio (2020: 613). False theories of rationality play an important role in Littlejohn (2018): his conclusion is that because rationally believed false theories of rationality make for akratic combinations, such theories cannot be rationally believed. We think the examples below suggest a different conclusion.

We find it implausible that you are being irrational here. No doubt some will demur, thinking that one is constrained by what one rationally judges to be rationally impermissible, even if that judgment is based on a false theory of rationality.[10] As further therapy, consider the following:

Unger Games: Following Unger, you believe that a belief in *p* is rational only if you have reasons for believing *p* and that something can be a reason for belief only if it is known. Moreover you reasonably trust an epistemologist—Peter Unger, in fact—who tells you that knowledge is unachievable. You thus believe that none of your beliefs are rational, thinking that the best that can be hoped for is some lesser status. But Unger has got all this wrong, and in fact many of your beliefs are rational.

Analysis: Here again epistemological errors need not involve a lapse in rationality. Once one has convinced oneself—rationally—that rationality is too much to be hoped for, it does not seem that the akratic state that results is irrational.

The cases we articulated provide substantial reason for doubting that all epistemically akratic agents are irrational. But we do not wish to declare the rationality of epistemic akrasia prematurely. Having presented a battery of recipes for constructing cases in which akrasia doesn't look that bad, let us turn to positive arguments for anti-akratic principles.

Part Two: **Defenses of Anti-Akratic Principles**

It is not atypical for authors to take anti-akratic principles for granted, and then set out to derive further conclusions from them.[11] Positive arguments often take the form of describing particular akratic agents and remarking that there is something clearly amiss. But the truth of a general anti-akratic principle is but one possible explanation, and those who reject such principles can offer alternative explanations.[12] Let us therefore explore some more systematic ways in which one might argue for anti-akrasia epistemology. If systematic arguments for the irrationality of akrasia called the status of our counterexamples into doubt then our case for the rationality of akrasia would be insecure.

[10] It may help to flesh out the case in more detail. Assume that you take yourself to believe, but not know, the Williamsonian theory (you think it is very hard to know philosophical theories). Then you might think to yourself: "Maybe the Williamsonian theory is wrong and my belief that it is right is permissible but false - in which case my belief that I ought not to believe my ticket will lose is false. Or maybe the Williamsonian theory is right and I shouldn't believe either that my lottery ticket will lose or that the Williamsonian theory is right. But given that I don't know the theory to be right it seems a bit much to throw out both beliefs: After all, if I throw them out then my failure to believe that I will lose the lottery will start to look pretty bad: there will be lots of evidence that I will lose and, in the absence of belief in the Williamsonian theory, I will likely then think that my failure to believe that I will lose is impermissible, and so I will be akratic anyway."

[11] For a recent example, one that we shall later discuss in more detail, see Titelbaum (2015).

[12] See e.g. Lasonen-Aarnio (2020, Forthcoming B).

II.1 *Moore-paradoxicality*

Perhaps the most popular argument for an anti-akrasia constraint appeals to the seeming Moore-paradoxicality of assertions like '*p*, but it is irrational for me to believe *p*'.[13] It is then taken that the best explanation of the seeming paradoxicality of such assertions is that the relevant contents can never be rationally believed.[14]

It is notable that sometimes the inference '*p* is unassertable, therefore *p* is not rationally believable' goes badly wrong. I can rationally believe *he's disingenuous, but I would never assert that*, but asserting 'he's disingenuous, but I would never assert that' would certainly be odd (it is like saying one never drinks while drinking). It is important to distinguish cases in which one believes an akratic conjunction from cases in which one asserts it. First, while it is easy enough to construct cases in which one believes a proposition *p* but does not know that one believes *p*, it is much harder to think of cases in which a subject genuinely asserts a proposition, but has no epistemic access to the fact that she asserted it.[15] After all, a subject who lacks access to her own belief in *p* typically doesn't go about asserting *p*.

Second, the argument from Moore-paradoxicality only works assuming that the norms governing assertion are the same as those governing belief. But there are reasons to dispute this. Perhaps, for instance, there is a knowledge or sureness norm on assertion, but no knowledge or sureness norm on belief. Compare (1) and (2) below:

(1) The train leaves at noon, though I'm not certain.

(2) Sally believes that the train leaves at noon, though she is not certain.

To many, (1) sounds like a worse indictment of oneself than (2) is of Sally, and one explanation of this is that the norms governing assertion are more stringent than those governing belief. Such a contrast is especially dramatic on the view of Hawthorne et al. (2015) that belief is weak.[16] On their view, suspecting that *p* counts as believing that *p*. But believing *p* while suspecting the belief is irrational does not seem nearly as problematic as flat out asserting *p* and flat out asserting that it is irrational to believe *p*. A suspicion that one's belief is problematic may induce further investigation and reflection on the matter, but need not, it seems, require one to drop that belief. (Consider, for example, a case where one suspects that one's belief in *p* is irrational but knows an easy way to double check whether it is or not. It would hardly seem so odd in this case to retain

[13] See, for instance, Feldman (2005: 108), Bergmann (2005: 424), Hazlett (2012: 211), Littlejohn (2018), and Worsnip (2018). Smithies (2012) focuses more on belief, taking on board the idea that "believing a Moorean conjunction is just as bad as asserting it" (p. 281), but this is something we wish to contest.

[14] The Moorean thought experiments are often constructed as ones in which one believes certain conjunctions. Even if the irrationality of believing the conjunction is conceded, the conclusion that the pair of propositions cannot be simultaneously rationally believed only follows straightforwardly on the assumption that rational belief is closed under logical consequence. But that assumption is not unassailable - consider, for instance, Preface-type cases.

[15] That is not to say that there are no such cases. Think for example about edge cases on a continuum between flat out asserting that *p* and a hedged claim that *p* or a case where one is deaf and is unsure whether one actually produced the words one intended to produce.

[16] We don't wish to take a stand here on whether that view is ultimately correct.

the belief but carry out the double-checking process in order to confirm or disconfirm one's suspicion.) By contrast, a flat out assertion that a certain belief is irrational seems to represent oneself as having settled the question whether one's belief is irrational and so creates far greater self-induced pressure to drop the belief. Similar remarks will apply to any view on which belief may indicate a rather more tentative attitude to a question than (sincere) flat out assertion.

But further, the kinds of considerations salient in the counterexamples above mitigate the paradoxicality of the assertions that fit standard templates for Moore-paradoxicality. Assume that you become persuaded by a very stringent theory of belief that belief requires absolute certainty, and that it is in practice impossible for ordinary mortals to be certain enough of contingent truths to count as genuinely believing them. If you then assert 'It is raining, but [of course] I don't *believe* that', your strange views of belief mitigates much of the strangeness of this assertion. Similar points can be made in connection with the examples given above. For instance, in *Unger Games* a subject falsely but rationally believes that knowledge is unachievable. If she then asserts 'It is raining, but I don't *know* that', the paradoxicality of the assertion will again be mitigated if we bear in mind her stringent view of knowledge.

There are thus several reasons to be skeptical of attempts to draw out general norms on belief based on considerations having to do with the seeming paradoxicality of certain assertions.

II.2 *Dubious luck*

Horowitz (2014) sharpens the case against rational akrasia, going beyond standard Moorean arguments. She addresses akratic states that involve believing a proposition *p*, while also believing that one's evidence doesn't support *p*. We will, at least while discussing her paper, follow Horowitz in thinking that an agent is rationally permitted to believe a proposition if and only if that agent's (total) evidence supports that proposition, and follow her in thinking that evidence supports a proposition just in case it makes it sufficiently likely. In this case, rational akrasia would require that an agent's evidence be radically misleading regarding itself in the following way: it makes *p* likely, while making it likely that it doesn't make *p* likely.

