Elusive Externalism

Bernhard Salow
Penultimate Draft – Please Cite Published Version

Epistemologists have recently noted a tension between (i) denying access internalism and (ii) maintaining that rational agents cannot be epistemically akratic, believing claims akin to 'P, but I shouldn't believe P'. I bring out the tension, and develop a new way to resolve it. The basic strategy is to say that access internalism is false, but that counterexamples to it are 'elusive' in a way that prevents rational agents from suspecting that they themselves are counterexamples to the internalist principles. I argue that this allows us to do justice to the motivations behind both (i) and (ii). And I explain in some detail what a view of evidence that implements this strategy, and makes it independently plausible, might look like.

Access internalism says that we can always tell, simply by reflecting, whether our beliefs are rational or justified. Evidentialism maintains that our beliefs are rational or justified insofar as they conform to our evidence. Together, the two views require that we can always tell by reflection what evidence we have.

Externalists object that this 'access' requirement on evidence is unsustainable: nothing, or at any rate not nearly enough, could count as evidence if we always have to be able to tell what our evidence is. In particular, for reasons I summarize in §1, all externalists should agree that we must reject the 'negative access' requirement:

Thanks to audiences at MIT, Oxford, and the 2016 Midsummer Workshop for helpful comments. Special thanks to Andrew Bacon, Alex Byrne, Nilanjan Das, Kevin Dorst, Jeremy Goodman, Alex Gregory, Abby Jacques, Matthias Jenny, Brendan de Kenessey, Justin Khoo, Harvey Lederman, Matt Mandelkern, Ram Neta, Agustin Rayo, Ginger Schultheis, Kieran Setiya, Alex Silk, Paulina Sliwa, Declan Smithies, Jack Spencer, Bob Stalnaker, Josh Thorpe, Jonathan Vogel, Roger White, Bruno Whittle, Tim Williamson, Steve Yablo, and several anonymous referees for helpful feedback and discussions.

Negative Access

If P is not part of one's evidence, one has conclusive evidence that P is not part of one's evidence.

But there are also strong reasons to accept access internalism. Some of these, such as the thought that only access internalism can provide epistemic norms able to 'guide' agents when forming beliefs, are too familiar to provide new leverage. Others, however, are more surprising. For example, the recent literature on 'level-connections' in epistemology has brought out that counterexamples to the access constraint on evidence make for cases in which agents have evidence supporting 'epistemically akratic' conclusions akin to P, BUT I SHOULDN'T BELIEVE P.[1] Believing such conclusions, however, looks irrational.

In response to this worry, some externalists argue that akratic beliefs can be rational after all;[2] others conclude that rationality and one's evidence sometimes make incompatible demands, thus rejecting evidentialism.[3] On the face of it, neither option is attractive.

But there is another way. We can reject negative access, accept evidentialism, and nonetheless avoid epistemic akrasia. To do so, we need to challenge the move from someone *having evidence for a proposition* to that person *being able to believe that proposition in line with her evidence* – a move which, I argue, is independently problematic. The move can be blocked in sufficient generality by appeal to a form of contextualism (or relativism, or expressivism) about what counts as part of someone's evidence.[4] On the resulting picture, negative access failures are real but elusive: rational agents can never suspect that they are happening to *them*. This allows us to subscribe to what is attractive in externalism, without accepting its worst excesses.

[1] The thought that an anti-akrasia constraint might require access internalism is explicit in Bergmann (2005), Gibbons (2006), Smithies (2012), White (2014), Lasonen-Aarnio (2015), Worsnip (forthcoming), and Dorst (ms).

[2] This is the response suggested by Lasonen-Aarnio (2014, 2015); it seems to be implicitly endorsed by Williamson (2011), and is explicitly left open by Horowitz (2014). Coates (2012) and Weatherson (ms) argue for this conclusion on slightly different grounds.

[3] Christensen (2007, 2010) floats such a response. Worsnip (forthcoming) advocates it explicitly, though he allows that there are some normative notions of which evidentialism is true.

[4] The appeal to contextualism makes my view a version of what Greco (2017) calls 'Contextualist Foundationalism'; Greco defends the general viability of this approach, and argues that it can also help with issues related to defeat.

1 Against Negative Access

Externalists reject the negative access principle, worrying that it implies that nothing, or at any rate not nearly enough, counts as evidence. While I think it compelling, I won't defend it in detail here. I will, however, recapitulate it briefly, to ensure that the later discussion remains sensitive to its spirit as well as its letter.

The worry can be put as follows.[5] All, or almost all, our information-gathering mechanisms are fallible: they sometimes malfunction and provide us with a falsehood, without giving us any warning that this has happened. In such a case, the deliverance of the mechanism, being false, cannot be part of our evidence. On pain of ruling out too many of our information-gathering mechanisms, this had better not imply that the deliverances of those mechanisms aren't part of our evidence when things are going well (even if we didn't know beforehand that they would). But then we get failures of negative access. For consider a case where someone is provided with P by a mechanism that is, on this occasion, malfunctioning. Since P is false, it isn't part of her evidence. But, since she has no reason to think that anything is awry, she has no reason to think that P is not part of her evidence. So, while P isn't part of her evidence, she has no evidence that this is so: she is a counterexample to negative access.

A paradigm instance of this problem is perception. Sometimes, my visual system tells me that a wall is red even though it's not, e.g. because it's a white wall misleadingly lit to look red. Let's call this the *bad case*, in contrast with the *good case* in which I see that the wall is red. In the bad case, it isn't part of my evidence that the wall is red. But, the externalist maintains, we shouldn't conclude that this claim is never part of my evidence, or is part of my evidence only if I knew beforehand that conditions are good. In particular, we shouldn't conclude that this claim is not part of my evidence in the good case. Instead, we should accept that, in the bad case, we have a violation of negative access: THE WALL IS RED isn't part of my evidence; but, since it is part of my evidence in the good case, and I have no reason to think I'm not in the good case, I have no

[5] Something like this argument is implicit in McDowell (1982, 1995, 2011) and Williamson (2000, ch.8). My presentation follows Weatherson (2011, p.451).

reason to think that THE WALL IS RED isn't part of my evidence.

While one can, of course, be an externalist and reject this particular example, I will use it as my paradigm case of a negative access failure in what follows.

2 From Externalism to Akrasia

We should reject the negative access principle. However, as indicated in the introduction, there is a powerful argument that doing so commits us to the rationality of intuitively problematic attitudes that are naturally described as 'epistemically akratic'. This section will lay out the problem; the rest of the paper will attempt to solve it.

The most straightforward case of an (epistemically) akratic agent is someone who believes a statement of the form 'P, but I shouldn't believe P' or 'P, but my evidence doesn't support P.' More generally, an agent is akratic if her attitudes conflict with her thoughts about what attitudes she should have.

What does this come to when we move to a graded framework which recognizes multiple levels of confidence or credences?[6] It's clearly akratic to have one credence, while believing a different one to be rational: 'I'm .3 confident in P, though my evidence supports P to degree .5' definitely sounds off. But there may be no particular credence which the agent thinks is rational; what is it to be akratic then? A natural proposal is that akrasia arises when one's credence diverges from one's *best estimate* of what the evidence supports: 'I'm .3 confident in P, though I estimate that my evidence supports P to degree .5' still seems like an odd thing to say. Probabilistic frameworks nicely capture estimates of a quantity as the expected value of that quantity: the weighted average of the possible values, weighted by how likely the quantity is to have that value. In our case, this means that an agent is akratic if her credence differs from the expected evidential support, as calculated using her credences.[7]

[6] I take a stance on this primarily to make the discussion more concrete. For I show in Appendix A that, once we reject negative access for the reason discussed in §1, it is hard to see what would prevent the possibility of bodies of evidence that support credence distributions that are clearly akratic on any plausible account.

[7] See Christensen (2010) and Horowitz and Sliwa (2015) for further discussion.

Epistemic akrasia looks like a paradigm instance of irrationality.[8] But, oddly enough, agents who violate negative access have evidence that supports an akratic state. For someone who violates negative access has evidence that differs from what she has reason to think it is. So her evidence supports one thing (supports P to one degree) and what she has reason to think her evidence is supports another (supports P to a different degree). Since the evidential support for her first-order beliefs depends on her evidence, and the evidential support for propositions about what's rational depends on what her evidence *says* her evidence is, the two will come apart. So her evidence supports an akratic state.[9]

This problem arises quite clearly in the case of the red wall.[10] Suppose that your background evidence establishes conclusively that you are either faced with a red wall or else with a white wall with red light shining on it. As it turns out, you are in the bad case, facing the white wall. So your evidence is just that the wall appears red, which (let us assume) supports the claim that it is red to degree .9. But you also know the relevant epistemological facts: in particular, you know that if the wall is red, the fact that it's red is part of your evidence and so your evidence supports the claim that it's red to degree 1. So your evidence supports to degree .1 that your evidence supports the claim that the wall is red to degree .9, and supports to degree .9 that it supports the claim to degree 1. So your evidence supports estimating the evidential support for the wall being red

[8] Cf Adler (2002), Feldman (2005), Kolodny (2005), Gibbons (2006), Christensen (2007, 2010), Smithies (2012), Elga (2007, 2013), Greco (2014a), Horowitz (2014), Titelbaum (2015), and Worsnip (forthcoming). For dissent, see Williamson (2011), Coates (2012), Lasonen-Aarnio (2014, 2015), and Weatherson (ms).

