ACCESS AND ANTICIPATION

by

Bernhard Salow

MMathPhil, University of Oxford (2009)
BPhil, University of Oxford (2011)

Submitted to the Department of Linguistics and Philosophy in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy in Philosophy

at the

Massachusetts Institute of Technology

September 2015

Signature of author: Signature redacted

Department of Linguistics and Philosophy
August 4, 2015

Certified by: Signature redacted

Roger White
Associate Professor of Philosophy
Thesis Supervisor

Accepted by: Signature redacted

Alex Byrne
Professor of Philosophy
Philosophy Chair and Chair of the Committee on Graduate Students

Access and Anticipation

by Bernhard Salow

Submitted to the Department of Linguistics and Philosophy on August 4, 2015 in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Philosophy.

Abstract

Can we always tell, just through reflection, what we should believe? That is the question of access, the central disagreement between epistemic internalists and externalists, and the focus of the dissertation.

Chapter 1 gives an argument for access, connecting it with the question of whether we can intentionally bias our own investigations to favour desirable hypotheses. I argue that we can't: since we have to take any known biases into account when evaluating the evidence obtained, attempts to bias our inquiries will be self-undermining. Surprisingly, this explanation fails for agents who anticipate violating access; and such agents can in fact intentionally bias their investigations. Since this possibility remains counterintuitive when we focus on alleged counterexamples to access, this is a serious problem for externalism.

Chapters 2 and 3 offer a solution to this problem and related, more familiar, ones. Chapter 2 lays some technical foundations, by investigating iterated knowledge in David Lewis's contextualist theory of knowledge. I show that his account has the surprising consequence that agents cannot attend to "negative access failures", cases in which someone fails to know something without knowing that they fail to know it. Whilst this prediction is prima facie unattractive, I show how it can be defended.

Chapter 3 uses this Lewisian treatment of negative access failures to solve our problems for externalism. For I show that these problems arise not from maintaining that, in some situations, agents are unable to tell what they should believe, but rather from maintaining that rational agents can sometimes suspect that they are currently in such a situation or anticipate that they will be in such a situation in the future. Externalists can reject this stronger thesis. To explain how, I sketch a theory of evidence which integrates the Lewisian treatment of negative access failures to predict that agents always have to think that they can tell what they should believe, even though this isn't always true. By rejecting access, but maintaining that agents can never anticipate violating it, this theory reconciles the most attractive features of externalism and internalism.

Thesis Supervisor: Roger White
Title: Associate Professor of Philosophy

Acknowledgements

I enjoy philosophy most when doing it with others. So it is hardly surprising that this dissertation has been shaped by a large number of people other than myself.

First, there is my dissertation committee. Roger White offered just the right balance of detailed and 'big-picture', substantive and presentational feedback. Bob Stalnaker usually knew where my ideas were going long before I did. Steve Yablo's unexpected examples and analogies reminded me of why I love philosophy in the first place.

While not officially on my dissertation committee, Alex Byrne, Declan Smithies, and Jonathan Vogel performed a similar role, reading multiple versions of my papers, and providing extremely helpful and detailed guidance on how to develop and present my thoughts.

I also benefited hugely from less formal conversations. Two groups of people deserve special notice here. First, there is MIT's flourishing epistemology team. Conversations with its members, especially Nilanjan Das, Kevin Dorst, Sophie Horowitz, and Ginger Schultheis, have shaped my ideas and their presentation more than I can articulate. Second, there are three of my all-time favourite people to talk philosophy with: Brendan de Kenessey, Jack Spencer, and Josh Thorpe; they deserve special credit for spending so much time with me thinking about topics that were often quite far removed from those that they themselves work on.

In addition to those mentioned above, I also had extremely helpful comments from and/or discussions with Stew Cohen, Jeff Dunn, J. Dmitri Gallow, Jeremy Goodman, Dan Greco, John Hawthorne, Brian Hedden, Abby Jacques, Justin Khoo, Maria Lasonen-Aarnio, Harvey Lederman, Jack Marley-Payne, Agustin Rayo, Damien Rochford, Kieran Setiya, Alex Silk, Ian Wells, and four anonymous referees. I am also grateful to audiences at the 2012 ARCHE/ CSMN graduate conference, the 2014 Formal Epistemology Workshop, and the 2014 Joint Session of the Aristotelian Society and the Mind Association. Finally, I am extremely lucky to have been a part of the wonderful MIT philosophy community over the last few years – the feedback I received at numerous workshops at which I presented these ideas was extremely helpful; but it hardly scratches the surface of the influence that this entire community has had on my philosophical development.

My last, and greatest, debt is to Ellen Salow. If it wasn't for her, I would be less patient, less experimental, and more argumentative; in short, I would be a

significantly worse philosopher. If it wasn't for her willingness to accompany me here, despite the unfamiliar and distant country and the unclear career prospects, I would not have come to MIT; the result would, again, have been a very different philosopher. For these reasons, and many others less directly relevant to the dissertation, I am extremely glad to have her in my life.

Chapter 1

The Externalist's Guide to Fishing for Compliments

Abstract

Suppose you'd like to believe that p (e.g. that you are popular), whether or not it's true. What can you do to help? A natural initial thought is that you could engage in *Intentionally Biased Inquiry*: you could look into whether p, but do so in a way that you expect to predominantly yield evidence in favour of p. The paper hopes to do two things. The first is to argue that this initial thought is mistaken: intentionally biased inquiry is impossible. The second is to show that reflections on intentionally biased inquiry strongly support a controversial 'access' principle which states that, for all p, if p is (not) part of our evidence, then that p is (not) part of our evidence is itself part of our evidence.

Some truths are bleak. Faced with the prospect of discovering such truths, we might prefer not to know; we might even prefer to have false beliefs. For the same reason, however, if the bleak claim turns out to be false, finding that out would be extremely reassuring. So, when faced with a question which may have a bleak answer, we often feel ambivalent about inquiring. Whether we want to know the answer depends on what the answer is.

Take an example. If my colleagues like me, I'd love to know. But if they don't, I still want to believe that they do. The negative effects which knowing they don't like me, or even just becoming less confident that they do, would have on my self-esteem and my ability to sustain reasonably productive relationships simply far outweigh any possible advantages such knowledge or doubts might generate. Given such preferences, what am I to do?

A somewhat plausible initial thought is that I could inquire into whether my colleagues like me, but do so in a way that is more likely to yield evidence in one direction rather than the other. I could, for example, talk primarily to people who I expect to have a high opinion of me if anyone does; and I could avoid reading the blogs and facebook comments which I expect to be particularly harsh. Moreover, doing this does not seem to require that I deceive myself or exploit any straightforward kind of irrationality, e.g. some predictable failure to correctly assess my evidence.

In this paper, I hope to do two things. The first is to argue that this initial thought is mistaken: intentionally biased inquiry is impossible. The second is to show that reflections on intentionally biased inquiry strongly support a controversial 'access' principle, which is a natural component of epistemological internalism:[1]

> For all *p*, if *p* is (not) part of one's evidence, one's evidence entails that *p* is (not) part of one's evidence.[2]

I will begin, in §1, with a preliminary precisification and defence of the claim that intentionally biased inquiry is impossible. I will then, in §2, turn to consider influential alleged counterexamples to the access principle, showing that, if they were genuine, they would support strategies for biasing one's inquiries that obviously wouldn't work. Finally, in §3, I will explain the more abstract connection between the two topics, by formalizing both questions in a probabilistic

[1] Titelbaum (2010) presents related, but somewhat less general, considerations. I should note that, in presenting the argument in this way, I am being a little disingenuous. I actually suspect that we might be able to do justice to our observations about intentionally biased inquiry even if we reject the access principle, provided we maintain that rational agents always *have to think* that it is true of them in the present and future. But it is far from obvious that this gap is interesting: if a principle sometimes fails, why couldn't someone reasonably think that it will fail for them? In Salow (msa), I argue that this question can be answered, and that other important arguments for the access principle can also be avoided in this way; but, since my views on this are idiosyncratic, I will set them aside here. The point of this footnote is thus simply to note that, while this paper is written as an argument for the access principle, readers sceptical of this conclusion can instead read it as an advertisement for the kind of view just hinted at.

[2] Here, and throughout, I assume that evidence consists of propositions.

framework and showing that the access principle and the impossibility of intentionally biased inquiry stand and fall together. This strengthens our objection against access deniers, by showing that the problems we uncovered are unavoidable consequences of denying the access principle rather than irrelevant issues arising from the particular examples; and it strengthens our initial defence of the impossibility of intentionally biased inquiry, by showing that the strategies exploiting the alleged counterexamples to access, strategies we previously saw to be patently absurd, were in fact the most promising ones.

1 Intentionally Biased Inquiry

We are interested in questions for which our desire to know the truth depends on what the truth turns out to be. If intentionally biased inquiry were possible, it would be natural to use it in these cases. We have already seen one example with this structure: I want to know about my popularity only if I'm popular; and so I always prefer believing that my colleagues like me, whether or not they actually do. Other examples aren't hard to come by. If I'm worried that I might have a fatal and untreatable illness, it would be great to discover that I need not be concerned. But that might not be sufficient reason to inquire into the matter in an open-minded way. For if I do have the fatal illness, I would prefer not to find out: that way, I can at least live out my final days in peace, without the constant awareness of my impending death ruining whatever small chance of happiness remains for me. In fact, if I am about to die, I want to be not only ignorant but positively mistaken about that; I would thus quite like to receive misleading evidence reassuring me that I am in good health.