In fact, the kind of case Horowitz considers is one in which the evidence makes *p* likely, while making it likely that it makes *p unlikely*.[17] Consider, then, a subject who believes *p*, while believing that her evidence supports ¬*p*. If such a subject could be rational, Horowitz reasons that the subject could come to rationally believe that her evidence regarding *p* is misleading (she assumes that, at least in the case she is imagining, a subject will be in a position to rationally believe something entailed by a pair of things each of which she rationally believes[18]). But how could it be rational to believe *p* on the basis of evidence one takes to be misleading regarding *p*? Supposing that such a subject could be rational, Horowitz similarly reasons that this subject could

[17] Such evidence thus generates a more radical kind of akrasia than the kind we have just been considering: the agent does not merely believe that the evidence does not support *p,* but believes that the evidence supports ¬*p*.

[18] Note that this assumption cannot be vindicated by multi-premise closure, since the status of being likely on the evidence does not respect such a general form of closure.

rationally conclude that she got lucky in coming to truly believe *p*, despite having evidence that fails to support *p*.

The kinds of considerations discussed above can be brought to bear on Horowitz's arguments. Consider cases involving lack of access to one's own beliefs. In order to rationally conclude, by using the kind of reasoning Horowitz envisages, that one got lucky in arriving at the truth regarding whether *p* despite having evidence that fails to support *p*, one must believe that one believes *p*. But the subjects in all of the counterexamples in our first class lack precisely this kind of access to their own beliefs. Indeed, when Horowitz (2014: 727) describes the oddness of believing that one's evidence regarding *p* is misleading, she claims that the person who believes this can point to a "a particular belief state of his that is, he thinks, unsupported by his total evidence". But in cases where access to the relevant belief is lacking, this is not something that the agent can do.

Various other attempts to bring out just why specific (putative) forms of incoherence are bad also make implicit or explicit assumptions about access. For instance, Worsnip (2018) motivates a requirement prohibiting epistemic akrasia on the grounds that it is difficult for epistemically akratic subjects to make sense of themselves. But in pressing this point, he considers a case in which a subject *recognizes* that she is epistemically akratic, which of course requires her to have access to her own mental states – it requires her to know, or to at least truly believe, that she both believes *p*, and believes that she lacks adequate evidence to believe *p*. Being epistemically akratic, Worsnip argues, at least involves *taking* oneself to be believing against the evidence. But if I have no access to the fact that I believe *p*, I don't seem to be taking myself to hold a belief that goes against my evidence.

Let us now turn to our second theme from section one, and see how faulty epistemic theories—in this case, false theories of evidence—can also make akratic states less puzzling. Here is a case that resembles those already discussed:

Anscombe: Following Elizabeth Anscombe, you believe that intending to ϕ often gives you an immediate kind of knowledge ("practical knowledge") that you will ϕ – and hence, rational belief not based on any evidence. In fact, you think that evidence is completely irrelevant for such knowledge: there are cases in which you know that you will ϕ by so intending, even though it is likely on your evidence that you won't. You know that you intend to go for a run. And you know, and (may even know that you know), that you will go for a run. However, you also believe that it is likely on your evidence that you won't go for a run, for you falsely hold a phenomenal conception of evidence, and believe that the phenomenal evidence you have makes it likely that you won't go for a run. Your belief about your evidence is false: it is in fact likely on your evidence that you will go for a run.

Assume that the subject in such a case concludes that her evidence regarding whether she will go for a run is misleading. Given what we know about the case, her believing this doesn't (pace Worsnip) seem that bizarre, nor does the reasoning strike us as irrational: though her belief in a

phenomenal conception of evidence might be rational, it is false.[19] Her evidence is not in fact restricted in the way she thinks it is. In short, insofar as one is confident in a kind of anti-evidentialist foundationalism according to which one can know *p* without having any evidence for *p*, there may be no internal tension in believing *p* while conceding that one's evidence points in a different direction. Given that one can rationally hold a false view of evidence, the apparent paradoxicality of believing that one's evidence for *p* is misleading tends to fade.

II.3 *Sleepy detectives and Williamsonian akrasia*

Horowitz (2014) does not, in fact, deny the possibility of rational epistemic akrasia. She argues only that paradigmatic instances of epistemic akrasia are irrational, but allows for the rationality of more eccentric instances. The paradigmatic instance she considers comes from a case she terms *Sleepy Detective* (p. 719), while the eccentric one comes from a case she terms *Dartboard* (p. 736), which is inspired by the clock cases in Williamson (2010) and Williamson (2014). Horowitz argues that these cases are different in several ways, and thus that different verdicts about them are warranted. Here are the cases Horowitz presents:

Sleepy Detective: Sam is a police detective, working to identify a jewel thief. He knows he has good evidence—out of the many suspects, it will strongly support one of them. Late one night, after hours of cracking codes and scrutinizing photographs and letters, he finally comes to the conclusion that the thief was Lucy. Sam is quite confident that his evidence points to Lucy's guilt, and he is quite confident that Lucy committed the crime. In fact, he has accommodated his evidence correctly, and his beliefs are justified. He calls his partner, Alex. "I've gone through all the evidence," Sam says, "and it all points to one person! I've found the thief!" But Alex is unimpressed. She replies: "I can tell you've been up all night working on this. Nine times out of the last ten, your late-night reasoning has been quite sloppy. You're always very confident that you've found the culprit, but you're almost always wrong about what the evidence supports. So your evidence probably doesn't support Lucy in this case." Though Sam hadn't attended to his track record before, he rationally trusts Alex and believes that she is right—that he is usually wrong about what the evidence supports on occasions similar to this one.[20]

Dartboard: You have a large, blank dartboard. When you throw a dart at the board, it can only land at grid points, which are spaced one inch apart along the horizontal and vertical axes. (It can only land at grid points because the dartboard is magnetic, and it's only magnetized at those points.) Although you are pretty good at picking out where the dart has landed, you are rationally highly confident that your discrimination is not perfect: in particular, you are confident that when you

[19] Of course those philosophers who think that the phenomenal conception of evidence is correct might take this to show that one can know *p* while knowing that the evidence does not support *p*.

[20] Horowitz does not explicitly say within the vignette that the Sleepy Detective believes that her belief that Lucy is a thief is not supported by the evidence. But it is clear from the surrounding discussion that we are to impute this belief to the detective.

judge where the dart has landed, you might mistake its position for one of the points an inch away (i.e. directly above, below, to the left, or to the right). You are also confident that, wherever the dart lands, you will know that it has not landed at any point farther away than one of those four. You throw a dart, and it lands on a point somewhere close to the middle of the board.

Supposing that the dart lands at point <3,3>, Horowitz says that you should be certain that the dart landed on either <3,3>,<3,2>, <2,3>, <4,3>, or <3,4>, but that you should be highly uncertain as to which of those points in particular was hit.[21]

We do not here want to take a stance on whether or not epistemic akrasia is rational in the *Sleepy Detective* case: we think the matter hinges on further details.[22] Consider, for instance, the following variant of the case. There are ten suspects S1-S10, and for each of S1 to S9, Sam remembers that they have an alibi, and remembers that Lucy (S10) doesn't. Alex then tells Sam: "About half of the time in the past when you knew that only one person lacked an alibi and did some late-night reasoning, you mixed up who had alibis and who didn't." It's hard to see why this testimony would indict any of the particular pieces of memory about who had alibis (just as general Preface claims do not indict any particular claim in a book). But if so, then even after receiving Alex's testimony, it will still be likely on Sam's evidence that Lucy was the thief. And if knowledge is preserved under conjunction introduction, he will also know a proposition that makes it likely that Lucy is the thief (namely, the proposition that all of S1-S9 have alibis, whereas Lucy doesn't). Meanwhile, if Sam does not know that he knows each of the memory beliefs, then upon receiving the testimony it may be likely on his evidence that he does not know the conjunction in question. Indeed, this may be a case in which Sam knows something that makes it highly likely that Lucy is the thief, but it is nevertheless likely on his evidence that it is unlikely on his evidence that Lucy is the thief. Other ways of filing in the details, however, may induce different verdicts. Irrespective of how the details of *Sleepy Detective* are spelled out, we want to look in more detail at whether cases like Ring can be singled out as special. The following are to our mind the two most important potential disanalogies between acceptable and unacceptable cases of epistemic akrasia that Horowitz points to.