[9] This intuitive argument doesn't *quite* show that all agents who are counterexamples to negative access will have akrasia-supporting evidence. For, as formulated, it suffers from presupposition failure: there may be no body which our agent has reason to think is her total evidence. This matters because the uncertainty in what our agent's evidence is could 'cancel out' so that evidential support and expected evidential support coincide. For example, our agent's evidence could be E_1, which supports p to degree $\frac{1}{2}$; and E_1 could assign $\frac{1}{3}$ probability each to her total evidence being E_1, E_2, and E_3, where E_2 conclusively refutes p and E_3 conclusively establishes p. Then E_1 would violate the access principle, but it would not licence akrasia. Nonetheless, the intuitive argument makes clear why it would be an incredible coincidence if the negative access principle were false and yet no agent's evidence ever supported an akratic state; this makes (P2) highly plausible, even if not undeniable. Moreover, Williamson (2011), Samet (forthcoming), and Dorst (ms) prove that, in fact, such a coincidence can't obtain systematically, so that if any agents violate negative access, then some agents have akrasia-supporting evidence.

[10] Cf White (2014, p.306-8).

at $.1 \times .9 + .9 \times 1 = .99$, while also being only .9 confident of the wall being red.[11] It thus looks as though, if you follow your evidence, you will be akratic.

We now have all the pieces for a powerful argument from externalism to the rationality of akrasia:

(P1) The negative access principle is false.

(P2) If the negative access principle is false, someone could have akratic evidence (that is: evidence supporting an akratic state).

(P3) If someone could have akratic evidence, someone could rationally be akratic.

(C) Someone could rationally be akratic.

We have already motivated (P1) and (P2). (P3) *looks* like a modest consequence of evidentialism, stating that conforming one's beliefs to one's evidence is sufficient for being rational. And the argument is clearly valid. We thus seem forced into the unattractive conclusion that having akratic beliefs can be rational.

Before responding to the argument, I should explain how it connects to the literature. The general tension brought out by the argument, between externalism and anti-akrasia, is quite familiar. The focus, however, is unusual. Most of the literature ignores examples like that of the wall, which are motivated by the argument against negative access from §1; instead it focuses on (alleged) counterexamples to *positive* access (the principle that if P is part of one's evidence, one has conclusive evidence that P is part of one's evidence), often supported by considerations about margins for error, or on (alleged) examples of agents with misleading evidence about what supports what.[12] This is understandable,

[11] A divergence of .09 may strike some as too small to worry about. However, I doubt that we can rule out arbitrarily large divergences in a principled manner once we deny the negative access principle. See Appendix A for discussion.
Note that nothing depends on the evidence, when the wall is white, supporting its being red to degree .9; it's sufficient that it neither rules out that the wall is red nor supports it conclusively. For suppose the evidence supports it to degree x. Then, when the wall is white, the evidence estimates the support at $(1-x) \times x + x \times 1 = 2x - x^2$, which equals x only if $x = 0$ or $x = 1$.

[12] For example, Christensen (2010), Williamson (2011), and Elga (2013) all focus on a case motivated by margins for error, while Christensen (2007), Elga (2007), Coates (2012), and Weatherson

since those examples are more immediately intuitive than the theory-laden case of the wall. Nonetheless, it is a mistake. For these examples are supported by arguments far less compelling than that of §1. In particular, there may well be viable externalist positions which embrace positive access or the view that rational agents cannot be misled about the evidential support relation;[13] but there is no way externalists can subscribe to negative access. As externalists, then, the version of the problem discussed here should be our primary concern.

My defence of the focus on negative access has an ulterior motive. For, while the gap I diagnose in the argument above is present in similar arguments based on failures of positive access or misleading evidence about evidential support, I see no obvious way of extending my strategy for systematically exploiting that gap to these other cases. But since mine is the hardest case, and since we might, even as externalists, have other resources for handling the other cases, this does not eliminate the interest of my approach.

3 The Way Out

The argument from externalism to akrasia is powerful; but, fortunately, (P3) is not as innocent as it appears. This premise lets us move from the observation that some agents have evidence which supports an akratic state to the claim that some agents are rationally akratic. But this transition is non-trivial: having evidence which supports a state does not always mean that one can rationally be in that state, *even if* we accept the evidentialist thesis that rationality is a matter of conforming ones beliefs to the evidence. I will illustrate this using a more familiar case, before explaining how it might apply to akratic states.

Consider the Moorean conjunction Q AND I DON'T BELIEVE THAT Q. It is often noted that a claim of this form can be true. More important for us is that it

(ms) focus on cases where agents have misleading evidence about the evidential support relation. Horowitz (2014) and Worsnip (forthcoming) discuss both types of cases, but don't discuss examples motivated along the lines of §1.

[13] McHugh (2010), Greco (2014b), Stalnaker (2015), and Das and Salow (forthcoming) suggest ways to reconcile externalism with positive access. Greco (2014a), Titelbaum (2015), and Smithies (2015) develop strategies for denying the possibility of misleading evidence about relations of evidential support, which look independent of the internalism/externalism debate.

can also be supported by one's evidence. In fact, anyone who (like most of us) has strong evidence for some Q, while obviously failing to believe it, has strong evidence for a Moorean conjunction. Clearly it would be a mistake to conclude, just from this, that someone could rationally believe such a conjunction.

In the Moorean case, it's easy to see why this doesn't follow, even given evidentialism. For it's plausible that (i) one has evidence supporting a Moorean conjunction only when one doesn't currently conform to one's evidence. After all, it's plausible that agents can tell what they believe; it follows that one has strong evidence for Q AND I DON'T BELIEVE THAT Q only when one doesn't believe Q. Moreover, one has strong evidence for Q AND I DON'T BELIEVE THAT Q only when one has strong evidence for Q. Putting the two together, we get that one has strong evidence for Q AND I DON'T BELIEVE THAT Q only when one doesn't believe Q despite having strong evidence for it. That is, one has strong evidence for the Moorean conjunction only when one fails to conform some relevant beliefs to one's evidence.[14]

It follows from (i) that (ii) in conforming to some evidence that supports a Moorean conjunction, one changes matters so that one's evidence no longer supports that conjunction. But it's worth looking a bit more carefully at why conforming to the Moore-paradoxical evidence is self-undermining in this way. To conform to evidence supporting Q AND I DON'T BELIEVE THAT Q, one has to, among other things, come to believe Q. But coming to believe Q changes what evidence one has; and, for the Moorean proposition, some of these changes matter. For one's new evidence no longer supports the claim that one doesn't believe Q, and thus no longer supports the conjunction Q AND I DON'T BELIEVE THAT Q. While one starts off with evidence for the Moorean conjunction, attempting to conform to it changes what evidence one has in such a way that the new evidence no longer supports that conjunction.

The Moorean case thus reveals a gap between (a) having evidence that supports a state and (b) being able to be in that state while conforming to one's evidence; it thus shows that (P3) doesn't immediately follow from evidentialism.

[14]This treatment of the Moorean case, especially in its assumption that agents can always tell what they believe, is controversial. However, I'm using this case primarily to illustrate the non-triviality of (P3); it can serve that purpose even if my treatment of it is, ultimately, mistaken.

To *reject* (P3), however, we need more: we need to argue that akratic evidence systematically falls into that gap, just like Moorean evidence does. We thus have to motivate analogues of the above claims about Moorean evidence: (i') one has akratic evidence only because one doesn't currently conform to one's evidence; and therefore, (ii') in conforming to one's current evidence, one would change things so that one's new evidence no longer supports an akratic state. Since the akratic evidence under discussion here is evidence violating the negative access principle, it would be sufficient to motivate:

> **Instability**
> An agent violates the negative access principle only if (and while) her beliefs don't perfectly conform to her evidence.