In cases like these, intentionally biased inquiry would be appealing: I would like to inquire into the matter in a way that I know will only, or at least predominantly, yield evidence in a particular direction. But can I? Several philosophers have endorsed a positive answer more or less explicitly. Parfit (2011, p.421), for example, writes:[3]

> [W]e might cause ourselves to have some beneficial belief by finding evidence or arguments that gave us strong enough epistemic reasons to have this belief. This method is risky, since we might find evidence or arguments that gave us strong reasons not to have this belief. But we might reduce this risk by trying to avoid becoming aware of such reasons. If we are trying to believe that God exists, for example, we might read books written by believers, and avoid books by atheists.

I will argue that, despite this initial appearance, intentionally biased inquiry is not, in fact, possible. But before I can do that, I will need to make more explicit what exactly would count as 'intentionally biased inquiry.'

1.1 What is Intentionally Biased Inquiry?

What is it for an inquiry into whether p to be biased? Perhaps the most obvious case is if the inquiry is a **sure-win investigation**: the total evidence produced by the investigation is guaranteed to be evidence for p. Somewhat more generally, the inquiry still seems biased if it is what Titelbaum (2010) calls a **no-lose investigation**: the inquiry is designed so that the total evidence produced might be evidence for p, but is guaranteed not to be evidence against p. But even this is insufficiently general – as is clear from Parfit's discussion, it can, intuitively, be enough to reduce (rather than eliminate) the risk of evidence against p. In particular, if an inquiry is very likely to yield powerful evidence for p and has only a small chance of yielding at most weak evidence against p, it still seems biased.

[3] Another, less explicit, endorsement seems present in Kripke's (2011) discussion of his dogmatism paradox. Kripke's paradox is an argument for (amongst other things) deciding to further favour a known claim over its negation, by selectively avoiding counter-evidence. He explains that what he has in mind is a resolution to avoid "reading the wrong books (for they contain nothing but sophistry and illusion), associating with the wrong people, and so on" (2011, p.49). This seems to assume that the actions described could be a way of intentionally favouring the known claim over its negation.

The most natural way of making this idea precise appeals to the notion of the *expected value* of a function V; this is the weighted average of all the values V might take, weighted by the probability that it will take that value. Intuitively, we want to say that the inquiry is biased in favour of p if, on balance, we expect an increase in the evidential support for p. So our function V should measure the extent to which the evidential support for p at the end of inquiry exceeds the evidential support for p at the start of inquiry. It's natural to measure the 'extent to which the evidential support increases' simply by subtracting the initial evidential support from the later evidential support; the expected value can then be calculated by assigning probabilities to the various values which this difference might take. We can then say that an inquiry is **biased towards** p if the expected value of the difference is positive, **biased against** p if it is negative, and unbiased if it is zero.[4]

We now know what it is for an inquiry to be biased.[5] Can an agent take steps to make sure her own inquiries are biased in this way? In two kinds of cases, she obviously can. The first is if her current actions themselves provide evidence

[4] This definition means that we count as unbiased an investigation which might nonetheless strike us as odd. Suppose I care very unevenly about evidence of my own popularity: the only thing I want is to have evidential support of degree at least .9 that I am popular. Support of degree .3 is no worse than support of degree .7, and support of degree .99 is no better than support of degree .91. As Kelly (2002, p.170) points out, I could then decide to keep inquiring until the first moment my evidence reaches support .9, and stop immediately after. That decision increases the chances that my preferences will be satisfied; but this form of investigation needn't count as biased on our definition.
The preference structure motivating this kind of odd investigation isn't plausible, I think, in the cases we described above, where more evidence is always better, and less evidence always worse, and so no stopping point is better than any other. But such preferences might make sense if the pay-offs associated with our attitudes are 'all or nothing.' Pascal's Wager might illustrate this kind of case: there is some attitude which is minimally sufficient to count as believing (and thus for being rewarded as a believer); greater confidence in God's existence has no (comparable) benefit, and any less committal attitude will keep you out of heaven forever.

[5] Or rather, what it is for a body of information to classify an inquiry as biased, since we need a body of information to calculate the expected evidential support (or, in the less general version discussed earlier, to determine what is 'guaranteed' to and what 'might' happen). I will henceforth suppress this qualification and say that A's inquiry is biased if it is biased relative to A's (initial) evidence.

about p, as when p is "I develop lung cancer at some point in my life" and the agent can decide to smoke. The second is if the agent expects to lose relevant information in the course of the investigation. She could, for example, ask a friend to tell her a year from now that she is popular, knowing that she will forget having given these instructions; it seems that, in this way, she will be able to manufacture evidence of her popularity for herself. But these ways of biasing one's inquiry seem intuitively very different from the one Parfit identifies: when I exploit the evidential relevance of my own actions, there is nothing weird about my *inquiry* (as opposed to the topic I am investigating), while when I exploit information-loss my biasing only succeeds because, at some point or another, I lack important information about what is going on.[6] I will thus set them aside, and restrict 'intentionally biased inquiry' to cases in which an agent succeeds at biasing her investigations only in intuitively 'open-eyed' ways.

A potential third kind of case is ruled out because, following Parfit, we characterized the relevant notion of 'inquiry' as yielding *evidence* rather than *beliefs*. It obviously is possible to exploit known biases in our perception or reasoning to set up an inquiry that is more likely to leave us with one belief rather than another. For example, it seems plausible that group pressures might ultimately instil in me a belief in God if I were to associate only with devout theists. I take it, however, that (if the belief really is just caused by group pressure) this would be a case in which I form a belief without adequate evidence, so associating only with devout theists doesn't count as biasing one's inquiry on our understanding.[7]

[6] Parfit (2011) himself explicitly sets aside the cases where one's own actions are evidentially relevant to the question.

[7] This is compatible with different epistemological accounts of such biases. On the most straightforward account, they lead to belief without supplying any additional evidence. But even if they do supply evidence (for example, by affecting what 'feels plausible' which arguably is a form of evidence), the evidential value of this evidence is plausibly nullified when we know that it arises *only* because of the biased mechanism. (Kelly (2008) offers a more sophisticated account of how biases can offer evidence. He agrees (p.629) that once we know about these biases, the import of this evidence is undermined.) But I need to know about my biases if I am planning to exploit them. So either way, my *total* evidence at the end of the investigation will not support theism any more than it does at the outset, even if I do form the belief.

Similarly, in cases in which I exploit arational changes in how I assess evidence (if these are possible[8]), I might succeed in biasing my future beliefs without exploiting irrationality. But, again, this would not qualify as intentionally biasing my inquiries, since it doesn't involve biasing what evidence I receive.

1.2 Intentionally Biased Inquiry is Impossible

With this sharpening of the question in mind, we can revisit the question of whether intentionally biased inquiry is possible. Parfit's brief discussion makes it sound straightforward. Suppose I can choose between reading a creationist book and a biology textbook; then surely I can bias my inquiries against evolutionary theory by choosing the creationist one. On reflection however, this thought starts to become less obvious. For it might well be that the facts I encounter in the creationist book are (even) less impressive than I expected; if this is the case, my epistemic position with respect to evolutionary theory isn't compromised, and may even be strengthened. Similarly, reading the biology textbook might well give me evidence against evolutionary theory, if the facts appealed to there turn out to be less conclusive than I thought they would be. When we think about it this way, it starts to seem plausible that each book will only give me additional evidence in the 'expected' direction if the facts presented there are actually more compelling than anticipated; and, of course, it is hard to see how I could think *that* to be particularly likely.

How can we reconcile these thoughts? It's a familiar observation that whether a proposition E is evidence for or against another proposition H often depends on what background information is available. And relevant background information can, amongst other things, include facts about how the evidence has been selected. Suppose, for example, that you are sitting on a jury. The prosecution has just finished making its fairly compelling case for the defendant's guilt. Nonetheless, the rational thing for you to do at this stage is to keep an open mind. After all, you knew all along that they were only going to bring up the

[8] See e.g. Kelly (2013) and Schoenfield (2014) for defence.

incriminating facts that presents the defendant in a particularly negative light. And the facts they did present were no more compelling than you would expect from such a one-sided advocacy. The evidence you received would, *against ordinary background information,* favour the defendant's guilt. However, given what you know about how the facts you were exposed to have been selected, they do not favour his guilt *against your background information.* Everything thus depends on the rebutting evidence which the defence is about to introduce. You can be pretty confident that, *against ordinary background information,* what the defence will present will be evidence of innocence. But, if that evidence ends up being weaker than you are currently expecting, you will (rationally) come to the conclusion that the defendant is probably guilty. There is thus no guarantee that what will be presented will be evidence of innocence *against your background information.*

The case of choosing which books to read is, I think, exactly analogous.[9] I expect the creationist book to contain facts which, against 'ordinary' background information, tell against evolution; that is to say, I think it likely that someone with no background expectations would be rational to be less confident of evolutionary theory after reading the creationist book than after reading the biology textbook. However, this doesn't mean that I expect the creationist book to contain facts which, against my background information, tell against evolution. This is because I have expectations and, as we saw above, these expectations change the evidential impact of the information I would obtain by reading the book. In particular, if I am confident that one of the books will contain facts of a certain kind (e.g. facts which are quite hard to account for in evolutionary terms), then I must already be confident that there are facts of that kind; so

[9] An important complication, which I will continue to ignore, is that academic books of the kind Parfit envisions often contain arguments as well as teaching one new facts. (You may have noticed that I changed the example from 'does God exist' to 'is the theory of evolution correct'; the intention was to make this point slightly less pressing.) If we don't think of arguments as presenting us with new evidence, but rather as enabling us to better satisfy the demands of rationality by showing us what our evidence really supported all along, we may have to think of those cases slightly differently.

my current view about evolution should already 'factor in' these anticipated facts. Finding out that the book doesn't contain facts of that kind (they are all easier to account for than I expected) might then suggest that I was 'factoring in' difficulties that, as it turns out, aren't genuine; in this way, the facts I learn favour evolution (relative to *my* background information), despite highlighting its problems.