First, Horowitz argues that in the two cases the akratic states are produced by different kinds of uncertainty. In Sleepy Detective Sam knows what his evidence is, but is uncertain of what it supports. By contrast, in Dartboard you do not know what your evidence is, but the case need not involve any uncertainty about evidential support-relations. However, it is not clear if anything about these cases forces these verdicts on us. Our worry is that the impression that Sam knows

[21] In adapting Williamson's case to the credence framework, Horowitz assumes that knowing *p* makes credence 1 in *p* rational, and more generally that rational credence matches what Williamson calls 'evidential probability'. Williamson himself is a little cagey about the relation between evidential probability and rational credence.

[22] We think it is important to distinguish questions regarding evidential probability and knowledge, on the one hand, and questions regarding whether a subject manifests good, knowledge-conducive dispositions (see Lasonen-Aarnio 2010, 2020, Forthcoming A, B). We think the notions of rationality and justification, as standardly deployed, cannot carry the theoretical burdens assigned to them: for any evidence-tracking success (knowing, proportioning one's beliefs or credences to the evidence), there are both cases of succeeding despite manifesting problematic dispositions and cases of failing despite manifesting good ones.

what his evidence is comes from deploying a casual notion of evidence, one on which knowing what clues you have without knowing what you know on that basis still counts as knowing what your evidence is. By contrast, Horowitz analyzed Dartboard with a stricter notion of evidence (deployed by Williamson), one according to which knowing that you have a visual experience without knowing what you know on its basis counts as not knowing what your evidence is. But we can flip which framework is used to evaluate each case. One could claim that in Sleepy Detective Sam doesn't know what his evidence is, either because there is something he knows about the case but does not know he knows, or because there is something he doesn't know about the case but doesn't know that he doesn't know. Indeed, Sam's late-night reasoning might tend to be sloppy because he either takes himself to know facts he doesn't know, or because he discounts known facts. And given a casual notion of evidence, one could claim that in Dartboard you do know what your evidence is (because you know that your evidence is your visual experience). Moreover, even given a situation sufficiently formalized to distinguish between uncertainty about what an agent's evidence is and uncertainty about what an agent's evidence supports, Horowitz does not provide any reason to think that such a distinction has bearing on whether or not an instance of epistemic akrasia is rational. To our eyes, both sorts of uncertainty seem on a par.

The most important disanalogy Horowitz points to is that in Sleepy Detective Sam's evidence is truth-guiding, whereas in Dartboard your evidence is *falsity-guiding.* She writes,

> In cases like Sleepy Detective, our evidence is usually "truth-guiding" with respect to propositions about the identity of the guilty suspect (and most other propositions, too). By this I mean simply that the evidence usually points to the truth: when it justifies high confidence in a proposition, that proposition is usually true, and when it justifies low confidence in a proposition, that proposition is usually false. If a detective's first-order evidence points to a particular suspect, that suspect is usually guilty. If it points away from a particular suspect, that suspect is usually innocent. (2014: 738)

> In Dartboard, however, the evidence is not truth-guiding, at least with respect to propositions like Ring. Instead, it is falsity-guiding. It supports high confidence in Ring when Ring is false—that is, when the dart landed at <3,3>. And it supports low confidence in Ring when Ring is true—that is, when the dart landed at <3,2> , <2,3> , <4,3> , or <3,4>. This is an unusual feature of Dartboard. And it is only because of this unusual feature that epistemic akrasia seems rational in Dartboard. You should think that you should have low confidence in Ring precisely because you should think Ring is probably true—and because your evidence is falsity-guiding with respect to Ring. Epistemic akrasia is rational precisely because we should take into account background expectations about whether the evidence is likely to be truth-guiding or falsity-guiding. (2014: 738)

We note that—if read flat-footedly—the claim that the evidence supports low confidence in Ring when Ring is true is false (as is often the fate of flat-footed interpretations). The evidence supports low confidence in Ring in nearly all situations—whenever the dart does not land on <3, 3>. Of the many situations in which the evidence supports low confidence in Ring, Ring is true in only four. If the dart lands nowhere near the ring the evidence will support low confidence in Ring and Ring will duly be false. Of course, these cases are ruled out by the agent's evidence, and so Horowitz's

claim must be relativized to epistemically possible cases.[23]

Let us say that a proposition is *falsity-guided* relative to a range of situations when it is supported by evidence in one of those situations just in case it is false. The hypothesis, then, is that rational high confidence in a proposition of the form '*p* and *p* is unlikely on my evidence' only ever occurs in Williamson style cases when *p* is a falsity-guided proposition relative to the possibilities that are live in that case, and in that sense the counterexamples are eccentric. However, although the instance of epistemic akrasia Horowitz considers does involve a falsity-guided proposition, that is an inessential aspect of the case. We can construct models in which one's evidence justifies high confidence in '*p* and my evidence supports ¬*p*' but where *p* is true.[24]

Here is a simple case. Suppose a pointer is pointing to one of 100 points arranged in a circle (which for convenience we shall number 1-100) with uniform prior over the points. The subject's margin-for-error is plus/minus 3. A pointer is pointing to 10, and thus the strongest thing the observer knows is that it is in the range 7-13. Now consider the proposition—call it Gappy—that the pointer is pointing to either 7, 8, 10, 12 or 13:

… 4 5 6 **7** **8** 9 **10** 11 **12** **13** 14 15 16 …
@

Gappy isn't falsity-guiding relative to the worlds epistemically accessible from @. At the actual world 10 Gappy is both likely on the evidence and true. Yet, at the actual world the proposition *Gappy is true and unlikely on the evidence* is likely (namely, 4/7 likely, since this akratic conjunction is true at worlds 7, 8, 12, and 13). In sum, Gappy is true and it is rational for the observer to be over .5 confident that Gappy is true and unlikely on the evidence.[25]

In response, one might suggest that being falsity-guided relative to a range of situations should be construed in a slightly more lenient way: falsity-guidedness doesn't require that *all* of the situations in which a proposition is true are ones in which it is unlikely, but just that it be unlikely in *most* of the situations in which it is true. And on this more lenient way of construing being falsity-guided, Gappy is falsity-guided (as is, of course, Ring).[26] The first thing to note is that if this is how falsity-guidedness is understood, the claim that rational confidence in a proposition of the form '*p* and *p* is unlikely on my evidence' only ever occurs in cases where *p* is a falsity-guided proposition follows trivially: it is indeed impossible for there to be a proposition *p* that is not falsity-guided in this sense and such that *p and p is unlikely on my evidence* is likely. For assume that this akratic conjunction is likely (at some world *w*). Then, in most of the worlds

[23]Since the agent's evidence doesn't rule out those cases not being ruled out (since the agent's evidence is uncertain about what the agent's evidence is) it remains to be seen if a restriction to epistemically possible cases is dialectically appropriate.

[24] We are not the first ones to argue that the points Horowitz makes based on the example involving *Ring* do not generalize: see also Weatherson (2019, Ch 10) for different, more complex counterexamples.

[25] The case still has the following feature: at all accessible worlds at which Gappy is false (namely, 9 and 11), it is likely. But this is also an inessential feature of the case (see fn. 28).

[26] We are indebted to an anonymous reviewer for this suggestion.

accessible from *w*, *p* is both true and unlikely. But *p* cannot then be likely in most of the accessible worlds where it is true. No one can quibble here: the notion of being falsity-guided has now been defined in a way that straightforwardly entails the hypotheses that cases in which the akratic conjunction is likely are special in that they involve falsity-guided propositions. The project was to find a feature that distinguishes the putatively eccentric cases in which epistemic akrasia can be rational. But first, we are not sure how helpful it is to point to properties that, as a matter of logic, will be had by any case of rationally believing an akratic conjunction. Second, we can still construct cases in which the akratic conjuncts are individually likely, but the relevant proposition that appears in each conjunct (in one case as the conjunct itself, in the other as a topic for normative appraisal) is not falsity-guided even in the more lenient sense.