By endorsing Instability, we can reject (P3) without rejecting evidentialism. On the resulting view, agents who perfectly conform to their evidence will never have evidence that violates negative access, and hence won't be akratic – or, at least, won't be akratic because of issues arising from negative access.[15]

4 Contextualizing the Red Wall

I have explained the tension between externalism and the claim that akrasia is irrational, and described a strategy for resolving it. The question now is whether this strategy can be successfully implemented, by explaining why akratic evidence would be unstable in the required way.

Moorean conjunctions illustrate one source of instability. Could the instability in akratic conjunctions be exactly analogous? It seems not. Moorean evidence is unstable because whether I have evidence for the second conjunct – that I

[15] Smithies (2012, p.288-292) discusses the related gap between having justification for a proposition and having a justified belief in that proposition; he even includes the analogy with Moorean propositions (crediting it to Ralph Wedgwood). However, the only proposal he discusses for exploiting this gap is Bergman's (2005) view that believing that one shouldn't believe P defeats one's justification for believing P. Since Bergman gives no explanation for why believing that one shouldn't believe P would change one's evidence about P, that defeat is naturally understood to only affect one's *doxastic* justification for believing P; this leaves the view open to Smithies' (2012, p.288-292) objection that it is insufficiently explanatory. By following the Moorean analogy more fully, and defending Instability, we can improve on Bergman's view in exactly this respect.

believe Q – depends on whether I believe the first conjunct – namely Q. Things seem different for akratic conjunctions: whether I have evidence for the second conjunct – that I shouldn't believe P – does not seem to depend on whether I believe the first conjunct P (or *vice versa*). If akratic evidence is unstable, this won't be because what I believe affects my evidence. But it could be for a similar reason identified by contextualists: what I believe might affect my epistemic standards, and hence what the word 'evidence' applies to when I use it.

How would this help? Let us say that an agent *takes a possibility seriously* if that possibility is compatible with her (all-out) beliefs. And suppose that whether I take seriously the possibility of funny lighting affects what 'evidence' refers to when I utter it as follows. When I don't take the error possibility seriously, I use 'evidence' to refer to evidence$_{lo}$. And evidence$_{lo}$ works exactly as we have been assuming, in that THE WALL IS RED can become part of one's evidence$_{lo}$ when one looks at the wall, provided only that looking is in fact reliable in one's circumstances. By contrast, when I do take the error possibility seriously, my usage picks out the more stringent evidence$_{hi}$; and claims like THE WALL IS RED can become part of one's evidence$_{hi}$ only if one has already checked that the lighting is normal. This, I claim, allows us to implement the Instability-based solution in the example of the red wall.[16]

(I have, and will continue to, put the point in contextualist terms. But one can say exactly parallel things if one adopts an analogous relativism or expressivism about 'evidence'.[17] I believe that invariantist views, even if 'subject-sensitive', cannot implement an equally attractive version of the strategy;[18] but I won't

[16]See e.g. Hawthorne (2004, p.73-77) and Neta (2005). Obviously it's implausible that there is a single relation, evidence$_{lo}$ (evidence$_{hi}$), which we always pick out when we don't (do) take seriously that the lighting is misleading. On any plausible version of contextualism, there will be myriad relations that the phrase 'p is part of X's evidence' can pick out, and our attitude towards this particular error possibility is only one constraint in settling which one we do pick out. But we can categorize these relations according to whether one can come to bear them to the proposition THE WALL IS RED just by looking at the wall, without having checked the lighting. The claim I actually require is then that, at least sometimes, when we don't take the possibility of misleading lighting seriously, we use 'evidence' to pick out a relation that has this property; and that, when we do take the possibility seriously, we always use 'evidence' to pick out a relation that lacks this property. And *that* claim is quite plausible. But, for presentational simplicity, I will continue with the simplifying assumption made in the main text.

[17]For epistemic relativism, see MacFarlane (2005); for expressivism, see Chrisman (2007).

[18]For subject sensitive invariantism as an alternative to contextualism, see e.g. Fantl and

argue that here.)

On the contextualist account, the case of the red wall divides into two, depending on whether the subject takes the error possibility seriously. Suppose first that she does. Then, as she uses it, 'my evidence' will pick out her evidence$_{hi}$. But checking the lighting is required before THE WALL IS RED can be part of her evidence$_{hi}$, regardless of how cooperative the circumstances are. And our agent can tell that she has performed no such checks. So she can tell that THE WALL IS RED isn't part of what she refers to with 'my evidence', regardless of what the actual circumstances are. She thus won't be worried that she should be more confident that the wall is red; and so she won't be akratic.

Suppose instead that she doesn't take the possibility of misleading lighting seriously and all-out believes that the wall is red. Then, as she uses it, 'my evidence' will refer to her evidence$_{lo}$, and whether THE WALL IS RED is part of her evidence$_{lo}$ *does* depend on what the circumstances are actually like. But she can't be worried about this dependence. By hypothesis, she all-out believes that the lighting is good; so, provided she draws the obvious consequence, she also believes that THE WALL IS RED is part of her evidence$_{lo}$, and hence that her all-out belief that it's red is fully justified.[19]

Importantly, treating the case this way does not eliminate the counterexample to negative access. If our subject fails to take the possibility of misleading lighting seriously when it is actual, she is a counterexample to negative access for evidence$_{lo}$ – the thing she refers to with 'evidence'. Since the wall isn't red, her evidence$_{lo}$ won't include THE WALL IS RED. But her evidence$_{lo}$ won't tell her that it doesn't include that claim, since it leaves open the possibility that the wall might be red and the lighting normal, in which case THE WALL IS RED would have been part of her evidence$_{lo}$. So she is a counterexample to negative access, and the information she refers to with 'my evidence' supports an akratic state.

McGrath (2002), Hawthorne (2004), and Stanley (2005).

[19]What if she does not draw the consequence? Then she could indeed be akratic, all-out believing that the wall is red while estimating that she should only be .99. But she will not be rational: either her first-order belief (if she is, in fact, in the bad case) or her higher-order belief (if she is, in fact, in the good case) will fail to conform to what she refers to with 'evidence'. (She will also be *structurally* irrational, incoherently taking the wall to be red when wondering what colour it is but not when wondering what her evidence says about its colour. As discussed in footnote 30, this is no accident.)

It's just that she doesn't suspect that this is going on, and hence won't herself be akratic. Moreover, if she were to suspect that this is going on, her usage of 'evidence' would change to pick out evidence$_{hi}$ instead; and the case of the red wall is no counterexample to negative access for evidence$_{hi}$.[20]

This treatment of the red wall implements the Instability strategy. The information our subject refers to with 'my evidence' violates negative access, and hence supports an akratic state, only when she fails at perfectly conforming her beliefs to it. For we get no negative access violation (for what the subject picks out with 'evidence') when the subject is merely .9 confident that it is red: in that case, she uses 'evidence' to pick out evidence$_{hi}$ and she can tell that THE WALL IS RED is not part of her evidence$_{hi}$. And we also get no violation when the subject believes the wall to be red and it really is: in that case, she uses 'evidence' to pick out evidence$_{lo}$, and THE WALL IS RED is part of her evidence$_{lo}$. It is only if the subject believes the wall to be red when it isn't that we get a violation. But in that case, her doxastic state doesn't perfectly conform to what she refers to as 'my evidence', since that calls for high confidence, rather than all-out belief, in the claim that the wall is red.

Moreover, this fact explains why our subject's attempts to conform to her

[20] But isn't it still true – as a referee suggested – that our agent can figure out that THAT WALL IS RED is not part of her evidence simply through reflection? After all, the process of taking new possibilities seriously looks rather like reflection, and the result of that process is that the agent knows that she lacks this piece of evidence. And if that's right, it looks like our account vindicates the core commitment of internalism.
However, I think this notion of 'being able to know by reflection what evidence one has' is too deflated to provide comfort. For note that reflection, even if extended to include the process of taking seriously new possibilities, cannot tell our subject whether THAT WALL IS RED is part of her evidence$_{lo}$. It thus isn't a way of answering the question she originally asked herself. At best, it's a way for her to shift her interest to the question of whether THAT WALL IS RED is part of her evidence$_{hi}$ – a question she finds much easier to answer. But that's cheating: we can answer *any* question by reflection if we're allowed to replace it with an easier one.
Suppose, however, that we waive this first point, granting that there is a good sense in which the questions are the same. Then a different worry arises, namely that this reflection succeeds only because it can 'change' what the answer is, making our subject 'throw out' evidence by raising her standards so that previously eligible propositions no longer qualify. This means that reflection puts us 'in a position to know' the answer to our question only in a highly attenuated sense. Suppose you buy an App which is advertised as 'putting you in a position to know' how much money you have in your account. But when you open the App, it turns out to consist entirely of a single button which will, when pressed, donate the entire contents of your account (if any) to charity. You would feel cheated. Since the reflective process which 'puts you in a position to know' what your evidence is is no better, we should reject the internalist marketing.

evidence, in the case where it does support an akratic conjunction, would be self-undermining, much like in our Moorean example. Suppose that our subject starts off in the worst case: she's faced with a white wall, but doesn't take that possibility seriously. Then the information she refers to with 'my evidence' (i.e. her evidence$_{lo}$) supports the claim she would express with the akratic conjunction 'the wall is .9 likely to be red, but I estimate that my evidence supports it being red to degree .99.' But it does so only because the information she refers to with 'my evidence' does not in fact rule out that the wall is white. So in order to conform to the body she calls 'my evidence', she would have to start taking seriously the possibility that the wall is white.