What matters to our success in biasing our inquiries is what the new evidence will support against our own background information. So these considerations show that, *pace* Parfit, it isn't obvious that we can intentionally bias our inquiries; the claim only looks obvious if we mistake it for the truth that one can manipulate one's inquiry to make its outcome favour *p relative to 'ordinary' background information.* In addition, the considerations suggest an explanation for why intentionally biasing one's inquiries might be impossible. The evidential impact of a proposition on *p* depends on our expectation that it is true, and that we would learn it, if *p* is true. But those expectations can change as we know different things about the set-up. It is no surprise *to us* that the prosecution can find some facts that make the defendant look bad; we would expect them to be able to do this even if he is in fact innocent. Similar things apply to the creationist literature. When we try to bias our own inquiry into *p* by affecting what evidence we might find, we are automatically changing what it is rational to expect, and are thereby changing which claims count as evidence for or against *p*. In other words, knowing of the bias (as we must, since the biasing is intentional and 'open-eyed') undermines its effects.

I haven't, as yet, given a direct argument against the possibility of intentionally biased inquiry. And I won't be able to provide a fully general argument until §3. Nonetheless, it will be helpful to have pursued the intuition in a few additional cases. In order to do so more effectively, let me return to a topic that is more important than the controversy over evolution: the question of whether I am popular. Intuitively, questions like this one, where our desire to know the answer depends on what the answer is, put us in a sort of bind; that is why, if we inquire at all in such cases, we do so only reluctantly. A very natural explanation of this bind is that intentionally biased inquiry isn't possible: there is simply nothing we can do to shift the expected outcome of the inquiry towards the answer that we want to see supported and away from the answer we want to avoid learning. Again, this point will become clearer if we consider some concrete examples.

Let's assume that I have a very loyal friend, who knows me and my beliefs extremely well. Suppose my friend also knows how popular I am with my colleagues. Could I somehow use my friend to set up a biased inquiry into my popularity? On reflection, it seems implausible that I could. I could, of course, ask him to tell me what he knows about my popularity if and only if I am popular. But now his silence will, for me, be evidentially equivalent to him telling me that nobody likes me, and so I have not reduced the risk of bad news.

Moreover, it doesn't help if we change the case so I don't know whether my friend knows whether I am popular. I could again ask him to tell me what he knows only if he knows that I am popular. Unlike in the previous case, his silence in the face of my request would not now be decisive evidence that nobody likes me; after all, he might just be silent because he doesn't know the answer. However, his silence is some evidence that I am unpopular, since his knowing that I am unpopular would explain it rather well. This will be so even if he is in fact silent because he doesn't know. So I have eliminated the risk of decisive evidence that I am unpopular only at the cost of turning a state which, had I not made the request, would have been evidentially neutral, into a state which yields weak evidence that I am unpopular. Where previously there was a small chance of decisive refutation, there is now instead a pretty good chance of weaker disconfirmation. Plausibly, the expected outcome of the inquiry hasn't changed at all. It is natural to expect that more realistic cases will merely add complications without escaping this underlying structure. My efforts to bias my inquiries can't help undermining themselves.

These considerations constitute at least a *prima facie* case that, if we restrict ourselves to fully transparent ways of biasing our inquiry, Parfit is mistaken. There is nothing we can do to steer our inquiry into p in a particular direction in such a transparent way. If there was, it would be unclear why cases where we 'prefer not to know' put us in the kind of bind described above; yet they manifestly do. Moreover, any initial impression that we *can* manipulate our inquiry disappears if we are careful to set aside the (irrelevant) notion of one claim being evidence for another relative to ordinary background information, and focus on which claims support which others relative to the agent's own background information. Together, these observations make a strong *prima facie* case that, contrary to Parfit, this kind of biasing isn't possible.[10]

2 The Access Principle

I have been building a case that intentional and open-eyed biasing of one's own inquiry isn't possible. Epistemologists rarely discuss this issue.[11] I think this is a significant oversight, since the issue can help shed important light on the debate between epistemological internalism and externalism.

[10] This might be a good point to mention an interesting case that initially seems to raise doubts for my position (Thanks to Roger White, Caspar Hare, Vann McGee, and Jack Spencer for discussion). Note that we can investigate using 'stopping rules' that intuitively seem biased. Suppose, for example, that I want to know whether a coin is fair or biased towards heads, knowing already that it isn't biased towards tails. I could decide to keep tossing the coin until it has landed heads more often than it's landed tails. Since I know that the coin isn't biased towards tails, I can be sure that this will happen at some point, so that such an investigation is bound to yield a result. (If I didn't know that the coin isn't biased towards tails, I could not be sure of this.) But couldn't I know in advance that this result will favour heads-bias? Interestingly, I cannot. The reason is that more heads than tails needn't favour heads-bias over fairness. Given normal background beliefs, a sequence containing 49 tails and 50 heads supports fairness over heads bias; and no matter what background beliefs I have, there will always be some length such that sequences of length larger than that will support fairness over heads-bias. Admittedly, these sequences are less likely to occur than the ones favouring heads-bias; but this is balanced out by the fact that the ones favouring heads-bias generally favour it only very weakly.

[11] Which is not to say that the considerations I have been raising are entirely original. They clearly connect to Popper's (1961) famous claim that falsifiability is a precondition for testability (and hence, we might add, for confirmation). For some recent related discussion, see also White (2006, p.543-549), Sober (2009), and Titelbaum (2010).

A natural component of internalism is the access principle:

> **The Access Principle:** For all p, if p is (not) part of one's evidence, one's evidence entails that p is (not) part of one's evidence.

For internalism says, very roughly, that we can always work out which of our beliefs are justified (i.e. supported by our evidence) merely by reflecting on what we've already got. But if that is to be true, then what we've already got (i.e. our evidence) had better tell us what our evidence is. In other words, the access principle had better be true.[12]

For some purposes, it will be helpful to have separated the access principle into the positive and negative access principles:

> **Positive Access:** For all p, if p is part of one's evidence, one's evidence entails that p is part of one's evidence.
>
> **Negative Access:** For all p, if p is not part of one's evidence, one's evidence entails that p is not part of one's evidence.

Each of these two principles is controversial. Williamson's (2000, ch.9) view that one's evidence consists of all and only the propositions one knows, abbreviated as E=K, helps to bring this out, since each of the corresponding 'introspection' principles for knowledge faces well-known objections.[13] However, it should be clear that simply rejecting E=K does not make the principles unproblematic. For the very same considerations that made the 'introspection' principle problematic for knowledge can also be used to argue directly against the access principle for evidence.

In the next two subsections, I will consider two different adapted arguments of this kind that seem, on first sight, quite convincing: one is based on an alleged

[12] Different theses go under the label 'internalism', and not all of them will be committed to the access principle; Wedgwood (2002), for example, emphatically denies it. But since I will be *defending* the principle, I will not pursue these subtleties.

[13] Hintikka (1962) already rejected the 'negative introspection' principle because it seems clear that, when we mistakenly believe something, we fail to know it without knowing that we so fail. Williamson's (2000) more recent arguments against 'positive introspection' (also known as the *KK* principle) have also proven influential.

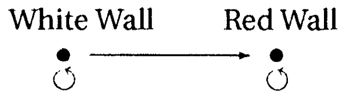
epistemic asymmetry between 'good' and 'bad' cases (§2.1), the other on our limited discriminatory capacities (§2.2). I will sketch how these considerations motivate particular counterexamples to the access principle; I will then show that if these really were counterexamples to the access principle, someone could exploit these cases to intentionally bias her inquiry in ways that obviously don't work. This refutes the (alleged) counterexamples, and thereby casts doubt on the considerations which motivated them. It thus constitutes an indirect defence of the access principle. It also sets the scene for §3, where I draw out the more systematic connection between the access principle and the possibility of biasing one's inquiry, a connection I take to support both the access principle and our tentative conclusion that intentionally biased inquiry is impossible.

2.1 Good and Bad Cases

The first kind of example targets only the negative access principle. To get an example of this kind, we try to exhibit a 'good case' and a 'bad case' such that (i) in the good case, my evidence entails that I am definitely in the good case but (ii) in the bad case, my evidence leaves open whether I am in the good case or the bad one. I will thus be in the bad case if and only if it isn't part of my evidence that I am in the good case. It follows that negative access must fail in the bad case. For if negative access held in the bad case, that case would yield the evidence that it isn't part of my evidence that I am in the good case. But this piece of evidence entails (combined with the description of the cases) that I am in the bad case; thus contradicting the stipulation that my evidence in the bad case leaves open that I might be in the good case.

One powerful motivation for accepting such examples arises from sceptical worries. It is part of my evidence that I have hands: after all, I can see that I have them just by looking, and denying that seeing yields evidence quickly leads to scepticism. A (handless) brain in a vat with my exact experiences could not have the corresponding claim as part of its evidence, since the corresponding claim is false.[14] Yet the brain in a vat is presumably in no position to tell that it lacks this evidence: if it were, it could conclude that it is in a very unusual situation, and the tragedy of the brain's predicament is exactly that it is in no position to figure this out.