Consider a case that is otherwise like the one just described, but where the subject's margin for error is plus/minus 2 instead of 3 (and hence, the strongest known proposition is that the pointer is in the range 8-12). And let Gappy* be the proposition that the pointer is pointing to either 8, 10, 12, or 14:

… 6 7 **8** 9 **10** 11 **12** 13 **14** …
@

Gappy* is true, and likely on the evidence (3/5 likely). And Gappy* is likely to be unlikely (with probability 3/5 of having probability less than ½, since it is unlikely at worlds 8, 9, and 11). Yet, Gappy* is likely in two of the three accessible worlds at which it is true (since it is likely at 10 and 12) - and, further, unlikely at all of the accessible worlds in which it is false (namely, worlds 9 and 11). Hence, Gappy* is not falsity-guided even in the more lenient sense.

In sum: Horowitz hopes to cordon off cases of rational epistemic akrasia by providing a condition showing that they are somehow eccentric, involving falsity-guided propositions. We have seen that falsity-guidedness can be defined in a way that logically entails that an akratic conjunction of the form '*p* and my evidence supports ¬*p*' (or of the form '*p* and my evidence doesn't support *p*') cannot be likely on the evidence. But there will still be cases in which one's evidence makes both a proposition *p* and the proposition that the evidence supports ¬*p* likely. Hence, there are cases in which an akratic combination of states is rational, but the relevant proposition is not falsity-guided.

Setting aside the notion of falsity-guidedness, it might be worth discussing a further potential epistemic defect of the cases so far discussed. The reader might point out that in none of the models discussed is the relevant proposition (Ring, Gappy, Gappy*) *known*. Let us say that a proposition *p* is *ignorance-dependent* across a range of situations when there are situations in which '*p* and *p* is unlikely on my evidence' holds, but they are all ones in which *p* is not known. One might hypothesize, then, that one can only ever have rational high confidence in an akratic conjunction of the form '*p* and *p* is unlikely on my evidence' if *p* is ignorance-dependent.

First, even this conjecture is clearly false in simple models in which discriminatory powers (and hence, margins for error) vary from world to world. Again, suppose that a pointer is pointing

to one of 100 points arranged in a circle (which for convenience we shall number 1-100) with uniform priors over the points. If the pointer points to 15, one's discriminatory powers are rather good: one knows that the pointer points to either 14, 15, or 16. But were the pointer to point anywhere else, one's discriminatory powers would be *very* bad: one could rule out the pointer pointing to the maximally distant point on the very opposite side of the circle, but nothing else. Let p be the proposition that the pointer points to either 14, 15, or 16. Assume that one is lucky, and the pointer actually points to 15: one knows p, and p is certain on the evidence. However, it is likely (⅔ likely) that p is very unlikely on the evidence. In fact, the conjunction *p and it is likely that* $\neg p$ is likely.

Further, the conjecture regarding ignorance-dependence is false even in some models involving constant margins for error. Suppose that the epistemic possibilities are suitably modelled as a finite 3-dimensional space. When a person is at a point in the space, the margin for error is given by a radius that generates a sphere around the point. Suppose the margin for error, relative to some unit of measure, is 1. The strongest thing the person knows, then, is given by a sphere of radius 1 around the point that the person occupies. Call the proposition that the person occupies one of the points in that sphere 'Sphere'. The person knows Sphere. Consider an inner sphere generated by a radius of 0.7 from the point that the person occupies. Subtract the inner sphere from the initial sphere. That gives us a thick crust of the sphere. Call the proposition that the person is somewhere in the crust 'Crust'. Most of the volume of the sphere is taken up by the crust. Thus the evidential probability (assuming uniform priors across regions of equal volume) of Crust is high; the probability of Crust is high conditional on Sphere, and Sphere is the strongest thing known. But here is a feature of every point c in the crust: each sphere centred on it with radius 1 will intersect the original sphere in such a way that most of the sphere around c will be outside of the original one. Thus if the agent is at any point in the crust, Sphere is unlikely to be true. We get the result that the conjunction *Sphere and Sphere is unlikely* is a proposition that is likely to be true. But notice that Sphere is known. Sphere is not ignorance-dependent, but nevertheless figures in an akratic conjunction that calls for rational high confidence.[27] There is no special sense in which epistemic akrasia is only rational for propositions that are falsity-guiding or even ignorance-dependent; subject to the construction of the model the relevant proposition may be neither.[28]

[27] Even stronger results follow in higher dimensions. In such cases, uniform discriminatory abilities will lead to the strongest proposition one knows being that one is somewhere in an n-dimensional hypersphere. In hyperspheres, the probability that the strongest proposition one knows is supported by one's evidence can approach 0. The greater the dimensions, the greater proportion of the original hypersphere the hypercrust can be (where the hypercrust is wholly within the hypersphere and such that at every point in it, the hypothesis that one is in the original hypersphere is unlikely on the evidence).

[28] Variants of the sphere example are also helpful for showing the inessentiality of various other kinds of falsity-guidedness to akrasia Here is one possible notion, distinct from either of the ones discussed above: A proposition is falsity-guided iff either it is unlikely at all of the accessible worlds where it is true, *or* likely at all of the accessible worlds where it is false. In the Gappy example, Gappy is not falsity-guided in the original sense but *is* falsity-guided in this new more lenient sense. But now consider the disjunction of the centre point and Crust, and let new proposition be Crust*. Suppose one is at the centre. Crust* is true and likely. So Crust* is likely at some accessible world where it is true. (And if you want the measure of the worlds where it is true and likely to be non-zero, then include a tiny sphere rather than merely the centre pont.) Further, it is unlikely at some accessible worlds where it is false: consider points within the sphere that are very close to but not in the crust. So there is no neat correlation either between the

In sum, we doubt that the plausibility of rational epistemic akrasia in Dartboard can be confined to a highly restricted class of eccentric cases.

II.4 *Justification-Knowledge links*

The proponent of anti-akratic epistemology might rely on some systematic structural connection between rationality and knowledge. The simplest such connection is the radical idea that a collection of beliefs is rational if and only if all the beliefs constitute knowledge. Assuming knowledge to entail rational belief, on such a view one cannot rationally both believe *p* and believe that believing *p* is not rational, for these propositions cannot be simultaneously known. If one knew *p*, then since knowledge (let us assume for now) entails rational belief, it would be false that believing *p* is not rational - and hence, one could not know that believing *p* is not rational. However, as we noted earlier, anti-akratic principles lose their distinctive interest in a setting where there is a general factivity requirement on rationality. There are, however, alternative structural connections between justification and knowledge that one might hope to exploit, which we now turn to.

II.4.1 *Duplication Theories*

A more interesting strategy is suggested by a range of views that frame rational (or justified) belief in terms of knowledge without invoking factivity. A prominent instance of such views relies on a duplication test on rationality. According to the simplest account, a belief in *p* is rational just in case it could be that a duplicate agent knows *p*.[29] These views have the consequence that unknowable propositions cannot be rationally believed. Assuming knowledge to entail rational belief (as these views commonly do), and that knowledge distributes across conjunction, the conjunction *p and it is not rational to believe p* cannot be rationally believed, as it is unknowable.[30] Of course, being epistemically akratic does not require believing the above conjunction, but merely believing both of the conjuncts. However, given that duplication is an equivalence relation, facts of possible knowledge by duplicates put interesting constraints on what propositions can be jointly rationally believed. Suppose that the agent is rationally epistemically akratic: she rationally believes *p*, and rationally believes that it is not rational to believe *p*. Then, by the duplication account, there is some duplicate agent A_1 who knows *p*, and some duplicate agent A_2 who knows that it is not rational to believe *p*. But since duplication is an equivalence relation, A_1 and A_2 must be duplicates. We have reached a contradiction: since A_2 has a duplicate who knows *p*, by the

unlikelihood of Crust* and its truth, nor between the likelihood of Crust* and its falsehood. Crust* is not falsity-guided even in this more lenient sense.

[29] For related views, see Bird (2007), Ichikawa (2014), and (to a lesser extent) Hirvelä (*manuscript*).

[30] Lasonen-Aarnio (2010, *forthcoming* A) has argued that given the most promising ways of construing much talk of rational, justified, or reasonable belief, knowledge is orthogonal to rational belief: there are cases of "unreasonable knowledge". If that is right, then the duplication strategy as a way of arguing for the irrationality of epistemic akrasia is particularly unpromising, for it assumes knowledge to entail rational belief.

duplication account it is rational for A_2 to believe *p*. Hence, A_2 cannot know that it is not rational to believe *p*, for knowledge is factive.[31] It follows that the agent cannot both rationally believe *p*, and rationally believe that it is not rational to believe *p*.