Let's suppose that, a little later, she does just that. She can then come to believe the claim that she would previously have expressed with the akratic conjunction, namely that the wall is .9 likely to be red, but that she estimates that her evidence$_{lo}$ supports it being red to degree .99. But this claim no longer has the form of an akratic conjunction for her, because when she now uses 'evidence', she refers not to evidence$_{lo}$ but to evidence$_{hi}$. That is, the norm she now feels bound by is one that tells her to conform her beliefs to her evidence$_{hi}$. So believing or asserting that the wall is .9 likely to be red, but that she estimates that her evidence$_{lo}$ supports it being red to degree .99, is no more akratic than doing ϕ whilst believing or asserting that, according to rule R *which one doesn't endorse*, one shouldn't ϕ. To be akratic now, our agent would have to believe that the wall is .9 likely to be red, while estimating that her evidence$_{hi}$ supports it being red to degree .99. But that isn't something she's inclined to do, since she knows full well that her evidence$_{hi}$ supports the wall being red to degree .9.

This explanation may make it seem as if this solution is a case of linguistic trickery: we only predict that agents will never sincerely utter akratic conjunctions. But that isn't right. The important point is that what propositions we aim to conform our beliefs to (call them 'the propositions that are special by our standards') depends on what possibilities we take seriously. Genuine akrasia would be a state in which we have one attitude, whilst also thinking that the propositions that are special by our standards support a different conclusion. Our account explains why a rational agent will never be in such a state, even though her evidence may support a proposition R, such that believing R whilst

having her current standards would be such a state. For in following her evidence, our agent would change her standards and hence which propositions she treats as special; so if she then comes to believe R, doing so doesn't put her into a state where her attitudes and her views about what the propositions that are special by her standards support come apart.

These points also help with a slightly different worry. Suppose that *we* are in a context where we use 'evidence' to mean evidence$_{lo}$. And suppose that we are discussing someone who is faced with the white wall, and takes the error possibility seriously. Then couldn't *we* describe her by saying 'she is .9 confident of P, but estimates that her evidence supports P to degree .99'? After all, when we use it, 'evidence' means evidence$_{lo}$, and she does estimate that her evidence$_{lo}$ supports P to degree .99.[21] Moreover, since the subject is conforming her belief to her evidence$_{lo}$ – the thing we describe as 'her evidence' – we can truly say that the subject is rational. But then it seems like we can truly attribute an akratic state to a subject we truly describe as rational.

The previous discussion shows that this worry is misguided: the state which we are ascribing to the subject is not an akratic state. There is no tension in thinking something, while also thinking that by some standards one rejects, one shouldn't think that – and, as we noted above, that is all that is involved in our subject's state. That there is no tension here is occluded by the words we've used to describe the subject. But we should not be taken in by the form of words. In the right sort of context, for example when 'should' expresses legal obligations, I might describe someone as rationally ϕ-ing whilst knowing full well that she shouldn't ϕ. Despite what a superficial look at the form of words might suggest, I have not thereby given an example of rational practical akrasia. The same, I submit, is true when – in the scenario above – we describe our subject as

[21] Contextualists could resist this. For even if the reference of 'evidence' is *normally* determined by what the speakers are taking seriously, this may not be true when the word occurs embedded in a propositional attitude ascription, as it does here. Such behaviour would not be unusual: perhaps 'might' normally means for-all-the-speakers-know, but 'Sarah believes that it might be raining' almost always means that Sarah believes that it's raining for-all-*she*-knows. Modelling 'evidence' on the behaviour of 'might' would suggest that 'she is .9 confident of P but estimates that her evidence supports P to degree .99' is true, even in our mouth, only if she estimates that her evidence$_{hi}$ supports P to degree .99 – which she does not. However, since there is a deeper reason why this worry is misguided, I will grant this step for the sake of argument.

believing one thing while estimating that her evidence supports another.

It's worth noting that, despite the structural analogies, the way in which attempting to conform to the problematic evidence is self-undermining is different in the akratic case and in the Moorean one. In the Moorean case, conforming one's beliefs to one's evidence changes what evidence one has; and the new evidence no longer supports the problematic proposition. In the akratic case, conforming one's beliefs to one's evidence changes what one means by 'evidence'. Interestingly enough, however, this leaves unaffected what evidence one could correctly self-ascribe, since, in the bad case in which one has the potentially akratic evidence, one's evidence$_{lo}$ and one's evidence$_{hi}$ are the same: both contain weak propositions such as THE WALL APPEARS RED, and neither includes a strong claim such as THE WALL IS RED. Instead, the change in what one means by 'evidence' changes which propositions count as akratic propositions to believe. To this extent, the two resolutions work in slightly different ways.

We have now seen how certain contextualist judgements allow us to give an Instability-based treatment of the red wall, a treatment that reconciles the assumption that this is (on some uses of 'evidence') a counterexample to negative access with the claim that it never rationalizes akrasia. The next, and final, step is showing how this treatment of the particular example can be generalized and integrated into a systematic contextualist theory.

5 A General Theory[22]

Can a plausible contextualist account predict Instability with complete generality? I will argue that it can, by sketching a schematic and oversimplified theory which does. But I do not claim that a theory like the one I sketch is the best or only version of contextualism that can do this – if other versions have the same feature, so much the better.

To simplify things, I will presuppose Williamson's (2000, ch.9) controversial view that one's evidence consists of all and only the things one knows (E=K, for short). This will make the task easier, since contextualism is more familiar and

[22]This section builds on Salow (2016), which focuses on the status of negative access specifically in Lewis's (1996) theory of knowledge.

better developed for 'knowledge' than for 'evidence'.[23] It is also a convenient way to make sure that the resulting theory fully respects the externalist objection to negative access, since no plausible account of 'knowledge' could vindicate the principle that whenever one doesn't know P, one knows that one doesn't know P.[24] However, the details of E=K won't matter; we could do equally well, for example, with a theory on which one's evidence also includes claims one is merely in a position to know (cf Williamson (2000, ch.9)) or is restricted to claims known non-inferentially.

To build up our contextualist theory, start by considering a schematic reliabilist account of knowledge:

> *Reliabilism*
> X's belief in P at w is knowledge iff X doesn't falsely believe P in possibilities that bear R to w.

This yields a more concrete view when we interpret R: for example as *being close to*, *being a relevant alternative to*, or *being at least as normal as*.[25]

This schematic account is naturally adapted into one that relativizes knowledge attributions to a set of possibilities S that play a similar role to the actual world in fixing which possibilities matter, thus affecting how broad a range of possibilities a belief has to be reliable in:[26]

[23] Though see Neta (2003) for contextualism directly about 'evidence', and for arguments that this is preferrable to contextualism about 'knowledge'.

[24] Some have claimed that Lewis's (1996) contextualism entails this principle; Holliday (2015) and Salow (2016) show that this is a mistake.

[25] 'Is close to' makes this a safety-theoretic account of the kind inspired by Sosa (1999) and Williamson (2000); 'is a relevant alternative to' makes it a relevant alternatives theory of the kind advocated by Dretske (1970) and Goldman (1976); 'is at least as normal as' makes it a normal conditions account of the kind later favoured by Dretske (1981) as well as by Stalnaker (2006). Our intuitive grasp on these notions may not guarantee that they are different; but, when they are theoretically embedded, differences emerge. (One important difference is that only some of these relations are plausibly transitive. When R is transitive, the account validates the KK principle; and, as discussed in §2, that principle may be required for a full vindication of anti-akrasia.) A fourth option is to endorse this account in a non-reductivist spirit, refusing to say anything non-circular about R; cf Williamson (2009).

[26] The contextualist account developed by Lewis (1996), and endorsed in adapted form by Blome-Tillmann (2009, 2014) and Ichikawa (2011a,b), has something very close to this structure. While the account sketched by DeRose (1995) could fit a similar formal mold, his theory of how S is determined differs significantly from the account I will be assuming.

Relativized Reliabilism
X's belief in P at w is knowledge relative to S iff X doesn't falsely believe P in possibilities that bear R to either w or any world in S.