If one takes this position regarding me and the brain in a vat, there is pressure to take a similar line in more ordinary cases. Although the particulars won't matter, I will take the following as representative: I can get conclusive evidence that a red wall is red by looking at it in normal circumstances; but if the wall is actually a white wall lit by a red light, my only evidence will be that it appears red. In particular, in the case where I am being fooled, my evidence does not allow me to work out that 'the wall is red' is not itself part of my evidence.[15] To isolate the structural feature, it might help to consult this diagram of the situation:



In the diagram, the dots represent the possibilities that might (for all your initial evidence entails) obtain, whilst an arrow from w to w' represents that w' is left uneliminated by the evidence to be obtained if w is the case.

Our reflections on intentionally biased inquiry reveal problems in this way of thinking about the case. For consider again my interest in whether people like me. I would like my evidence to support that they do, whether or not people actually do like me. Normally we tend to think that this puts me in a bind when it comes to inquiring into the matter, even if I have a reliable and cooperative source I could consult. But if the above verdicts about the wall are correct, this is

[14] Here and elsewhere I assume that only truths can be evidence. Williamson (2000, ch.10) influentially defends this claim; Goldman (2009) seems to deny it. See Littlejohn (2013) for a discussion of the more recent literature.

[15] There are other cases which we might go to for counterexamples to negative access of this kind. For example, one might want to maintain that when one sees (as opposed to hallucinates) that p, it becomes part of one's evidence that p, yet also maintain that in the hallucination case one doesn't have evidence that one isn't seeing. Similarly, one might want to maintain that when a person A testifies (knowingly) that p, it becomes part of one's evidence that p even though had A been lying, one would not have been able to tell.

an illusion. What I should do is the following. I should find a white wall, and ask my friend (who knows about my popularity) to paint it red if I am in fact popular, and shine a red light on it otherwise. Once my friend has finished setting things up, I will take a peek. If people actually do like me, I will see a red wall, and will thus receive conclusive evidence that they do. But if I am unpopular, I will get no evidence at all, and, in particular, no evidence confirming my unpopularity. This means that the strategy described may give me evidence for popularity, and certainly won't give me evidence against it; it is a no-lose investigation. Since I want to be more confident that I am popular, and increasing what evidence I have favouring my popularity seems a good way of achieving this goal,[16] it seems clear that I should initiate the strategy.[17]

This conclusion is obviously absurd: the procedure just outlined is no way out of the bind we have been discussing. If I actually were to engage in it, I would be wasting my time since looking at the wall would tell me nothing about its colour, and hence nothing about my popularity.[18] If that's right, this is not a

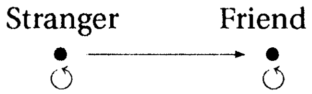
[16] One might challenge this step, since the cases in question are precisely cases in which I might not be able to tell what my evidence is, and it isn't clear that in such cases increased evidence really does lead to increased confidence. I explain why a response along such lines doesn't allow us to avoid all versions of the problem in §2.3.

[17] See Titelbaum (2010) for a similar observation about cases like that of the wall. Titelbaum presents this as a diagnosis of why the 'bootstrapping' reasoning apparently supported by the wall case seems so bad. (For classic discussion of the bootstrapping problem, see Vogel (2000) and Cohen (2002).) Since, as I will argue in §2.2 and §3, we can set up similarly biased investigations *whenever* we have a violation of the access principle, and not all these cases seem to exploit 'bootstrapping', I am less sure about the connection between these two problems.

[18] Someone otherwise sympathetic to the case might try to accommodate this result by saying that, in the scenario where I try to manipulate my inquiries, knowledge of the painting strategy acts as a defeater of the strengthened evidence even in the 'good case' in which the wall really is red (though see Lasonen-Aarnio (2010) for worries whether the externalist views that might be sympathetic to cases like that of the wall can really make sense of defeat). The challenge for this response is, of course, to give an account of why knowledge of the painting strategy, but not knowledge of e.g. one's susceptibility to hallucination or of the possibility of brains in vats, can function as a defeater. (Note that strategies which rely on the relative 'closeness' of error possibilities are not particularly promising, since the worlds in which the wall isn't painted red needn't, in any intuitive sense, be particularly 'close.' I might be safely popular, and my friend might be a very reliable executor of my instructions.) Our discussion sharpens this familiar challenge, by bringing out that an adequate account would have to entail that all cases in which

counterexample to negative access after all. Of course, it is a deep puzzle how to accept this without falling into scepticism. But that is not a puzzle I can resolve in this paper.

To see how robust this problem is, it is worth looking at how matters play out in the more intuitive counterexamples to negative access recently proposed by Lasonen-Aarnio (forthcoming).[19] The basic idea behind her examples is that, as she puts it, it's not unusual that "coming across a fake, one can mistake it for the real thing, but when one sees (feels, hears, tastes, smells) the real thing, one can tell that it is not a fake" (forthcoming, p.160). Here's one such case (not quite the one used by Lasonen-Aarnio, but similar enough, I think): you're meeting a friend from school that you haven't seen in many years. As you sit in the agreed coffee shop, several people walk in who look familiar enough that you think they might well be that friend. None establish eye-contact though, so you stay seated. Eventually your friend walks in and, despite the significant changes she's undergone, you recognize her immediately. At least at first sight, this suggests a similar structure to the case of the red wall: seeing your friend (the good case) yields conclusive evidence that it's her, while seeing the stranger (the bad case) gives you no evidence either way.



Because the case has the same structure as that of the red wall, it can be exploited in exactly the same way. I ask my helper to present me with my friend from school if I'm popular, and with a somewhat similar looking stranger if I'm not. (The person will then leave before I have a chance to talk to her.) If the above description were correct, this should ensure that I have set up a no-lose investigation. But, intuitively, it is clear that I haven't: unless seeing the person

one tries to use the failures of the negative access principle to manipulate one's own inquiries involve such defeat. It is hard to see how to predict this in a principled fashion.

[19] Thanks to Jack Spencer and Maria Lasonen-Aarnio for discussion, and to Jack Spencer for the particular case.

triggers a powerful feeling of recognition, and if I'm not popular it won't, I will have received evidence that I'm not popular after all.

Yet Lasonen-Aarnio's thought about cases like that of the coffee shop also seems right. What is going on here? In an ordinary case of the type she describes, I don't know beforehand that I will definitely recognize the real thing. For example, I don't know beforehand that seeing my friend will trigger a powerful sense of recognition; for all I know at the outset, she will only look vaguely familiar. I also know little about whether a stranger will strike me as entirely unknown or as vaguely familiar, though I do know that a stranger would not trigger a powerful sense of recognition. There are thus (at least) four possibilities for what might happen when someone enters the coffee shop: it could be my friend, and I have a powerful sense of recognition; it could be my friend, and she looks vaguely familiar; it could be a stranger who looks vaguely familiar; it could be a stranger who strikes me as entirely unknown. When I see my friend and have a powerful sense of recognition, I can rule out all but the first of these, and thus get conclusive evidence that it is my friend. When I see a familiar looking stranger, I can rule out all but the middle two, and thus don't get much evidence about whether the person is my friend or a stranger. But all of this is consistent with me always knowing exactly what evidence I have; in particular, when I see a familiar looking stranger, I know that my evidence doesn't entail that I'm faced with my friend.

To get a counterexample to negative access, we would need to argue that the situation remains unchanged if I know beforehand that I will definitely recognize my friend when I see her. For with that background knowledge in place, negative access would allow me to reason from 'my evidence doesn't entail that this is her' to 'it isn't her', and it was supposed to be counterintuitive that I can reach this conclusion. But if we imagine actually having the background knowledge in question, that reasoning actually seems attractive: if I know that I would recognize my friend if I saw her, it's perfectly legitimate to reason from 'that person only vaguely looks like her' to 'that's not her'.[20] This is well brought out by the attempt of using the case to set up a no-lose investigation. To do so, I have to think that I would definitely recognize my friend when I see her, since otherwise seeing someone who only looks vaguely familiar is (some) evidence that I am unpopular (since I'm guaranteed to see someone like that if I'm unpopular, and less likely to see such a person if I'm popular). But once we make this assumption explicit, the intuition the example relied on disappears.

2.2 Limited Discriminatory Capacities

Our second class of examples target the positive access principle as well as the negative one. Williamson (2000) has influentially used considerations stemming from limited discriminatory capacities and margins for error to argue for the existence of such cases; I will focus here on an example proposed in Williamson (2011), which has received sympathetic discussion even by philosophers not otherwise committed to Williamsonian epistemology.[21]

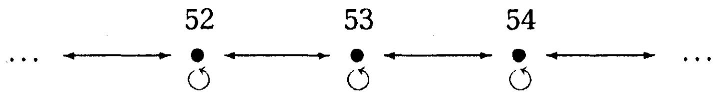
The basic case can be described as follows. Imagine that you are faced with an unmarked clock, with a single hand that can point in any one of 60 slightly different directions. Your ability to discriminate where it is pointing is good, but not unlimited. If you are to be reliable in your guesses, you need to leave yourself a margin of error. For example, if the hand is in fact pointing at 53 minutes, you can reliably judge that it is pointing somewhere between 52 and 54 (inclusive), but are unreliable about claims stronger than that. The same is true of every other position the hand could be in.[22]

[20] My dentist recently told me that if I needed further treatment, I'd know that I did. This was useful information: it allowed me to reason from "I feel (only) mild pain" to "I don't need further treatment", which I would not have been able to do otherwise.

[21] See, for example, Christensen (2010), Elga (2013), Horowitz (2014), and Horowitz and Sliwa (forthcoming). Our discussion could easily be adapted to other alleged counterexamples to 'positive introspection' for knowledge, such as the case of Mr Magoo in Williamson (2000, ch.5).