It is not clear how to extend even the simple duplication view to the second anti-akratic idea—that it is never rational to both not believe *p* and also believe that not believing *p* is rationally forbidden—for the view does not immediately deliver an account of when not believing *p* is rationally forbidden. (It obviously won't do to say that not believing *p* is rationally forbidden when and only when a duplicate knows ¬*p*. Consider the fact that you are rationally forbidden to form a belief about the outcome of a fair coin-flip.) But even the suggested argumentative strategy for defending the first akratic principle is rather suspect.

The simple version of duplication views appeals to full metaphysical duplication, checking for rational belief in a proposition by checking whether some perfect duplicate knows that very proposition. But one would have thought, for example, that it is possible to rationally believe metaphysical impossibilities—for instance, that Hesperus is not Phosphorus, or that thinkers have no proper parts (perhaps Descartes was rational in believing this!). A related worry is that rational false beliefs regarding any of the facts regarding a subject that are shared by her duplicates are impossible. A person's duplicates share all of their intrinsic properties.[32] Intrinsic properties presumably specify an agent's neuronal patterns, but it should surely be possible for an agent to hold a rational false belief about the microphysical properties of her neurons.

Possible fixes will either run into similar problems or complicate the argument for anti-akrasia or both. Perhaps, for example, the relevant kind of duplication just involves duplicating one's mental states (as opposed to perfect metaphysical duplication). But even setting aside the issues of whether knowledge counts as a mental state, we think it can be perfectly rational for subjects to hold false beliefs about some of their mental states, such as their beliefs.[33] Another kind of refinement—one that obviously helps with the Hesperus-Phosphorus example—appeals to counterpart beliefs, the rough idea being that to rationally believe *p* is to be such that a duplicate has a counterpart belief that is knowledge, where the counterpart belief might be in a different proposition.[34] While such a move makes the resulting view of rationality rather more plausible, it significantly complicates the possibility of mounting an argument for the irrationality of epistemic akrasia. In order to allow for rationally believing metaphysical impossibilities, counterpart beliefs may be in different propositions. But then it is not at all clear whether the resulting accounts can deliver the result that it can never be rational to both believe *p* and that believing *p* is rationally forbidden, for that result relied on the impossibility of jointly believing specific propositions that cannot be jointly known for reasons having to do with the factivity of knowledge.[35]

---

[31] We are indebted to a discussion with Jeremy Goodman.
[32] Lewis (1983).
[33] See Srinivasan (2015).
[34] Bird (2007: 87) and Ichikawa (2014: 194) appeal to counterpart beliefs in order to avoid the counterintuitive result concerning necessarily false propositions.
[35] As well as the fact that *p* is not held fixed, the content of 'rational' may vary as well across duplicates. Suppose, for example, that 'rational' is vague and picks out different relations according to slight variations in use by one's

In summary, we doubt whether any version of the duplication test will be both independently plausible and anti-akrasia entailing.

II.4.2 *Justification, Closure and the Possibility of Knowledge*

The simple duplication theory is but one instance of a class of theories according to which rationally believing *p* entails the possibility of knowing *p*. Call this the 'Possible Knowledge Principle'. That barebones claim, no matter how it is embedded in further theory, generates interesting results regarding akrasia. In particular, given (i) the factivity of knowledge, (ii) the distribution principle that necessarily, if one knows *p* and *q* one knows *p* and one knows *q*, and (iii) the principle that necessarily, if one knows *p*, one rationally believes *p*, we get the impossibility of rationally believing a conjunction of the form:

C: *p* and it is not rational for me to believe *p*.

Note that this line of thought does not require a metaphysical modality. So long as 'Possibly' and 'Necessarily' are interpreted as duals and the principles of factivity, distribution and knowledge-to-rationality are plausible for the relevant notion of necessity, the argument will go through.[36]

We can immediately see a further result. Suppose we accept the closure principle that necessarily, if it is rational to believe *p* and it is rational to believe *q*, then it is rational to believe *p* and *q*. Then rational akrasia of the form we have been discussing will also be impossible. For in that case rationally believing *p* and rationally believing that it is not rational to believe *p* will entail the rationality of believing a conjunction of form C. Since it is implausible to reject either the principle that knowledge distributes over conjunction or factivity, the main options for making room for akrasia in a setting where the Possible Knowledge principle is accepted will be either (i) denying that knowledge entails rational belief or (ii) to denying closure.

One of us has explored (i) at length elsewhere and, for reasons of space, this is not the place to rehearse the considerations adduced in that paper.[37] But we shall dwell on (ii) a little further. Even if multi-premise closure holds for knowledge, it is not hard to construct interesting toy models of justification on which justified belief entails the possibility of knowing but for which closure fails dismally. Suppose, for example, that one is justified in believing *p* just in case for all one knows one knows *p* (i.e. one doesn't know that one doesn't know *p*).[38] Let's deploy a standard framework according to which one knows *p* iff there is no epistemically accessible world where

community. Assuming that setup, various counterpart beliefs are not about rationality but about a slightly different relation, rationality*.

[36] Note that even on an epistemic possibility gloss, the *Unger Games* scenario is now blocked, as it is not epistemically possible that one knows that one knows nothing.

[37] Lasonen-Aarnio (2010). See also Lasonen-Aarnio (2008) for a discussion of closure.

[38] Note that if we want a notion of being justified in believing that entails belief (which we do in the present context of exploring akratic combinations of beliefs), then we may wish to add an additional belief requirement. After all, given that believing is not luminous, one's commitment to a proposition *p* might fall short of belief in *p* and yet, for all one knows, one knows *p*.

¬*p*.[39] (Again it doesn't matter whether the worlds are metaphysically possible or not.) Then it is easy enough to construct akratic models in which there is an epistemically accessible world where one knows *p* (and so one is in fact justified in believing *p*) and an epistemically accessible world where one knows that one knows that it is not the case that one knows *p* - and so one is also justified in believing that one is not justified in believing *p*.

Of course the model of justification just stated does not even preclude being simultaneously justified in believing *p* and being justified in believing ¬*p*, and the theorist may wish to posit constraints on accessibility that preclude this. The natural way to do this is to impose a convergence constraint on the accessibility structure so that when two worlds w_1 and w_2 are each accessible, there is a world w_3 that each of them can access. This corresponds to what is commonly known as the .2 axiom in modal logic, which says that anything that is possibly necessary is necessarily possible.[40] Then, there can never be a pair of accessible worlds one of which is such that one knows *p* in it and the other of which is such that one knows ¬*p* in it: the accessibility constraint would require some world that each world sees, but since the know-*p*-world would only see *p*-worlds and the know-¬*p*-world would only see ¬*p*-worlds, this constraint could not be satisfied. Still, even with that constraint in place, there is no block on being justified in believing *p* while also being justified in believing that one is not justified in believing p. To be justified in believing *p* we need an accessible world where one knows *p*, and to be justified in believing that one is not justified in believing *p* we need an accessible world where one knows that one knows that one does not know *p*. But it is perfectly consistent with this to suppose that there is a world that each of these two worlds sees—it can be a world where *p* is true but knowably unknown. So the natural accessibility constraint blocking justifiably believing each of a contradictory pair of propositions (blocking the truth of both J*p* and J¬*p*) imposes no block on akratic pairs (it doesn't block the truth of both J*p* and J¬J*p*).