This amounts to a contextualist account if we assume that, when agents make unrelativized knowledge attributions, S is supplied by the context of utterance. To have a term for S, let us call it the set of possibilities *salient* to the speaker (keeping in mind that this is a technical notion).[27]

To derive concrete predictions from the account, we need to say something about what makes a possibility salient. A plausible starting point is:

Salience Constraint
If X takes w seriously, then w is salient for X.

This condition follows from Blome-Tillman's (2009; 2014) account of salience as compatibility with one's presuppositions if we assume that, as seems plausible at least in single-individual contexts, if one presupposes something, one believes it. It correctly predicts that, when, for practical reasons or because of a felt need for greater intellectual responsibility, we take more error possibilities seriously, our standards for knowledge rise. And it also suggests a nice story about why they would do so. It's natural to think that we attribute knowledge that P to those we are willing to rely on to speak and act correctly when things depend on P; but we would not want to rely on someone concerning P if we took seriously possibilities in which their beliefs about P aren't hooked up with the facts.

While schematic, this account is constrained enough to support the judgements used in §4. Every world bears R to itself; so if X believes the wall to be red when it isn't, she cannot be described as knowing that it's red regardless of which possibilities are taken seriously. Similarly, if X believes that the wall is red based on looks, and the lighting is unreliable, she won't know in any sense of 'knows'; for there will be worlds bearing R to the actual world (ones in which the lighting is the same, but the wall is white) in which she believes this falsely. By contrast, if the wall is red, and is red in all worlds bearing R to the actual

[27] There are interesting questions about how S is determined if different possibilities are salient to different participants in the conversation; see DeRose (2004) for discussion. I side-step these complications by focusing on 'conversations' with only one participant.

world, it starts to matter which worlds the speakers take seriously. If they take seriously possibilities in which the wall is white but appears red because of odd lighting, they could not correctly describe a belief that the wall is red as 'knowledge' if it was formed just by looking. If, on the other hand, they don't take such possibilities seriously, they could.

Our schematic theory thus vindicates the judgements used in §4 to give an akrasia-free treatment of the case of the red wall. But, what is more, our account generalizes our treatment of that case to every counterexample to negative access, thus providing a completely general escape from the argument of §2. For our account entails a 'blindspot' thesis stating that agents can never take seriously the possibility that they are violating negative access:[28]

> **Blindspot Thesis**
> Suppose that X is *fully coherent*: her beliefs are consistent and closed, and she knows under which circumstances a proposition P and a person Y are related in the way that would make 'P is part of Y's evidence' true as X uses those words. Then X all-out believes that she is not currently a counterexample to negative access.

And the Blindspot Thesis implies the Instability principle from §3. For, by the Blindspot Thesis, any agent whose beliefs perfectly reflect her evidence will all-out believe that she is not a counterexample to negative access. But an all-out belief in P perfectly reflects one's evidence only if one has conclusive evidence for P. Since everything for which one has conclusive evidence is true, it follows that our agent is not a counterexample to negative access. So Instability is true.

Before looking at why exactly our account implies the blindspot thesis, it's worth noting that this thesis is suprising. After all, some people are counterexamples to negative access. Given that some people violate negative access, what's to stop us from suspecting that we might be among them? But the account provides an answer. Suppose someone begins to suspect that she is currently a

[28]The 'blindspot' terminology is from Sorensen (1988). My usage is slightly idiosyncratic since I classify something as 'falling in a blindspot' only if the agent has to believe it even when it's false; it's more common to classify something as 'falling in a blindspot' as long as the agent can't believe that it is false even when it is, which is consistent with her not taking a stance either way.

counterexample to the negative access principle. That is, she begins to take seriously certain possibilities in which a particular information-delivery mechanism of hers is unreliable. Then she will use 'evidence' to refer to a more demanding notion, so that the outputs of this particular information-delivery mechanism never fall in the extension of 'evidence' as she is now starting to use that word. But then she can tell that this mechanism isn't delivering anything falling under her use of 'evidence', whether or not it is unreliable. So the possibility in which this mechanism is unreliable will no longer be one in which she lacks something she would call 'evidence' but can't tell that she does.

This is exactly what we saw happening in the case of the red wall. Ordinarily, agents do not take seriously the possibility that the lighting is currently misleading. They thus do not take seriously the possibilities in which they would be counterexamples to the negative access principle. But suppose someone starts entertaining doubts. Then she will raise her standards for what falls under 'evidence' so that simply looking at the red wall is not enough to have THAT WALL IS RED as part of one's evidence – before that can become part of one's evidence, one first has to check that appearances are not misleading. But now the possibility in which the lighting is misleading is no longer a counterexample to negative access for the notion the agent picks out with 'evidence'. For, whether or not the lighting is misleading, the agent can tell that she hasn't checked that it is. And so she can tell that THAT WALL IS RED fails to be part of the information she picks out with 'my evidence' regardless of how good the lighting is.

This informal explanation can be turned into a more rigorous argument. Let S be the set of possibilities salient in X's context, and let w be a world which X takes seriously and in which X doesn't know$_S$ P; we will show that, at w, X knows that she doesn't know$_S$ P. We can assume that, at w, X believes P (otherwise she would surely know that she doesn't know$_S$ it).[29] By *Relativized Reliabilism*, this means that there is some v in which X believes P though P is false, such that

[29] One might worry that externalists can't accept the principle that we can always tell when we fail to believe something. I think that the strategies for reconciling externalism with positive access mentioned earlier can also reconcile it with this introspection principle. But, at any rate, whether X believes P is only relevant because we are working with E=K instead of the very similar view that one's evidence consists of what one is in a position to know; so the theory could easily be modified to avoid relying on this introspection principle.

either (i) vRw or (ii) for some $x \in S$, vRx. But if (i), then (ii), since, by the *Salience Constraint*, $w \in S$. So there is some v in which X falsely believes P which bears R to some world in S. But then it trivially follows from the account of know$_S$ that X doesn't know$_S$ P *at any world*. So if X is at all reflective at w, she will know at w that she doesn't know$_S$ P. Since w was an arbitrary world consistent with X's all-out beliefs in which X doesn't know$_S$ P, it follows that X all-out believes that if she doesn't know$_S$ P, she knows that she doesn't know$_S$ it. In other words, X believes that she is not a counterexample to negative access.

The argument just given is somewhat abstract; but it's just the obvious generalization of what happens in the case of the red wall. Suppose there is a possibility in which, according to how X actually uses 'know', X doesn't know the wall's colour and also doesn't know of her ignorance. Why doesn't she know the wall's colour? We can set aside the possibility that she doesn't have a belief about the wall's colour; if that were the case, she'd know that she doesn't know. So it must be that her belief is insufficiently reliable: the wall is actually a different colour, or the lighting is misleading, or she is hallucinating. But, crucially, the source of the unreliability must not itself be salient if X is to be unaware of her ignorance. Suppose, for example, that the belief is unreliable because the lighting is misleading; if that possibility were salient, the standards for falling in the extension of 'knowledge' would be high enough that anyone whose belief isn't reliable when the light is misleading is ignorant, and X can tell that she doesn't clear *that* bar. So the possibility in which X is a counterexample to negative access must be one in which her belief is unreliable in some way that isn't currently salient. But, by our *Salience Constraint*, any way of X being unreliable that is consistent with X's actual beliefs will be salient in her context. So the possibility in which X is a counterexample to negative access must be inconsistent with X's actual beliefs.

We thus have the outline of a theory, consisting of *Relativized Reliabilism* and the *Salience Constraint*, which entails the Blindspot Thesis and Instability, and can thus implement the strategy of §§3-4 with full generality.[30] Unfortunately,

[30] In fact, the Blindspot Thesis has implications for epistemic akrasia that go beyond its ability to implement the Instability strategy. Instability guarantees only that you won't be akratic (due to issues arising from negative access) when you perfectly conform your beliefs to the evidence. As I show in Appendix B, the Blindspot Thesis allows us to go further and show also that you won't be

I cannot quite leave things there. Before resting my case, I will address two objections, the first of which requires me to refine the theory just described.

5.1 Generalized Blindspots

The first objection observes that *Relativized Reliabilism* and the *Salience Constraint* together entail not just the Blindspot Thesis, but also a more general principle. For note that, in our derivation of the Blindspot Thesis, it never played a role that subject and attributor of the knowledge ascriptions were the same. So the argument really established

> **Generalized Blindspot Thesis**
> Suppose X is fully coherent. Then X all-out believes that no one is (or has ever been, or will ever be) a counterexample to the negative access principle.