[22] This last claim is, of course, an idealization, since actual humans are probably better at discriminating when it's pointing at 15, 30, 45, or 60, than when it's pointing in other directions. Moreover, it is also unrealistic to suppose that our evidence allows us only to rule out some

It is somewhat natural to identify your evidence with the strongest claim about the hand's position which you can reliably get right.[23] But this means that the total evidence propositions we learn in the various scenarios partially overlap. If the hand is in fact pointing at 53, my evidence will be that it is within [52,54]; and if it is pointing at 52, my evidence will be that it is within [51,53]. Each scenario yields evidence which is compatible with being in the other, and yet they yield different evidence. Again, it might help to present this as a diagram:



It isn't hard to see why this case involves a violation of the positive access principle. For note that, given the above description, there is a 1-1 correspondence between positions of the hand and what evidence I have.[24] Given knowledge of the set-up, the truth about what evidence I have thus entails the truth about where exactly the hand is pointing. But my evidence isn't good enough to single out the hand's position: after all, I can't reliably do so (even if I do know the set-up), and this isn't a failure of rationality. So positive access must fail.

To see why this description of the case is problematic, let us return to the topic of my popularity. My friend knows whether I'm popular; and I would like to have additional evidence that I am, regardless of whether it is true. So I ask him to construct an unmarked clock of the kind Williamson describes, and I ask him to set the hand in the following way: if people like me, he will set it to 53; if they don't, he will flip a coin to decide whether to set it to 52 or to 54. Then I

positions; presumably it also supports the remaining possibilities to a degree which is proportional to their proximity to the true value. Finally, the description is a little misleading since it is presumably vague what margins are sufficient for my guesses to count as 'reliable'. But these idealizations won't matter.

[23] See also Williamson (2011), Christensen (2010), and Elga (2013) for slightly different routes to the same conclusion. (If there are worries about the clock being part of the external world, we can instead change the case to be about where the hand is pointing *according to your visual experience*. I find it plausible that the point about limited discriminatory abilities applies here too.)

[24] There is also a 1-1 correspondence between positions of the hand and what evidence I lack; so a similar argument establishes that the case violates the negative access principle.

take a look. If people actually like me, it will be set to 53, and so my evidence will only tell me that it is somewhere between 52 and 54, which I knew beforehand by knowing the set-up. So if people like me, I get no new evidence. But if people do not like me, it will be set either to 52 or to 54. Suppose it is set to 52; then my evidence will allow me to rule out that it's set to 54, since 54 far enough away from the actual setting. But I knew that there was a 50-50 chance that it would be set to 54 if people didn't like me. So seeing that it isn't set to 54 gives me some evidence that I am popular. Moreover, my evidence cannot discriminate between the hand being set to 52 and it being set to 53, so that I get no evidence against my being popular. So, if the hand is set to 52, I will get evidence that I am popular; by similar reasoning, I will also get such evidence if the hand is set to 54. So if people don't like me, I will get evidence that I am popular. Again, I have successfully set up a no-lose investigation into my popularity.

This method is, I take it, slightly less satisfying than the one involving the wall. Both are no-lose investigations: I might get evidence that I am popular, and run no risk of getting evidence that I am not. But, in the clock case, I will get evidence for my popularity only if it is misleading; if I actually am popular, I will get no evidence at all. Fortunately, it doesn't matter too much to me. Perhaps added evidence that I am popular is even better when it's pointing me towards the truth; but given the desirability of self-confidence, it is welcome to me even when it is misleading.

The evaluation we just went through is, of course, absurd. I cannot boost my evidence for my popularity in the way just described. The unmarked clock is no more a way out of my bind than the wall was. That much is obvious; what is surprising, perhaps, is that this is inconsistent with the Williamsonian judgements about the clock. The culprit, I think, was to identify my evidence with the strongest claim I am reliable about given the actual setting, which was the move that gave rise to the failures of the positive (and negative) access principle we were exploiting. What else could my evidence be in the scenario described above? One possibility, following Stalnaker (2009), might be that my evidence is instead determined by my best guess about the exact position of the hand: if my best guess it that it's pointing at X, my evidence is that it's pointing somewhere between X-1 and X+1.[25] Crucially, my best guess won't always match the actual position (that, we might say, just is what it is for my discriminatory capacities to be limited), so that different hand positions will sometimes yield the same best guess and the same hand position will sometimes yield different best guesses. We thus lose the tight connection between the hand's actual position and what my evidence is. This, in turn, allows us to hold on to the thought that I am always in a position to know what my evidence is, thereby preventing failures of the access principle and the absurd conclusions those entail.

2.3 Beliefs, Rationality, and Evidence

I have argued that certain cases which supposedly illustrate failures of the access principle shouldn't be thought to do so, since they would otherwise vindicate strategies for intentionally biasing one's inquiries which obviously wouldn't work. It is time to consider a response to this line of argument. In my discussion so far, I have moved somewhat freely between a desire for evidence of my popularity and a desire to be more confident that I am popular, on the assumption that (to the degree that I am rational) these go together. The cases of imperfect access to one's evidence just discussed, however, might put pressure on the connection between the two. Arguably, it is only rational (or only rationally required) to raise one's confidence in a claim if *one knows that* it is now better supported by one's evidence than it previously was;[26] and an agent using the strategies described

[25] Cohen and Comesaña (2013) also defend an approach along these lines. For criticism of this approach, see Hawthorne and Magidor (2010), Goodman (2013, p.34) and Williamson (2013, p.80-83). Since I can't discuss how to respond to such criticisms here, gesturing towards this alternative construal of the case remains a promissory note.

[26] This line is somewhat analogous to Hawthorne and Stanley's (2008, p.580-85) view that one should only act on one's evidential probabilities if one knows what they are. It is also suggested, albeit somewhat loosely, by Williamson's (2009b, p.360-61) view that there are different senses or precisifications of 'justified confidence', one which requires merely that the confidence matches one's evidence and others which require in addition that one knows this (and perhaps knows that

above plausibly doesn't know that he received evidence for popularity even when he did. One could then accommodate the observation that the clock and wall strategies for boosting one's evidence are an absurd way of pursuing one's goals, but maintain that this is not because they don't yield evidence but rather because that evidence won't yield the desired beliefs in agents like ourselves.

Different worries could be raised about this strategy. We might complain that it does violence to the theoretical role of 'evidence' to allow that rational agents sometimes ignore their evidence, regardless of whether they know that they have it. Or we might point out that, in the case of the wall at least, it's not so clear why the agent who receives evidence of his popularity shouldn't also know that this is what happened. Fortunately, however, we need not enter such difficult territory to show that the response is inadequate. For we can use the kinds of examples we have been discussing to build not just *no lose investigations* but also *sure win investigations*: ones that are guaranteed to give me evidence that I am popular no matter what. And if I have set up a sure win investigation, I will know at the end of the investigation that I just received additional evidence of popularity. I should thus raise my confidence even if one should only do so when one knows that one has received evidence. Yet the strategies seem equally absurd in these only slightly more complicated cases.

What do these sure win investigations look like? The simplest one just combines the strategies discussed in the clock and wall cases: I ask my friend to both arrange a wall and a clock in line with the instructions described above, and then look at one and then later at the other. The result is a sure win investigation. For either I am popular or I am not. If I am, looking at the wall will yield evidence that I am popular and looking at the clock will yield nothing. If I am not popular, looking at the wall will yield nothing but looking at the clock will yield evidence that I am popular. Either way, the total effect will be additional evidence that I am popular. So, since I know the set-up, I will know at the end of inquiry that I

one does, etc.). For this view implies that, if the agent doesn't know of the evidence boost, the increased confidence would not be justified in at least one sense of 'justified'; it is thus plausible that increased confidence isn't rationally required, in at least one sense of 'rationally required'.

just received additional evidence of popularity. But if I *know* that my evidence supports my popularity at least to a specific degree x (and, when we fill in some details, I can put such a lower bound on how strong the evidence was), then I would surely be irrational not to have at least the corresponding level of confidence in the claim that I am popular.[27] If that is true, however, one would be irrational not to become more confident of one's popularity after looking at both the clock and the wall.[28]

The combined case just described is effective against an opponent who is sympathetic to both the wall and the clock case. However, the epistemological motivations for these two cases are quite different and, at least on first sight, one might be inclined to accept one without accepting the other. Would this argument break down against such an opponent?

It would not, since we only need one kind of case to set up a sure win investigation. Let us look first at the cases motivated by our limited discriminatory capacities. The original instructions in the clock case were designed to generate no evidence either way when I'm popular, and evidence that I'm popular when I'm not. But they are easily adapted to provide a way of receiving evidence that I'm popular when I'm popular and no evidence either way when I'm not. The new instructions to my friend will simply be to set the hand to 52 or 53 if I am popular (to be decided by a coin flip) and to 51 or 54 if I am not. If I am popular, the evidence I'll get will then allow me to rule out exactly one of the possibilities in which I'm not, and will hence generate evidence of my popularity. And if I'm

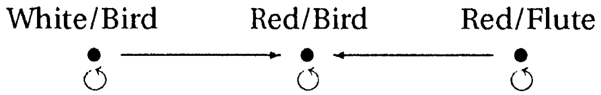
[27] Note also that if I can have many iterations of knowledge about the set-up (and why shouldn't I be able to?), I could have many iterations of knowledge that my evidence supports popularity at least to degree x. We could thus weaken this conditional, by strengthening the antecedent to require many iterations of knowledge, and still run our objection.