Such blocks require additional accessibility constraints. One way of blocking the truth of both J*p* and J¬J*p* is to additionally impose the KK principle (i.e. a transitivity constraint on accessibility). This principle says that knowledge is luminous. (That will mean that there will be no worlds that a K*p* and a K¬K*p* world both see, since, given KK, a K*p* world only sees K*p* worlds. So the earlier trick for respecting the .2 axiom is unavailable.) Another principle that will preclude rational akrasia is a luminosity constraint on justification itself, i.e. the principle that if J*p* then

[39] This of course involves a bit of idealization. A proposition *p* may be true at all epistemically accessible worlds and yet one does not know *p* because one has not come to know what one is in a position to know. (Relatedly, one might intuitively lack justification because, while in fact one doesn't know that one doesn't know *p*, one is in a position to know that one is not in a position to know *p*). For that reason 'K' is sometimes interpreted as 'being in a position to know' in these models (see, for example, Rosenkrantz (2018) and Dorst (2019), where justification can also be given a position-to-know theoretic gloss). Furthermore, in some cases, *p* may be true at all epistemically accessible worlds and yet one not even be in a position to know *p* because one is unable to grasp *p* (see Rosenkrantz (2018), who also has his own distinctive way of dealing with cases involving ungraspable proposition). We shall continue to operate within the more idealized setting where we ignore ungraspable propositions and pretend that one knows everything one is in a position to know, as the points that we make in what follows would not be much affected by complicating things.

[40] When the epistemic logic is normal and imposes factivity for K and adds the .2 axiom, we get KT.2 as the epistemic logic. The .2 is called 'G1' in Cresswell and Hughes (1996).

KJ*p*.[41] But such luminosity constraints are highly controversial.[42] Another route is to directly impose the constraint that if J*p* then ¬J¬J*p*, but that would be dialectically uninteresting in the context of this paper.

By way of illustrating our hesitancy about imposing additional constraints on accessibility to block akrasia, we offer a case inspired by the false theory of rationality theme of section one. It presents a challenge for the anti-akratic theorist, adapted now to the knowledge-theoretic conception of justification under consideration. The case illustrates how hard it is to think intuitively about akrasia in this context—prepare yourself for some confusingly iterated epistemic operators. Here goes.

An unmarked clock appears to be pointing to 3, and indeed it is. The margin-for-error is 1. So you know that the clock's hand is somewhere between 2 and 4. (And since K entails J, you J that the clock's hand is pointing between 2 and 4.) A generally reliable person has told you (incorrectly) that the margin-for-error is roughly 2. It is far from unusual for one not only to come to know, but to come to know that one knows various propositions by acquiring beliefs through testimony. And there are no tell-tale signs that this case is any different: For all you know, you know that you know that the margin-for-error is roughly 2, and so for all you know, you know that that you know that you don't know that the clock is pointing to between 2 and 4. (This would not mean that the clock's hand is not pointing between 2 and 4, just that you don't know that it is pointing between 2 and 4.) So you know (and thus have justification for) the proposition that the clock's hand is pointing between 2 and 4, and you don't know that it's not the case that you know that you know that you don't know that the clock's hand is pointing between 2 and 4. So you are justified in believing that you are not justified in believing that the clock's hand is pointing between 2 and 4 (since substituting 'J' for '¬K¬K' in '¬K¬KK¬K*p*' gives us JK¬K*p* and substituting '¬J' for 'K¬K*p*' in the latter formula gives us J¬J*p*). So—interpreting 'J' as 'You don't know that you don't know'—we have a case of akrasia. Yet, there seems to be nothing especially odd about the subject. The case thus provides little intuitive pressure to provide additional accessibility constraints that block the akratic combination suggested by the case. (Think also of a case where one comes to know *p* but a normally reliable person tells you that you won't actually know *p* until you have double checked *p*. Here we can easily generate the same structure based on the assumptions that one knows *p* but for all one knows, one knows that one knows that one does not yet know *p*.) In short, the accessibility frameworks provide intriguing ways of linking the Possible Knowledge principle to anti-akrasia, but we remain skeptical that a case for anti-akrasia can be made along these lines.

[41] As an accessibility constraint this amounts to the principle that what is possibly necessary is necessarily possibly necessary.

[42] Williamson (2000) provides the best known recent case against such luminosity principles. Stalnaker notably disagrees, being very friendly to KK and indeed to the logic 4.2 for knowledge that one gets by adding the .2 axiom just discussed to the characteristic S4 axioms. His 2006 "On Logics of Knowledge and Belief" is a locus classicus for the exploration of the logic of 'For all you know you know'. In recent years, a group of philosophers, in large part former students of Stalnaker, have been working to revive the KK principle. These include Greco (2014), Salow and Goodman (2018), and Dorst (2019).

We have seen that, for the closure-denying theorist, there is all the difference in the world between the hypothesis of conjunctive akrasia (which involves believing C) and the kind of akrasia with which we started this paper (which involves believing each of the conjuncts of C individually). It bears emphasis that the theorist who rejects conjunctive akrasia but not our non-conjunctive version will not be troubled by many of Horowitz's arguments, since those require reasoning from justified belief in conjunctions like C.[43]

The Possible Knowledge principle offers an interesting case against conjunctive akrasia. And in combination with closure for justification it delivers full blown anti-akrasia. Such views are rather out of the spirit of much of the anti-akratic literature. As we have just seen, for example, Horowitz is interested in the limits of rational high confidence, and uses rational high confidence as the heuristic for determining the limits of rational belief. But it is obvious enough that one can have rational high confidence in propositions that one cannot know. Suppose, for example, that one holds a ticket in a 100-ticket lottery. One has rational high confidence in the conjunction expressed by 'I will lose and I don't know that I will lose'. But assuming that knowledge distributes over conjunction, it is not possible to know this conjunction. Epistemological orthodoxy has tended to assume that there is a viable notion of rational belief that is not governed by the Possible Knowledge principle. But in any case, the prospects for defending a general anti-akratic outlook on the basis of that principle does seem especially promising.

Part Three: Idealization and Fixed Point Theses

Defenders of an anti-akrasia constraint might appeal to idealized agents in order to defend some version of their anti-akratic position. It is important here to distinguish between two ways that idealizations may be deployed in epistemology.

First, consider, for example, a Bayesian who deploys an idealization of logical omniscience (and various Bayesian theorems that rely on such an idealization). Such idealizations can be helpful without presuming anything to the effect that mistakes about logic are *ipso facto* lapses in irrationality. For example, they can be helpful as a way of exploring which kinds of epistemic structures can arise even without any failures of logical knowledge, or as a way of exploring how best to bet in a certain kind of situation when every logical consequence is fully in view. It may be particularly messy, for instance, to judge how one is to bet in a situation where competing

[43] It is also worth noting in passing that Horowitz readily assumes something like closure for justification. In the initial paragraph of her paper she glosses akratic states as involving "high confidence in something like *p*, but my evidence doesn't support *p*", but then immediately switches to glossing it in terms of believing each of those conjuncts individually. In the setting of her paper this may be harmless enough. Perhaps insofar as one can devise cases where one has rational high confidence in *p* and also has rational high confidence in the proposition that *p* is not supported by one's evidence, one can devise cases where one has rational high confidence in the conjunction *p* and *p* is not supported by my evidence. (The familiar Bayesian point that high confidence isn't generally closed under conjunction need not pose a problem for this existential thought). But for anyone who believes in the Possible Knowledge principle, the difference between conjunctive akrasia and regular akrasia cannot be ignored.

considerations are complex and one also has a limited grasp of logic. Some illumination will be achieved by factoring out noise from one of the parameters, keeping the competing considerations in place but rendering the agent logically omniscient. (Of course in special cases this won't be possible, e.g. when one is betting on one's level of competence in logic.) However, appeal to such idealization won't further the cause of the anti-akratic: their claim isn't merely that anti-akrasia principles hold in models in which certain sources of uncertainty are ignored.