But that principle looks obviously false. Suppose that I've stolen my friend's lunch from the fridge, and that he has not yet gone to look for it. Then surely I can, without incoherence, say that he doesn't know that his lunch is in the fridge, but also doesn't know that he doesn't know this.

I think that's right: the Generalized Blindspot Thesis really is absurd.[31] And I think that it's a problem even for the use I'm trying to make of the theory, which doesn't require it to be exactly right – the problem with the Generalized Blindspot Thesis is simply too close to the desired application to be dismissed as a problem with the details.

The problem is that the *Salience Constraint* is too strong.[32] Consider again the case of the stolen lunch. Since I know that my friend's lunch has been stolen, I take the lunch theft possibility seriously. The *Salience Constraint* then entails

akratic provided only that your credences obey the probability axioms. This is a welcome result: it means that akratic states are classified not only as failing to conform to the evidence, but also as exhibiting a failure of *structural* rationality, as intuitively they do. Cf Worsnip (forthcoming).

[31] Though see Salow (2016, §2.1) for some defensive considerations.

[32] However the problem is not that we, loosely following Blome-Tillmann (2009), connected salience to belief, rather than to attention as Lewis (1996) does. For the analogous principle connecting salience to attention entails that we can never attend to any counterexamples to the negative access principle, which is just as implausible.

that I use 'know' to pick out a very demanding relation: one that, no matter his environment, my friend could have borne to the proposition that the lunch is in the fridge only if he had checked that it wasn't stolen.[33] But that's wrong. Nothing prevents me from using 'know' in an undemanding way which, for example, allows me to say that my friend would have known where his lunch was, if only I hadn't stolen it.

What, then, is the difference between the possibility that my friend's lunch has been stolen (which I can take seriously without raising my standards) and the possibility that the light in this room is misleading (which, intuitively, I cannot)? A natural thought is that the latter threatens not just to make *someone* unreliable, but to make *me* unreliable. Picking up on this suggestion, we can weaken the Salience Constraint to:

> *Weak Salience Constraint*
> If X takes w seriously, and w bears R to a possibility in which X herself has a false belief, then w is salient for X.

This weaker principle retains much that was attractive about the original one. In particular, it upholds the link between what standards I set for 'knowledge' and which of my own beliefs I'm willing to rely on. It is more parochial, since it allows me to recognize error possibilities faced by others whilst nonetheless adopting standards for my own beliefs which work well only if I myself am not in a similar predicament. But, given the pervasiveness of error, and the need to continue with life regardless, such parochialism may be exactly what's required.

Relativized Reliabilism and the *Weak Salience Constraint* do not entail the Generalized Blindspot Thesis. Consider, for example, the case of the stolen lunch. I take the possibility that my friend's lunch has been stolen very seriously. That possibility is one in which my friend is unreliable; but it is not a possibility in which I am unreliable. So the *Weak Salience Constraint* is consistent with saying that the stolen lunch possibility is not salient for me. So I can continue to pick out a relatively undemanding relation in my use of 'know'. And since my friend is obviously unaware that he fails to bear even this undemanding

[33] Note that my friend *can* tell that he doesn't bear *that* relation to the proposition that his lunch is in the fridge – so once we get to this point, the Generalized Blindspot Thesis really does follow.

relation to the proposition that his lunch is in the fridge, nothing prevents me from registering this.

However, *Relativized Reliabilism* and the *Weak Salience Constraint* still entail the ordinary Blindspot Thesis. For to take seriously that she herself might be a counterexample to negative access, X would have to take seriously a possibility in which she herself is unreliable. And since X herself is unreliable in that possibility, the *Weak Salience Constraint* would then kick in to force that possibility into salience. But when the possibility is salient, the standards are raised so that 'know' picks out a more demanding relation. And X is easily able to tell that she does not bear this more demanding relation to the proposition in question.

The availability of the *Weak Salience Constraint* thus shows that natural contextualist theories can entail the Blindspot Thesis without entailing its generalization. This is enough to make my point that a plausible version of contextualism can vindicate the Blindspot Thesis. That said, I want to remain neutral on whether the retreat to the *Weak Salience Constraint* is ultimately the right response for contextualists whose theories entail the Generalized Blindspot Thesis. The problem is a new one, and it may be too early to tell how best to modify otherwise attractive theories to avoid it. My hope, supported by the naturalness of the retreat to the *Weak Salience Constraint*, is that the right response will preserve the Blindspot Thesis. That this thesis can do the work I have carved out for it is, I believe, further abductive reason to think so.[34]

5.2 Blindspots and Prefaces

However, one might worry that even the weaker Blindspot Thesis is still implausible. The thesis says that I always have to believe that I satisfy negative access, even though negative access isn't always true. This just sounds incredible. Now, we saw a contextualist explanation for why the very real negative access failures are 'elusive' in just this way. But still. I know that virtually no one else satisfies the negative access principle, and remember many cases in the past where I violated it myself. Since I know that I-right-now am not relevantly different, why

[34] Salow (2016, §2.2) and, implicitly, Schultheis (ms) explore other responses. I suspect that these can also be used to vindicate the Blindspot Thesis; but I won't defend that suspicion here.

should I think that I now satisfy negative access?

This worry should remind us of the preface paradox. Since I believe each of my beliefs, I believe of each of my beliefs that it is true. Yet I know perfectly well that virtually no one else has only true beliefs, and remember many cases in the past when even I myself did not. Since I know that I-right-now am not relevantly different, why should I think that I now have only true beliefs?

This is not the place for an in-depth discussion of that paradox. But it's worth filling out the analogy, just to show that this problem with the Blindspot Thesis really is a problem for (nearly) everyone.

In the case of the preface paradox, it is natural to distinguish between (a) believing of each of one's beliefs that it is true and (b) believing that all of one's beliefs are true. We can draw a similar distinction in the case of the blindspot thesis, between (a') believing with respect to each proposition that one satisfies the instance of negative access that concerns that propositions and (b') believing that one satisfies every instance of the negative access principle. The contextualist account only directly motivates the weaker (a'). Moreover, to show that an agent won't be akratic with respect to a particular proposition P, we only need to assume that the agent believes that she satisfies negative access with respect to each proposition that she takes to be relevant to P. So something close to (a') is enough to show that agents won't be akratic with respect to any particular proposition. The distinction between believing each instance and believing the generalization thus seems to help in our case just as much as it does in the case of the preface paradox.

That distinction, of course, is difficult to maintain in the presence of principles like multi-premise closure. This matters for our purposes because the notion of 'belief' I have been using is supposed to be a notion of *all-out* belief, the kind of belief that perfectly conforms only to conclusive evidence, and that corresponds to maximal credence. And these notions (conclusive evidence, maximal credence) seem like they will have to obey multi-premise closure. It is thus unclear that the distinction between (a') and (b') can really be sustained in the kind of background theory I have been working with. But then the same thing will be true about the distinction between (a) and (b), and so the blindspot thesis will still present no new problem over and above the one presented by the

preface paradox.

We can make the same point in a more concessive spirit. The preface paradox is easily solved if one rejects the notion of all-out belief, or maintains that there is hardly anything that it is ever rational to all-out believe. A theorist of this kind might be puzzled by the blindspot thesis, and by the account I have been offering. But a theorist of this kind is likely not to see the point of the current project in the first place; after all, the primary motivation of externalism, as I have presented it, is to maintain that we can receive conclusive evidence from fallible mechanisms. Someone who maintains that we should never form all-out beliefs will presumably also think that conclusive evidence is not something that we have or need; they are thus likely to get off the boat from the very start. But anyone who, like my externalist, thinks that we often have conclusive evidence, should accept that we have many rational all-out beliefs; any such person should thus think that the preface paradox is a difficult problem, and that some of the less intuitive features of the blindspot thesis are simply an instance of this difficulty.

6 Conclusion

We began with a challenge. Externalists want to allow that we can have plenty of evidence, even if some of that is delivered by fallible mechanisms. To allow that, externalists reject the negative access principle. But, when negative access fails, agents have evidence that supports an akratic conclusion. Externalists thus seem forced to choose between allowing for rational akrasia or rejecting evidentialism, neither of which is an attractive option.

I offered a way out: we can argue that agents only ever have evidence that supports an akratic conclusion when (and while) they don't perfectly conform to their evidence. The fact that some agents have akrasia-supporting evidence then doesn't show that akrasia can be rational any more than the fact that some agents have evidence for a Moorean conjunction establishes that believing such conjunctions can be rational.

To implement this strategy with the desired generality, I sketched a theory of evidence that predicts the blindspot thesis: while negative access can fail,

agents always have to believe that it is true of them. I showed that a schematic reliabilist-contextualist account of evidence naturally makes such a prediction, and explained how it offers an attractive akrasia-free treatment of the red wall, my toy counterexample to negative access. Of course, there is much that's controversial about that account, and about the argument that it predicts negative access failures to be 'elusive' in the way described, and I have not been able to defend every detail. But I think that the proposed implementation passes the threshold of plausibility required to show that my general strategy is viable.