[28] This style of argument also applies to responses, perhaps inspired by Gallow (2014), Bronfman (2014), and Schoenfield (ms), which maintain that agents should generally update by a rule other than conditionalization in situations where the access principle isn't antecedently known to hold. For an alternative update rule can prevent agents from manipulating their confidence in the combined case only by requiring that agents sometimes can't raise their confidence in a claim even though they are certain that their total evidence now supports it more than it did previously; and that still strikes me as an unfortunate consequence. (Thanks to Dmitri Gallow for discussion.)

not popular, the evidence will let me rule out one possibility of either kind, and will thus leave the initial probabilities unchanged. I will thus get evidence for popularity if it is true, and no evidence at all if it isn't. Alternating this strategy with the original instructions, I can ensure that I'll receive evidence of my popularity *no matter what*; I can thus know, at the end of the process, that I've just received evidence that I am popular (though, I will not know what exactly that evidence was).

Let us look next at the cases motivated by the thought that the good case yields strictly more evidence than the bad one; this time I will construct a case which, just by itself, allows for a sure win investigation. If you liked the wall case, you should also like the following case: when I hear a bird call nearby, I receive conclusive evidence that there is a bird nearby; this is true even though I cannot distinguish bird calls from the noises produced by some sophisticated bird flutes. It also seems clear that my ability to tell red walls by sight and nearby birds by their sound are quite independent: one of them malfunctioning should not prevent the other from yielding the evidence it usually does.

If that is right, I can give my friend the following instructions. If I am popular, he is to present me with a red wall and a genuine bird call; if I am unpopular, he is to toss a coin to decide which of these to replace with the corresponding illusion. There are thus three possibilities consistent with my knowledge of the set-up; and the evidential relations amongst them are as represented in this diagram:[29]



What will happen to my evidence about my popularity? If I am popular, I will get conclusive evidence that the wall is red and that there is a bird nearby, and thus that I am popular. If I am not popular, there are two possibilities. One is that I am in White/Bird; in that case, my evidence rules out Red/Flute, since

[29] This case has the same structure as Williamson's (2000, ch.10) 'simple creature' example; but the epistemological story is quite different.

it entails that there is a bird nearby and there are no nearby birds in Red/Flute. The other is that I am in Red/Flute; my evidence then rules out White/Bird, since it entails that the wall is red, which it isn't in White/Bird. In either of the two possibilities in which I am unpopular, my evidence eliminates exactly one possibility in which I am unpopular and nothing else. So in those possibilities, I also get evidence that I am popular. So I will get evidence that I am popular whether I am popular or not. Afterwards, I will thus not only have more evidence of popularity but also know that I do.

There are other interesting variants that we could discuss. But we have done enough to make the required point. Explaining the absurdity of the strategies described in the alleged counterexamples to access by forcing a gap between rational beliefs and evidence has some initial appeal. It is not entirely unnatural to think that one shouldn't become more confident of one's popularity even if one just received evidence for it when one doesn't (and isn't in a position to) know that one received such evidence. But we have just seen that the same motivations to those driving the initial cases can be used to generate cases in which one does know that one received evidence for one's popularity. Yet the strategies described in these cases seem equally absurd. The attempted alternative explanation of the absurdity thus fails.[30]

[30] A related defensive strategy is available to someone who thinks that belief in *p* is a necessary condition for *p* being part of one's evidence (as follows naturally from E=K). For one might then say that whilst in the above cases one is in a position to know the various claims, it would be quite unreasonable to form the relevant belief. Since a reasonable subject wouldn't form such beliefs, they wouldn't get the problematic evidence. The methods for biasing one's inquiries described above thus aren't methods that are available to reasonable people, and this is why they strike us as absurd. (Thanks to Alex Byrne for discussion.)

Here, unreasonableness might be understood along the lines suggested by Lasonen-Aarnio (2010): forming a belief is unreasonable if it would manifest a disposition which often leads to beliefs that are false (or otherwise fall short of knowledge). But, on that understanding, the belief-forming methods needn't be unreasonable: my policy might be to form the crucial belief in the wall case only if my prior confidence that I am unpopular (and that I will thus be subject to an illusion) is less than 5%. This policy will only rarely lead me astray; but this method for biasing one's inquiry still strikes us as absurd.

A second problem with this defence is more general. The tight connection between evidence and belief seems independently problematic: can I really avoid receiving evidence just by failing

3 The Systematic Connection

In the paper so far, I have done two things. I have given *prima facie* reasons to be sceptical about the possibility of intentionally biased inquiry. And I have shown that we can cast doubt on otherwise compelling counterexamples to the access principle by showing that, if genuine, they would enable us to intentionally bias our inquiries in highly counterintuitve ways.

What I haven't done, so far, is establish a systematic connection between access and intentionally biased inquiry. This gives rise to questions that challenge the force of the arguments. Can *every* counterexample to the access principle be exploited to bias one's inquiries? If not, the problems with the particular cases discussed might stem from idiosyncratic features of, or simplifying assumptions about, those cases, and so need not be taken to support the access principle more generally. Could there be ways of biasing one's inquiries that do not exploit failures of the access principle? If not, the fact that the intentionally biased inquiry made possible by access failures is so counterintuitive does little to reinforce our tentative early conclusion that intentionally biased inquiry is impossible.

In this section, I will show that there is a systematic connection between our topics. In §3.1, I explain how to formalize the notion of intentionally biased inquiry within a probabilistic theory of evidential support. This will allow me, in §3.2, to show (i) that the impossibility of intentionally biased inquiry in fact follows from the access principle (together with any assumptions implicit in the probabilistic theory) and (ii) that, if the access principle is false, we should expect that people can exploit the counterexamples to the principle to bias their

to form certain beliefs, e.g. because I am prone to wishful thinking? (Nick Hughes raised this interesting worry to me in a different context.) There are natural responses for a defender of E=K: we could reject the connection between knowledge and belief; we could maintain that, in the relevant sense of 'belief', reflective endorsement (or even a disposition to reflectively endorse) is not necessary for belief; or we could retreat to the claim that one's evidence consists of all the claims that one is in a position to know. But it seems that, if we opt for any of these, the defensive strategy imagined here will fail.

own investigations. We thus get affirmative answers to both of the questions raised above, strengthening both the case for the access principle and the case against the possibility of intentionally biased inquiry. Along the way, we will see some interesting connections between the possibility of biased inquiry and the 'reflection principle' widely discussed in formal epistemology.

3.1 Formalizing Biased Inquiry[31]

Recall the lessons of the discussion of §1. We were wondering whether you could embark on a course of action (e.g. set up your inquiry in a certain way) that would (from your perspective) make your investigations favour a claim p over its negation, even though which action you choose does not itself provide evidence regarding p. The action was not supposed to achieve this effect in a way that exploits irrational biases or information loss; it was instead meant to have its effect by influencing what evidence would become available to you. And 'favouring' was supposed to be understood in terms of expected value: p would be favoured over its negation if the expected difference between future and present evidential support for p, given that you decide to inquire in this way, was positive.

To formalize the thought that one cannot, in this sense, manipulate the force of the evidence that will become available, we need to introduce some technical notions. Let $a_1, a_2, \ldots$ be the total, and thus pairwise incompatible, courses of action you might take (for all that your evidence entails); and let $A_1, A_2, \ldots$ be the corresponding propositions stating which (if any[32]) of those actions you perform. Moreover, let $Pr(p)$ represent p's current evidential probability. Then the expected value of some function V on the hypothesis that I perform action a_k

[31] Thanks to Jeremy Goodman and Harvey Lederman for extremely helpful discussion on this section.

[32] This qualification, that you might not perform any action at all, means that one proposition in the list (the one stating that you don't perform an action) will not correspond to anything on our list of actions. I will assume that one can't expect failing to act to bias one's inquiries for the same reason that one can't expect particular actions to.

will be the weighted average of the values $V_1, V_2, \ldots$ which V might take, weighted by $Pr(V = V_i|A_k)$, the probability that V will take that value if I perform a_k.

The function whose expected value we're interested in measures the difference between the initial and future evidential support for a proposition p. Each of these is plausibly determined by two factors: which propositions are part of the agent's evidence at the relevant time, and what those propositions support. It seems possible to imagine uncertainty about either of those factors; I can be unsure both about the colour of the hundred emeralds I will examine, and about the extent to which the observation that all of them are green would support the hypothesis that the next emerald to be mined is green. But, for our purposes, it makes sense to idealize away from the second kind of uncertainty. For uncertainty about the evidential support relation is orthogonal to both the access principle and our reasons for analysing biased inquiry in terms of expected probabilities, namely that we don't typically know beforehand what particular evidence some investigation will yield. And if we idealize away from uncertainty about the evidential support relation, we can take the values of the initial and future evidential support, and thus the value of the difference between them, to be fully determined by what our initial and future evidence is.