There is a second spirit in which one might make idealizations. One might think that rationality requires not making any mistakes about logic and think on those grounds that an idealization to the logically omniscient is mandatory insofar as one is exploring how a rationally ideal agent that is not subject to *any* lapse in rationality might proceed. When it comes to mistakes not about logic but about the rationality of one's own beliefs, the defender of anti-akrasia needs to deploy idealizations in this spirit. One might think that it is a requirement of rationality that—at a first pass—one not make any mistakes about the requirements of rationality, and thus that agents who are epistemically akratic are somehow automatically guilty of a failure of rationality.[44] The idea that mistakes about the requirements of rationality are mistakes of rationality has been pursued by Michael Titelbaum (2015), and it is instructive to see how that idea plays out in his hands.[45] Titelbaum recognizes that the idea needs immediate qualification. While he begins with the sweeping idea that all mistakes about the requirements of rationality are mistakes of rationality, he quickly retreats to the thesis that an *a priori* false belief concerning the requirements of rationality is never permitted. (As this principle is restricted to *a priori* matters, call this the *Restricted Fixed Point Thesis*.) Such a retreat is wise because the sweeping thesis is indefensible: Suppose I see what I take to be Anya walking on the other side of the street, but unbeknownst to me it is Anya's identical twin. Since I know that Anya permissibly believes *p*, I believe that the person walking on the other side of the street permissibly believes *p*. But my belief is false. It would be highly implausible to deem me irrational on the basis of my mistake about what Anya's twin rationally believes since the mistake is rooted in a reasonable mistake about their identity.[46]

What, then, should we make of the *Restricted Fixed Point Thesis*, and how does it bear on the rationality of akratic states? Here we would like to make a number of observations. First,

[44] Quite aside from the rationality of epistemic akrasia, it's an excellent question as to whether we should take the standard idealization of logical omniscience in the first or second spirit. We are tempted by a view on which logical relations and operators are part of the world and as such can be amenable to rational error just like any other part. But it is beyond the scope of this paper to pursue that issue properly.

[45] Some related ideas are pursued by Clayton Littlejohn (2018), though for reasons of space we shall confine our attention in the text to Titelbaum's discussion of fixed point theses. One matter worth remarking on, however, is that Littlejohn claims that his version of a fixed point thesis - that "if you believe that you are rationally required to believe *p*, this belief is either true or rationally prohibited" - follows from the "enkratic requirement" that "you don't both believe that you're rationally required to believe *p* and refrain from believing *p*" (p. 261). But the inference is far from straightforward. Even if we accepted the enkratic requirement, we could think a certain belief that believing *p* is rationally required is false but rationally permitted - and the reason that it is false can be that the *combination* of refraining from believing that *p* is rationally required and refraining from believing *p* is rationally permitted.

[46] And here is an example from Titelbaum himself: *s* mistakenly believes that what Frank wrote on a napkin is a requirement of rationality but is wrong because he has reasonable but mistaken beliefs about what was written on the napkin.

Titelbaum's own argumentative path to the *Restricted Fixed Point Thesis* begins from what he calls the *Akratic Principle*:

Akratic Principle: No situation rationally permits any overall state containing both an attitude A and the belief that A is rationally forbidden in one's current situation.

From this point he argues (in part abductively), for the *Restricted Fixed Point Thesis*. Taken at face value the Akratic Principle is precisely the kind of principle we have been arguing against and so, without an independent argument for it, would be of little relevance here. But we should recall that Titelbaum tells us early on that "from now on when I discuss beliefs about rational requirements I will be considering only beliefs in *a priori* truths or falsehoods".[47] So what is really going on is an argument based on an Akratic principle restricted to propositions that can be settled *a priori* (call this the "*Restricted Akratic Principle*"). What should we make of this more modest anti-akratic view and the *Restricted Fixed Point Thesis* that is argued for on that basis? Here we would like to raise four issues.

First, there is an issue defining what counts as a "mistake about the *a priori* truths about rationality". The crux of the problem is to say what it is for a claim to be *about* the requirements of rationality. Suppose that p is any *a priori* knowable proposition. Then,

> q_1: The requirements of rationality are such that p
> q_2: Rational beliefs are rational iff p

will be *a priori*, and if one believes $\neg p$, one will make mistakes about q_1 and q_2, at least insofar as one has opinions about q_1 and q_2 at all. But if either of q_1 or q_2 count as propositions about the requirements of rationality then any mistake about *a priori* matters will induce a mistake about the requirements of rationality. In sum, unless we are given a very refined notion of what it is for a proposition to be about the requirements of rationality, the *Restricted Fixed Point Thesis* will collapse to the more sweeping:

> *A Priori Fixed Point Thesis*
> A mistake about a proposition that is *a priori* true (i.e. *a priori* knowable) is a mistake of rationality.

Second, quite apart from the issue of defining the relevant notion of aboutness, one wonders why the *Restricted Fixed Point Thesis* should be true unless the *A Priori Fixed Point Thesis* is true as well. Suppose, for some mathematical proposition, we convince ourselves that even if a mathematical genius could come to know *a priori* that $\neg p$, it may be rational to believe p through testimony. Why not take the same attitude to *a priori* propositions about rationality? Suppose some

[47] And in a footnote the point of the use of 'situation' in the Akratic Principle is clarified: "... there will be *a priori* truths about which situations rationally permit which overall states. They will take the form "if the empirical facts are such-and such, then rationality requires so-and so", Titelbaum (2015) ftn 27.

sophisticated a priori argument showed that Unger was wrong to believe that a belief is rational iff one has reasons for it. Just as in the mathematical case, it is nevertheless quite natural to think that someone wouldn't automatically be irrational for believing this.

Third, even if we could somehow convince ourselves of the *A Priori Fixed Point Thesis*, that may have limited relevance to the more general questions of this paper concerning the rationality of akratic states. After all, the akratic person described at the outset need not automatically be making a mistake about *a priori* matters. Consider *Good News*—there is no reason to think that the person in that case is making a mistake about *a priori* matters. Similarly, if a Lockean thesis is true but the threshold for belief is not *a priori* knowable, then the above thesis need not make trouble for examples like *Missed It By That Much*.

Fourth, it is far from clear that *Restricted Akrasia* and the more general *Restricted Fixed Point Thesis* have similar motivations to more standard anti-akrasia principles. An initially compelling way to motivate anti-akrasia principles is by appealing to the idea that epistemic akrasia involves a distinctive kind of incoherence. But the fixed point theses discussed locate the trouble in an entirely different place—the problem is holding an *a priori* false belief (about the requirements of rationality). Assume that it is *a priori* knowable that believing *p* is rationally required in my current situation. Then rationality forbids me from believing that believing *p* is rationally forbidden. In fact, insofar as I believe that believing *p* is rationally forbidden, I am *already* irrational, irrespective of whether I am akratic. If I then become akratic by forming a belief in *p,* I cannot be faulted on the basis of a fixed point thesis of a further breach of rationality (in fact, that would seem to make me *more* rational, as I come to believe something I am rationally required to believe). The trouble with violations of the *Restricted Akratic Principle* is not that they involve a kind of incoherence, a failure by one's own standards; it is believing something that contradicts what is *a priori* entailed by what one believes. In sum, Titelbaum's *Fixed Point Thesis* (and *Restricted Akrasia*) stand in need of further justification, and even if they can be justified, they amount to heavily restricted versions of the original anti-akratic idea.

Conclusion

Anti-akratic constraints in epistemology, while popular, are subject to a wealth of counterexamples. Meanwhile, arguments for those constraints—when given in the first place—are surprisingly few, and face resistance from these counterexamples. Attempts to shore up those anti-akratic constraints through idealization are unpromising. The apparent failure of such constraints suggests more general difficulties for the project that attempts to lay down structural requirements of rationality.[48]

Afterword:

[48] We are grateful to Cian Dorr, Julian Dutant, Jeremy Goodman, Jaakko Hirvelä, Clayton Littlejohn, Jeffrey Russell, audiences at USC and Cincinnati, and an anonymous referee for comments on an earlier draft of this paper. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 758539.

Someone tentatively on board with our discussion above might worry that our general theoretical orientation proves too much. If rational mistakes about the nature of rationality might allow someone to be in a rational state of epistemic akrasia, why shouldn't we think that a rational mistake about the nature of negation might allow someone to rationally believe each of *p* and not-*p* (or even their conjunction), and why shouldn't we think that a rational mistake about the nature of the 'It is true that' operator might allow someone to rationally deny '*p* just in case it is true that *p*'? Are we moving toward the conclusion that no proposition is such that it is intrinsically irrational to believe it/deny it and no pair of propositions is such that it is intrinsically irrational to believe/deny both of them? If this is right then the whole project of laying out structural constraints on rational belief begins to look suspect.