If I am right that externalism and the rejection of akrasia can be reconciled in this way, then that is a significant result. One immediate consequence is that it strengthens the case for externalism, by showing how externalism can avoid some of the counter-intuitive consequences it is naturally taken to entail. In this way, my application of contextualism is quite similar to the more familiar application of contextualism to the sceptical puzzle. In both cases, the contextualist sides with the theorist who maintains that (in ordinary contexts) we have plenty of knowledge or evidence; but she does so whilst respecting the intuitions moving us towards scepticism, or internalism, in the first place.

Potentially even more important, however, are the consequences for debates amongst externalists. As I mentioned in §2, the denial of negative access is not the only potential tension between externalism and an anti-akrasia norm. If there is no hope for externalists to preserve anti-akratic intuitions, there is little point in exploring conciliatory resolutions to those other tensions. Once the non-negotiable denial of negative access has been reconciled with the rejection of akrasia, however, that project regains much of its interest and can serve as a useful dialectical lever in resolving disputes amongst externalists.[35]

On a more impressionistic level, the view suggested here also promises to offer a deeper insight into the internalism/externalism debate. It is easy to feel that how one comes down in this dispute depends on how one approaches

[35]Another interesting project for externalists to examine is whether the response suggested here can be extended to deal with problematic 'diachronic' consequences of denying negative access, such as issues about diachronic reflection principles (see e.g. Williamson (2000) and Salow (forthcoming)), or about the connection between accuracy and conditionalization (see e.g. Bronfman (2014) and Schoenfield (forthcoming)). I suspect that it can, but I cannot defend that suspicion here.

epistemology.[36] On the one hand, one can approach it from the first person perspective, asking questions such as 'what should I believe?' – internalism then seems almost inevitable. Alternatively, one can approach epistemology third-personally, asking questions such as 'what belief-forming mechanisms should creatures like us living in a world like ours be employing?' – externalism then seems very natural. The Elusive Externalism defended here offers a new way of vindicating this intuition. For the view predicts the access principle, and thus internalism, to be 'true from the first person perspective,' since any rational agent must believe that it holds of her. But the view also maintains that, when we step back and examine epistemological principles more impersonally, we can see that the access principle, and hence internalism, often fails. Showing how exactly this insight can help us to make progress elsewhere in the debate will, however, have to remain a project for future work.

[36] See, for example, many of the essays in Kornblith (2001).

Appendix A Externalism and Radical Akrasia

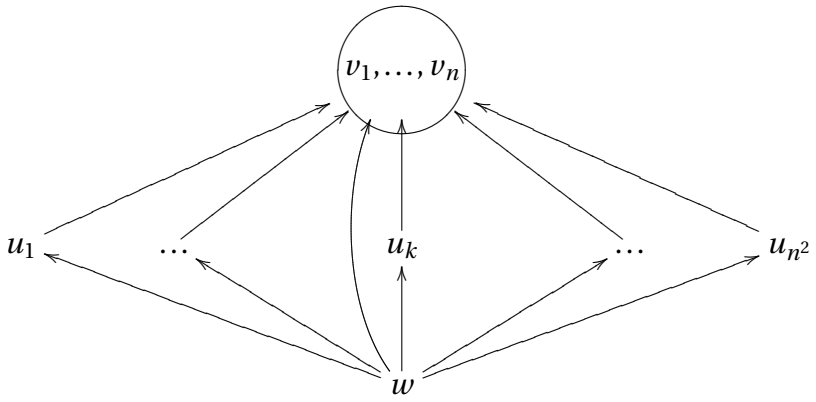
In the case of the red wall, an agent's evidence can support an akratic state. However, that state isn't *radically* akratic because the probability and expected probability of 'the wall is red' come apart only by .09. This may make the problem seem less serious. And it suggests that a different understanding of what it is for a graded state to be akratic might not classify this state as akratic at all.

This line of thought is not, I think, very promising, because we can construct structurally similar cases which do yield radical akrasia by any plausible account. In particular, we can describe situations in which an agent's evidence tells against p to an arbitrarily high degree whilst also making it arbitrarily likely that it tells in favour of p to an arbitrarily high degree. The state supported by such evidence is clearly *radically* akratic on any remotely reasonable account of akrasia.

I will first describe the case abstractly, using the tools of epistemic logic. We have a set W of possible worlds and an accessibility relation R between worlds such that wRw' only if the evidence had by the relevant agent at w does not entail that she is not in w'. For simplicity, W will be finite and the prior probability distributions uniform, so that the probability of $p \subseteq W$ at a world w is just the proportion of worlds accessible from w which are also in p.

The abstract model goes as follows. For some finite n, $W = \{w, v_1, v_2, \ldots, v_n, u_1, u_2, \ldots u_{n^2}\}$ and R is such that xRy iff (i) $x = w$ or (ii) $y = v_j$ for some j or (iii) $x = y$. That is: w can access every world; the v_j can access all and only each other; and each u_i can access itself and every v_j. As a diagram:[37]

[37] Pictorial conventions: any two worlds in the same circle can access each other, and any world that can 'access' a circle can access every world in that circle. To avoid clutter, I have omitted the reflexive arrows indicating that each world can see itself.



Now let $p = \{v_1, v_2, \ldots, v_n\}$. Then the probability of p at each u_i is $\frac{n}{n+1}$, which will approach 1 as n increases. Moreover, the probability at w that some u_i or other is actual is $\frac{n^2}{n^2+n+1}$ which approaches 1 as n increases. Finally, the probability at w of p is $\frac{n}{n^2+n+1}$, which approaches 0 as n increases. So for large n, the probability at w of p is arbitrarily close to 0, while the probability at w that the probability of p is extremely close to 1 is itself arbitrarily close to 1. This surely makes for radical akrasia by any measure.

Note that R is transitive and reflexive, and the prior probability distribution is the same at every world; the radical akrasia is thus entirely due to failures of symmetry and thus, ultimately, to failures of the negative access principle. Moreover, there is a (somewhat abstract) way to motivate the model which relies on a similar epistemological picture to the one motivating the example of the red wall.[38] For imagine a creature with n^2 independent mechanisms for learning facts that are, in this case, independent; and let us suppose that a mechanism yields knowledge if it is, in fact, delivering the truth, even if some other mechanism is being fooled.[39] Then w is the ultimate sceptical scenario, in which all the mechanisms are led astray; each v_j is a possibility in which all mechanisms are reliable; and each u_i is a possibility in which exactly one of the mechanisms delivers a falsehood. This yields the desired accessibility relations, since every working mechanism delivers information which rules out the possibility in which that mechanism is malfunctioning.

[38] Note also that the model is 'convergent' as well as transitive and reflexive; it thus validates S4.2, the strongest logic for knowledge advocated in the literature. So, given something like E=K, it is hard to see what could prevent the existence of situations for which this is the correct model.

[39] We can make it part of the agent's background knowledge that there won't be any Gettier cases; so that whenever a mechanism delivers the truth, it does so reliably.

Appendix B Blindspots and Akrasia

I claim that a probabilistically coherent agent who (i) knows exactly what the evidential support relation is, (ii) knows that her evidence obeys the positive access principle and (iii) all-out believes that her evidence satisfies the negative access principle will never be akratic. Since, as noted in §2, failures of (i) and (ii) are independent sources of akrasia, this establishes my claim in footnote 30 that, given the blindspot thesis, probabilistically coherent agents cannot be akratic due to failures of the negative access principle.

To prove this rigorously and in general, we need some formalism. We will look at models that are 4-tuples $< W, P, R_K, R_B >$, where W is the (finite) space of possibilities, P the evidential support relation, R_K the accessibility relation for the agent's evidence, and R_B the accessibility relation for what the agent all-out believes. Since our agent knows exactly what the evidential support relation is, it makes sense to model her prior credences as matching the a priori evidential support P. And it's also natural to assume that our agent's credences in a particular world are obtained by conditionalizing her priors on the set of possibilities consistent with her all-out beliefs at that world. We can thus define the agent's credence at a world as $Cr_w(p) =_{def} P(p|R_B(w))$. The evidential support a proposition has at a world is clearly given by $Pr_w(p) =_{def} P(p|R_K(w))$.