To make use of these conceptual points, we need further terminology. Let $E_1, E_2, \ldots E_n$ be the propositions which might, for all your initial evidence entails, be your total initial evidence; and let $E_1^+, E_2^+, \ldots E_m^+$ be the propositions which might, for all your initial evidence entails, be your total evidence at the relevant future time, the time at the end of the investigation. Furthermore, for each $1 \leq i \leq n$, let $E = E_i$ be the proposition that your total evidence at the initial time is E_i; and for each $1 \leq j \leq m$, let $E^+ = E_j^+$ be the proposition that your total evidence at the future time is E_j^+.[33] Finally, let P be (what you know to be) the evidential support relation, so that $P(p|E_j^+) - P(p|E_i)$ is the difference between

[33] Recall that, if the access principle fails, E_i and $E = E_i$ can be very different. If you are in the bad case, your evidence is entirely uninformative about the colour of the wall. But the proposition that you have this uninformative evidence is itself highly informative: it entails that you are in the bad case, and hence that the wall is white.

the future and the initial evidential support if your total initial evidence is E_i and your total later evidence is E_j^+. Then we can write the claim that, conditional on any A_k, the expected difference is 0 as[34]

$$\sum_{i,j} Pr(E^+ = E_j^+ \wedge E = E_i | A_k)(P(p|E_j^+) - P(p|E_i)) = 0. \qquad (\neg\text{IBI})$$

Such large equations are difficult to survey, so it will be helpful to have an alternative notation. I will use '$exp_Q V_i$' as an abbreviation for the expected value of V, as calculated by Q; and I will use '$Q(.|X)$' as a label for the probability function Q' obtained by setting $Q'(Y) = Q(Y|X)$ for every Y. (¬IBI) can then be rewritten as:

$$exp_{Pr(.|A_k)}(P(p|E_j^+) - P(p|E_i)) = 0.$$

In what follows, I will always give both ways of writing each equation.

It's easy to see that rearranging and simplifying (¬IBI) yields

$$\sum_i Pr(E = E_i | A_k) P(p|E_i) = \sum_j Pr(E^+ = E_j^+ | A_k) P(p|E_j^+)$$
$$exp_{Pr(.|A_k)} P(p|E_i) = exp_{Pr(.|A_k)} P(p|E_j^+)$$

From this equation, we can make our way towards something more recogniz-

[34] If we hadn't idealized away from uncertainty about P, we would have had to take a different approach. We would have used Pr^+ to stand for the agent's future evidential support; and we would have taken $X \subseteq [0,1]$ to be a finite set containing all the values the future evidential support might take and $Y \subseteq [0,1]$ to be a finite set containing all the values the initial evidential support might take. (¬IBI) would then have been written as

$$\sum_{x \in X, y \in Y} Pr(Pr^+(p) = x \wedge Pr(p) = y | A_k)(x - y) = 0.$$

The other equations in the text can be rewritten in similar ways, and the same connections hold between the rewritten principles as between the ones I discuss. The rewritten principles, however, strike me as less illuminating: they fail to capture the idea that, in determining whether an inquiry is biased, we wonder about how likely we are to learn various things, and how those things would impact on the proposition in question. Moreover, the connection between the rewritten principles and the access principle is a bit less straightforward.

able. For note that the actions in question, being total courses of actions for the relevant time-span, are mutually incompatible; moreover, since the list is a complete list of the actions you might take, your evidence entails that you either perform one of them or fail to act at all. This means that the set of propositions $\{A_1, A_2, \ldots\}$ forms a partition of the set of possibilities compatible with your evidence (which is to say that every such possibility is one in which exactly one of these propositions is true). But it is a straightforward theorem of the probability calculus, known as the law of total probability, that the probability of H is equal to the weighted average of the conditional probability of H on the members of a (finite) partition of the underlying space of possibilities. In symbols, $Q(X) = \sum_k Q(A_k)Q(X|A_k)$. But then we can use (¬IBI) to get a principle from which the action propositions have dropped out altogether, namely[35]

$$\sum_i Pr(E = E_i)P(p|E_i) = \sum_j Pr(E^+ = E_j^+)P(p|E_j^+) \qquad \text{(EQEXP)}$$

$$exp_{Pr}P(p|E_i) = exp_{Pr}P(p|E_j^+)$$

And *this* claim, stating that the expected future probability is equal to the expected initial probability, is the obvious consequence of two instances of van Fraassen's (1984) Reflection Principle, one synchronic and one future-

[35] Proof:

$$\begin{aligned}
\textstyle\sum_j Pr(E^+ = E_j^+)P(p|E_j^+) &= \textstyle\sum_j P(p|E_j^+)\sum_k Pr(E^+ = E_j^+|A_k)Pr(A_k) && \text{by total probability}\\
&= \textstyle\sum_k Pr(A_k)\sum_j P(p|E_j^+)Pr(E^+ = E_j^+|A_k) && \text{rearranging}\\
&= \textstyle\sum_k Pr(A_k)\sum_i P(p|E_i)Pr(E = E_i|A_k) && \text{using } (\neg\text{IBI})\\
&= \textstyle\sum_i P(p|E_i)\sum_k Pr(E = E_i|A_k)Pr(A_k) && \text{rearranging}\\
&= \textstyle\sum_i Pr(E = E_i)P(p|E_i) && \text{by total probability}
\end{aligned}$$

directed:[36,37]

$$Pr(p) = \sum_i Pr(E = E_i)P(p|E_i) \qquad \text{(S-REF)}$$

$$Pr(p) = exp_{Pr}P(p|E_i)$$

$$Pr(p) = \sum_j Pr(E^+ = E_j^+)P(p|E_j^+) \qquad \text{(F-REF)}$$

$$Pr(p) = exp_{Pr}P(p|E_j^+)$$

In fact, I want to go slightly further, and say that (EQEXP) is the result of 'subtracting' a commitment to (S-REF) from a commitment to (F-REF), the natural way of holding on to 'what we really wanted out of' (F-REF) without presupposing (S-REF). To explain what I mean by this, I should emphasize a non-standard feature of (F-REF) and (S-REF). Unlike other formulations of reflection principles, mine ask that an agent's probabilities match expected evidential support, not expected credences. This is a good thing, because it means that we side-step objections to reflection arising from the fact that future credences might be formed a- or ir-rationally.[38] And, crucially, it means that (S-REF) is a version of Christensen's (2010) rational reflection principle, stating that the

[36] I am here stating the reflection principle as a claim about expected values; that claim is usually treated as an important and immediate consequence of the principle rather than as the principle itself: see e.g. van Fraassen (1995, p.19) and Weisberg (2007, p.180). (Though Williamson (2000, p.230-237) also focuses on the claim about expected values when discussing reflection.) In our terminology, the 'standard' version of the (future-directed) principle would be $Pr(H|P(H|E^+) = c) = c$, where E^+ is a non-rigid designator for the agent's future evidence. Given that we are abstracting away from uncertainty about the evidential support relation, this 'standard' principle entails, but is not entailed by, (F-REF).

[37] Kadane et al. (1996) use a principle similar to (F-REF) to capture the thought that agents cannot 'reason to a foregone conclusion.' I explain below why (EQEXP) is the better choice if we want to avoid begging the question against access deniers.

[38] See Briggs (2009) for an excellent survey of those objections. Another well-known worry about reflection, also discussed by Briggs, arises from cases where an agent might lose evidence; since we've noted from the very beginning that (¬IBI) is only plausible in cases where we can be sure this won't happen, these worries also aren't relevant for our purposes.

probability of a proposition matches the expected current evidential support. But it is well known that it is hard to reconcile rational reflection with denials of the access principle, and rational reflection is widely rejected for this reason.[39] Moreover, violations of (S-REF) will quickly make for violations of (F-REF), for example if the agent knows that she will get no evidence in the relevant time span. So anyone, such as myself, who is interested in the specifically diachronic aspects of (F-REF), needs to find a way of articulating the specifically diachronic thought underlying (F-REF) in a way that insulates it from more immediate worries about (S-REF). I submit that (EQEXP) is the natural candidate for such a principle: in the presence of (S-REF) it's equivalent to (F-REF); but, unlike (F-REF), it is not subject to 'cheap' counterexamples constructed by considering a case in which (S-REF) fails and adding to the story that the agent knows she will receive no new evidence.[40]

This is not to say that (EQEXP) is or should be neutral on whether the access principle is true.[41] Williamson (2000, ch.10) and Weisberg (2007) highlight a serious tension between (F-REF) and denials of the access principle;[42] and, as we will see in the next section, the particular tension they discuss doesn't go away when we move from (F-REF) to (EQEXP). The point of replacing (F-REF) with (EQEXP) thus isn't to preserve neutrality, but to gain dialectical effectiveness. By focusing on (EQEXP), we make clear that the problem of intentionally biased inquiry is an additional problem for access deniers, over and above any problems they might face because of their rejection of rational reflection. In particular, it shows that this problem cannot be dismissed as arising from an intuitively attractive, but ultimately mistaken, endorsement of 'level bridging'; after all, the

[39] See e.g. Christensen (2010), Williamson (2011), Elga (2013), and Lasonen-Aarnio (forthcoming).

[40] The claim that (EQEXP) articulates the key insight behind (F-REF) can be further bolstered by showing that it, rather than (F-REF), is the principle which other arguments for (F-REF) really support, once we give up on (S-REF). I plan to do this in future work.

[41] Moreover, since, as we will see shortly, (S-REF) follows from the access principle together with our decision to ignore uncertainty about the evidential support relation, it will ultimately turn out that (EQEXP) is, in some derivative sense, committal even on whether (S-REF) holds.

[42] See also Hawthorne (2004, p.75-77) and Weatherson (2011) for briefer discussion.

two sides of (EQEXP) are both at the same epistemic 'level.'

3.2 Reflection and Access

What exactly is the connection between (EQEXP), (F-REF), and (S-REF) on the one hand, and the access principle on the other? It is relatively easy to see that (S-REF) will be true whenever the current evidence satisfies the access principle and that (F-REF) will be true whenever the current evidence entails that the total future evidence will satisfy the access principle.[43] The point about (S-REF) is rather trivial. For suppose the agent has evidence E_j. Since E_j obeys the access principle, E_j entails $E = E_j$.[44] So $Pr(E = E_i) = 0$ whenever $E_i \neq E_j$. (S-REF) thus holds trivially.