While we do not wish to engage fully with these larger issues in this paper – we have achieved our purpose if we have at least got the reader significantly worried about epistemic anti-akrasia principles – we are somewhat sympathetic with these broader conclusions. We give voice to these sympathies in what follows.

Assume that Di believes both a proposition *p*, and believes its negation not-*p*. However, assume further that Di is a dialetheist: she thinks that some contradictions are true, and that *p* and not-*p* is a true contradiction. Perhaps, for instance, she thinks that the sentence 'This sentence is not true' is both true and not true. We can assume this to be a belief that Di has reached as a result of careful philosophical investigation. With this background story in place, Di's beliefs look less irrational, irrespective of whether her belief in dialetheism is true: she holds beliefs that, by the lights of her own responsibly formed belief, are rational. It is notable that John Broome, one of the most prominent proponents of structural requirements of rationality, feels the pull of such considerations, responding to such cases by weakening the condition against contradictory beliefs:

> 'Necessarily, if you are rational, you do not believe *p* and believe not-*p*, unless you believe *p* is special' (Broome 2013: 91)

The dialethist may point out that one cannot read off from the content of a proposition *p* whether it warrants dialethist treatment. The proposition that the proposition asserted by the tallest woman in Alaska is false is fodder for dialethic treatment when asserted by the tallest woman in Alaska (along with no others) but not otherwise. And so there isn't, for the dialethist, a straightforwardly delineable category of 'special' propositions. Here is a refined version of Broome's thought: it *is* incoherent to merely 1) believe *p* and 2) believe not-*p*, but it is *not* incoherent to 1) believe *p* 2) believe not-*p* and 3) believe that for reasons having to do with some special aspect of one's situation, there is nothing irrational about believing *p* and believing not-*p*. (Indeed, some of the

examples above might be seen as on a par with the case involving Di's.)

Insofar as we qualify the structural prohibition on believing contradictions in this way, it is natural, even if we are initially sympathetic to anti-akratic principles, to lay down similar qualifications on anti-akratic principles, thereby abandoning them in their original form. But more importantly, once this first step is made, we think the program of laying down requirements of structural rationality begins to crumble. Broome's retreat encourages us to think that for each putatively incoherent state, a subject can form some further belief that renders her overall mental state coherent. But the resulting view scarcely aligns with the coherence requirements we started out with. This considerably undermines any attempt to support such requirements on intuitive grounds.

The thesis that no proposition or pair of propositions are intrinsically irrational to believe – a thesis that Broome is tending towards conceding as contradictions seem to be pretty much the best case against such a view – is a natural companion to the thesis that there are no conceptual truths. As an illustrative example, Williamson (2006: 9-13) uses semanticists who think either that 'every' entails an existential to illustrate the thesis that it is not a conceptual truth (on any of various standard glosses on conceptual truth) that all vixens are female foxes (since if you accepted that entailment thesis doubts about whether there are any vixens would spread to the universal generalization). And this entailment thesis would similarly recommend caution about the claim that all vixens are vixens. The very same considerations can be used to challenge the claim that it is intrinsically irrational to deny that all vixens are vixens. The general perspective we recommend is that logical relations and operators – along with normative relations and operators – are part of the world and as such can be amenable to rational error just like any other part. Of course an idealization to logical omniscience is still useful for some purposes, as noted above, and an idealization to omniscience about certain other limited subject matters may be similarly useful. But such idealizations will not encode intrinsic requirements of rationality. Our preferred orientation will not end up proving too much. It will rather force us to view various positive proposals about intrinsic rationality with proper suspicion. We end with a tendentious prediction: Theses of intrinsic irrationality for belief will ultimately suffer the same fate as the myth of conceptual truth.

References

Anscombe, G. E. M.,
1963, *Intention*, 2nd edition, Oxford: Blackwell.

Bergmann, Michael
2005 "Defeaters and higher-level requirements", *Philosophical Quarterly*, *55* (220): 419–436.

Bird, A.
2007, “Justified Judging”, *Philosophy and Phenomenological Research, 74*(1), 81-110.

Broome, John
2013. *Rationality Through Reasoning*. Wiley-Blackwell.

Cresswell, M.J. & Hughes, G.E.
1996. *A New Introduction to Modal Logic*. Routledge.

Dorst, Kevin
2019, “Abominable KK Failures”, *Mind* 128 (512):1227-1259.

Feldman, Richard
2005 "Respecting the evidence". *Philosophical Perspectives* 19 (1):95–119.

Foley, Richard
2009. "Beliefs, Degrees of Belief, and the Lockean Thesis". In Franz Huber & Christoph Schmidt-Petri (eds.), *Degrees of Belief*. Springer. pp. 37-47.

Goodman, Jeremy & Salow, Bernhard
2018, “Taking a chance on KK”, *Philosophical Studies* 175 (1):183-196.

Greco, Daniel
2014, “Could KK Be OK?”, *Journal of Philosophy* 111 (4):169-197.

Hawthorne, John., Rothschild, Daniel, and Spectre, Levi.
2015. ‘Belief is weak,’ in Philosophical Studies 173, 1393-1404.

Hazlett, Allan
2012. "Higher-Order Epistemic Attitudes and Intellectual Humility". *Episteme* 9 (3):205-223.

Hirvelä, Jaakko
Manuscript “Justification and the Knowledge Connection”

Horowitz, Sophie
2014 “Epistemic Akrasia”, *Nous*, 48.4: 718-744.

Ichikawa, J.
2014 “Justification Is Potential Knowledge”, *Canadian Journal of Philosophy,* 44.2:184-206.

Lasonen-Aarnio, M.
2008 “Singe premise deduction and risk”, *Philosophical Studies*, Vol. 141, No. 2, pp. 157-173.
2010 “Unreasonable Knowledge”, *Philosophical Perspectives* 24(1): 1–21.

2020 “Enkrasia or Evidentialism? Learning to Love Mismatch”, *Philosophical Studies* 177(3): 597-632.
Forthcoming A “Dispositional Evaluations and Defeat”, in Brown, Jessica and Simion, Mona (eds.), *Reasons, Justification and Defeat,* Oxford University Press.
Forthcoming B “Coherence as Competence”, *Episteme*

Lewis, David.
1983*, “*Extrinsic Properties*”, Philosophical Studies,* 44: 197–200.

Littlejohn, Clayton
2018 “Stop Making Sense? On a Puzzle About Rationality”, *Philosophy and Phenomenological Research* 96.2: 255-513.

McGee, Vann
1985. "A counterexample to modus ponens". *Journal of Philosophy* 82 (9): 462-471.

Rosenkranz, S.
2017 "The Structure of Justification", *Mind, 127*(506), 629-629.

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

Srinivasan, Amia
2015 “Normativity Without Cartesian Privilege”, *Philosophical Issues* 25-1: 273–299.

Stalnaker, Robert
2006 “On Logics of Knowledge and Belief”, *Philosophical Studies*, 128 (1):169-199.

Titelbaum, Michael G.
2015 “Rationality’s Fixed Point (Or: In Defence of Right Reason)”, *Oxford Studies in Epistemology 5*: 253-294.

Unger, Peter
1975. *Ignorance: A Case for Scepticism*. Oxford University Press.

Weatheson, Brian
2019 *Normative Externalism*, Oxford University Press.

Williamson, Timothy
2000. *Knowledge and its Limits*. Oxford University Press.
2006 “Conceptual Truth”. The *Aristotelian Society Supplemental Volume* 80.1:1-41. 2011 “Improbable knowing”. In T. Dougherty (ed.), *Evidentialism and its Discontents*. Oxford University Press.
2011 “Improbable knowing”. In T. Dougherty (ed.), *Evidentialism and its Discontents*. Oxford University Press.

2013 “Gettier Cases in Epistemic Logic”. *Inquiry: An Interdisciplinary Journal of Philosophy* 56 (1): 1-14.
2014 “Very Improbable Knowing”. *Erkenntnis* 79 (5): 971-999.

Worsnip, Alex
2018 “The Conflict of Evidence and Coherence”, *Philosophy and Phenomenological Research* 96.1: 3-44.