Since evidence is true, R_K is reflexive. By positive access, R_K is transitive. By the blindspot thesis, we have that, for each w, R_K is symmetric on $R_B(w)$. Finally, since the agent knows that she satisfies the positive access principle, we can assume that if our agent doesn't believe that the she knows p, she also won't all-out believe p. Since R_B is supposed to represent what the agent all-out believes, this yields the constraint that $R_K(u) \subseteq R_B(w)$ whenever $u \in R_B(w)$. So $R_K(u) = R_K(u) \cap R_B(w)$ whenever $u \in R_B(w)$.

Now, let w be an arbitrary world. Transitivity, reflexivity, and restricted symmetry guarantee thatR_K is an equivalence relation on $R_B(w)$, and hence partitions $R_B(w)$. Call the cells of the partition $c_1, \ldots, c_n$. Note that, for any $u \in c_i$, $R_K(u) \cap R_B(w) = c_i$.

Then the law of total probability gives us that

$$\begin{aligned}
P(p|R_B(w)) &= \sum_{1\leq i\leq n} P(c_i|R_B(w))P(p|R_B(w)\cap c_i) \\
&= \sum_{1\leq i\leq n} P(c_i|R_B(w))P(p|c_i) \\
&= \sum_{u\in R_B(w)} P(\{u\}|R_B(w))P(p|R_B(w)\cap R_K(u)) \\
&= \sum_{u\in R_B(w)} P(\{u\}|R_B(w))P(p|R_K(u)) \\
&= \sum_{u\in R_B(w)} P(\{u\}|R_B(w))Pr_u(p) \\
&= \sum_{x\in[0,1]} P(Pr(p)=x|R_B(w))x.
\end{aligned}$$

But, given our definition of Cr_w, that is just the anti-akrasia constraint that an agent's credence match her estimate of the evidential probability, i.e. that

$$Cr_w(p) = \sum_{x\in[0,1]} Cr_w(Pr(p)=x)x.$$

References

Jonathan Adler. *Belief's Own Ethics*. MIT Press, Cambridge MA, 2002.

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

Michael Blome-Tillmann. Knowledge and presuppositions. *Mind*, 118:241–295, 2009.

Michael Blome-Tillmann. *Knowledge and Presuppositions*. Oxford UP, Oxford, 2014.

Aaron Bronfman. Conditionalization and not knowing that one knows. *Erkenntnis*, 79:871–892, 2014.

Matthew Chrisman. From epistemic contextualism to epistemic expressivism. *Philosophical Studies*, 135:225–254, 2007.

David Christensen. Does murphy's law apply in epistemology. In T. Szabo Gendler and J. Hawthorne, editors, *Oxford Studies in Epistemology*, volume 2. Oxford UP, Oxford, 2007.

David Christensen. Rational reflection. *Philosophical Perspectives*, 24:121–140, 2010.

Allen Coates. Rational epistemic akrasia. *American Philosophical Quarterly*, 49: 113–24, 2012.

Nilanjan Das and Bernhard Salow. Transparency and the KK principle. *Noûs*, forthcoming.

Keith DeRose. Solving the sceptical puzzle. *Philosophical Review*, 104:1–52, 1995.

Keith DeRose. Single scoreboard semantics. *Philosophical Studies*, 119:1–21, 2004.

Kevin Dorst. Enkrasia demands perfection. ms.

Fred Dretske. Epistemic operators. *The Journal of Philosophy*, 67:1007–1023, 1970.

Fred Dretske. *Knowledge and the Flow of Information*. MIT Press, Cambridge MA, 1981.

Adam Elga. Reflection and disagreement. *Noûs*, 41:478–502, 2007.

Adam Elga. The puzzle of the unmarked clock and the new rational reflection principle. *Philosophical Studies*, 164:127–139, 2013.

Jeremy Fantl and Matthew McGrath. Evidence, pragmatics, and justification. *Philosophical Review*, 111:67–94, 2002.

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

John Gibbons. Access externalism. *Mind*, 115:19–39, 2006.

Alvin Goldman. Discrimination and perceptual knowledge. *The Journal of Philosophy*, 73:771–791, 1976.

Daniel Greco. A puzzle about epistemic akrasia. *Philosophical Studies*, 167: 201–219, 2014a.

Daniel Greco. Could KK be OK? *Journal of Philosophy*, 111:169–197, 2014b.

Daniel Greco. Cognitive mobile homes. *Mind*, 126:93–121, 2017.

John Hawthorne. *Knowledge and Lotteries*. Oxford UP, Oxford, 2004.

Wesley Holliday. Epistemic closure and epistemic logic I: Relevant alternatives and subjunctivism. *Journal of Philosophical Logic*, 44:1–62, 2015.

Sophie Horowitz. Epistemic akrasia. *Noûs*, 48:718–744, 2014.

Sophie Horowitz and Paulina Sliwa. Respecting *all* the evidence. *Philosophical Studies*, 172:2835–2858, 2015.

Jonathan Ichikawa. Quantifiers and epistemic contextualism. *Philosophical Studies*, 155:383–98, 2011a.

Jonathan Ichikawa. Quantifiers, knowledge and counterfactuals. *Philosophy and Phenomenological Research*, 82:287–312, 2011b.

Niko Kolodny. Why be rational? *Mind*, 114:509–563, 2005.

Hilary Kornblith. *Epistemology: Internalism and Externalism*. Wiley Blackwell, Malden MA, 2001.

Maria Lasonen-Aarnio. Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88:314–345, 2014.

Maria Lasonen-Aarnio. New rational reflection and internalism about rationality. In T. Szabo Gendler and J. Hawthorne, editors, *Oxford Studies in Epistemology*, volume 5. Oxford UP, Oxford, 2015.

David Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74:549–567, 1996.

John MacFarlane. The assessment sensitivity of knowledge attributions. In T. Szabo Gendler and J. Hawthorne, editors, *Oxford Studies in Epistemology*, volume 1. Oxford UP, Oxford, 2005.

John McDowell. Criteria, defeasibility and knowledge. *Proceedings of the British Academy*, 68:455–479, 1982.

John McDowell. Knowledge and the internal. *Philosophy and Phenomenological Research*, 55:877–893, 1995.

John McDowell. *Perception as a Capacity for Knowledge*. Marquette University Press, Milwaukee WI, 2011.

Connor McHugh. Self-knowledge and the KK principle. *Synthese*, 173:231–257, 2010.

Ram Neta. Contextualism and the problem of the external world. *Philosophy and Phenomenological Research*, 66:1–31, 2003.

Ram Neta. A contextualist solution to the problem of easy knowledge. *Grazer Philosophische Studien*, 69:183–206, 2005.

Bernhard Salow. Lewis on iterated knowledge. *Philosophical Studies*, 173:1571–1590, 2016.

Bernhard Salow. The externalist's guide to fishing for compliments. *Mind*, forthcoming.

Dov Samet. On the triviality of higher-order probabilistic beliefs. *Journal of Philosophical Logic*, forthcoming.

Miriam Schoenfield. Conditionalization does not (in general) maximize expected accuracy. *Mind*, forthcoming.

Virginia Schultheis. Worlds not properly ignored. ms.

Declan Smithies. Moore's paradox and the accessibility of justification. *Philosophy and Phenomenological Research*, 85:273–300, 2012.

Declan Smithies. Ideal rationality and logical omniscience. *Synthese*, 192:2769–2793, 2015.

Roy Sorensen. *Blindspots*. Oxford UP, Oxford, 1988.

Ernest Sosa. How do defeat opposition to Moore. *Philosophical Perspectives*, 13: 141–154, 1999.

Robert Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128: 169–199, 2006.

Robert Stalnaker. Luminosity and the KK thesis. In S. Goldberg, editor, *Externalism, Self-Knowledge, and Skepticism*. Cambridge UP, Cambridge, 2015.

Jason Stanley. *Knowledge and Practical Interests*. Oxford UP, Oxford, 2005.

Michael Titelbaum. Rationality's fixed point (or: In defence of right reason). In T. Szabo Gendler and J. Hawthorne, editors, *Oxford Studies in Epistemology*. Oxford UP, Oxford, 2015.

Brian Weatherson. Stalnaker on sleeping beauty. *Philosophical Studies*, 155: 445–456, 2011.

Brian Weatherson. Do judgements screen evidence? ms.

Roger White. What is my evidence that here is a hand? In D. Dodd and E. Zardini, editors, *Scepticism and Perceptual Justification*. Oxford UP, Oxford, 2014.

Timothy Williamson. *Knowledge and its Limits*. Oxford UP, Oxford, 2000.

Timothy Williamson. Probability and danger. *The Amherst Lecture in Philosophy*, 4:1–35, 2009.

Timothy Williamson. Improbable knowing. In T. Dougherty, editor, *Evidentialism and its Discontents*. Oxford UP, Oxford, 2011.

Alex Worsnip. The conflict of evidence and coherence. *Philosophy and Phenomenological Research*, forthcoming.