The argument for (F-REF) is only slightly more complicated. Even without the access principle, the plausible claim that only truths can be evidence already means that each $E^+ = E_j^+$ entails the corresponding E_j^+. As we just saw, the access principle for the future evidence allows us to establish the converse, that each E_j^+ entails the corresponding $E^+ = E_j^+$. The access principle thus guarantees that, for any H, and any E_j^+ you might receive, $P(H|E_j^+) = P(H|E^+ = E_j^+)$. Moreover, since E^+ subsumes your initial evidence E, $P(H|E^+ = E_j^+) = Pr(H|E^+ = E_j^+)$. It is also clear that $\{E^+ = E_j^+ : 1 \leq j \leq m\}$ is a partition of the possibilities left open by the current evidence, so that

$$Pr(H) = \sum_{j=1}^{m} Pr(E^+ = E_j^+)Pr(H|E^+ = E_j^+)$$

is simply an instance of the law of total probability. Substituting $P(H|E_j^+)$ for $Pr(H|E^+ = E_j^+)$, this yields (F-REF).[45]

[43] Assuming, as I will henceforth, that there will be no information loss.

[44] Why? Consider any $E_j \neq E_i$. Then either there is some p which is part of E_i but not E_j or there is some p which is part of E_j but not E_i (or both). If the former, E_i rules out $E = E_j$ via positive access; if the latter, E_i rules out $E = E_j$ via negative access. Since this is true for every $E_j \neq E_i$, the only remaining possibility compatible with the initial evidence, and thus with E_i, is $E = E_i$.

[45] Note that, in guaranteeing the equivalence of E_j^+ and $E^+ = E_j^+$, the access principle ensures

If the access principle holds in general, then, we would expect both (S-REF) and (F-REF) to be true. (EQEXP) trivially follows from their combination; and a directly analogous argument shows that (¬IBI) also holds. This settles one of the two questions motivating our more formal investigations: rather than being a weird case to focus on, counterexamples to the access principle are in fact the only possible cases for intentionally biased inquiry. Or, more cautiously, they are the only such cases compatible with the idealizing assumptions we have made explicit (finitely many evidence propositions, no possibility of information loss, no uncertainty about the evidential support relations) and those which are implicit in the formalism we have been employing (logical omniscience, no discovery of new possibilities, evidential support measured by exact values). Since these assumptions seem like adequate idealizations for modelling a large number of relatively 'ordinary' investigations, even this more cautious conclusion is still a significant result.[46]

To settle our second question, and generalize the argument from the impossibility of intentionally biased inquiry to the access principle, we need something like the converse entailment: that (EQEXP) fails unless the access principle is true. Unfortunately, the connection is not quite so straightforward. For (EQEXP) trivially holds of an agent who is certain not to receive any evidence, regardless of whether she satisfies the access principle.[47]

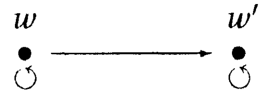
There is, however, a less straightforward connection that is strong enough for our purposes. Suppose that there are possibilities in which the access principle

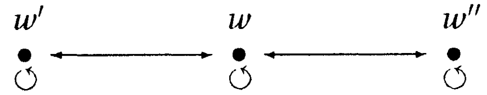
that our future evidence forms a partition of the (current) epistemic possibilities. That the future evidence forms such a partition is a standard assumption in attempts to derive reflection principles: see e.g. van Fraassen (1995, p.17) and Briggs (2009, p.69). Weisberg (2007, p.183-184) discusses the partitionality assumption in such proofs in detail, and concludes that it can only be motivated by something like the access principle.

[46] Kadane et al. (1996) show that principles like (EQEXP) may fail if we move to formal models which relax some of these assumptions; they leave open whether this is a problem for the principle or for the models, and I will too.

[47] In Salow (msf), I show that there is a good sense in which cases where the agent learns nothing are the only cases in which (EQEXP) holds despite failures of the access principle. However, the proof requires additional formalism, and so I will not appeal to the result here.

fails. That is, suppose there is some possibility w such that the total evidence we have in w does not allow us to rule out that we are instead in possibility w', in which we have different evidence. Then either the evidence in w' allows us to rule out that we are in w or there is a third possibility w'' which the evidence in one, but not the other, of w and w' allows us to rule out. In other words, the 'ruling out' relation between the possibilities will either fail to be symmetric or will fail to be transitive.[48] So, without loss of generality, we can assume that the 'ruling out' relation between the possibilities will exhibit one of the following structures:





These structures should look familiar: they are, respectively, the structure of the wall case and of the clock case (once we add to the clock case the background information that ensures that only three settings are possible).

But then it seems possible to imagine an agent whose initial evidence establishes all and only that he is in one of w, w', or w'', and who knows he is about to receive whatever evidence is associated with those possibilities. Then the evidential probabilities of our agent will violate (EQEXP). For the expected initial probabilities will match the initial probabilities (since in all of the possibilities, the agent's initial evidence establishes all and only that he is in one of w, w', or w'', so that there is no uncertainty about the initial evidence). And the expected future probability of w is lower than its initial probability in both cases, regardless of which (non-zero) initial probabilities are assigned to each world.

[48] Another way to see this is that the access principle, together with the claim that only truths can be evidence, implies that 'one's evidence entails that' obeys an S5 logic. And it's a well-known theorem of modal logic that a modal operator obeys an S5 logic if and only if the corresponding accessibility relation is reflexive, symmetric, and transitive.

That was, in fact, precisely why we earlier thought it to be desirable to somehow associate *w* with the possibilities in which I am unpopular.

This argument gives us a general recipe for converting counterexamples to the access principle into examples of cases where agents can bias their inquiries. In fact, it should now be clear that our discussion of the particular examples in §2 was simply an application of this general recipe. But the full force of our new defence of the access principle requires both this general argument and its specific application to cases. If we consider only the general argument, it may not be obvious why we shöuld resolve the tension between the denial of the access principle and the claim that intentionally biased inquiry is impossible in favour of the latter. And if we consider only the particular cases, it is natural to worry, as we did earlier, that the oddities we observe arise simply from idiosyncratic features of the particular example. But we have now seen both (i) that it really is the (supposed) violation of the access principle which makes it possible to use a case to intentionally bias one's inquiries and (ii) that it really is absurd, even in the cases which are the best candidates for such access violations, to think that intentionally biased inquiry is possible. This makes it hard to see a credible alternative to accepting the access principle.

4 Conclusion

We have covered a lot of territory. We began with the question of whether we can intentionally bias our own inquiries so as to favour one hypothesis over another. Our discussion suggested that the intuitive answer is 'no', at least once we have the relevant kind of biased inquiry clearly in view. This answer is particularly clear when we imagine trying to use such biased inquiry, for example to try to reassure ourselves of our own popularity. We then observed that certain popular counterexamples to the access principle would, if genuine, enable agents to bias their inquiries after all. But the relevant reasoning in those cases was clearly absurd; we should thus conclude not that such biasing is possible, but rather that we were wrong about the examples. Finally, we saw that the connection between the access principle and the possibility of intentionally biased inquiry is in fact both tight and perfectly general: formalizing the thought that we can't bias our inquiries, in a way closely related to the Reflection Principle, allows us to see that intentionally biased inquiry is possible if and only if the access principle is false. This connection, I suggested, both reinforces our argument that intentionally biased inquiry is in fact impossible and provides a powerful new reason to believe in access.

It is worth emphasizing that accepting the access principle is a radical conclusion. We have already encountered several reasons to reject the access principle when motivating the alleged counter-examples above. But let me add, in closing, what I think might be the best reason for denying access. This reason, nicely formulated by Weatherson (2011, p.451), adapts the obvious argument against 'negative introspection' for knowledge into a direct argument against the negative access principle. The argument has two premises: (i) rational agents can be mistaken about any (non-epistemological) subject matter and (ii) only truths can be evidence. Now let p be a proposition of the kind that could be evidence. Then a rational agent might mistakenly believe p, even though it is false, and thus fail to realize that p isn't part of her evidence. But if negative introspection were true, her evidence would entail that her evidence doesn't contain p, and so our agent's failure to realize that it doesn't would seem to be a failure of rationality (at least if she considers the question). Neither of the premises is undeniable,[49] but both are intuitively appealing.[50]

In addition to defending the view that intentionally biased inquiry is impossible and offering a novel argument in favour of the access principle, I hope to have offered a new perspective on Reflection-like Principles such as (F-REF) and (EQEXP). This shift might become clearest if we contrast our discussion of Reflection with Williamson's. For Williamson, in addition to being perhaps

[49] For example, Smithies (2012a) would deny (i) by maintaining that ideally rational agents would never be wrong about their phenomenal states, whilst Goldman (2009) seems to deny (ii).

[50] This is why I hold out hope for a way out along the lines hinted at in footnote 1.

the most prominent critic of the access principle, also discusses the Reflection Principle at some length, and my own discussion owes a lot to his. Despite this debt, we obviously disagree about whether Reflection is true. I think this is, at least in part, because we think of the principle quite differently.

Williamson presents Reflection as something that it would be nice to have, if only we could have it. The defender of Reflection, as Williamson sees him, is an overly enthusiastic optimist, who wants to reach ahead and make use of information he hasn't yet received. This allows Williamson to cast himself in the role of the cautious and sensible, if somewhat sombre, realist:

> But we cannot take advantage of the new knowledge in advance. We must cross that bridge when we come to it, and accept the consequences of our unfortunate epistemic situation with what composure we can find. Life is hard. (2000, p.237)

My own view of Reflection is less rosy than that of the opponent Williamson imagines. I have presented Reflection primarily as a limitation we face in the pursuit of our goals. From this perspective, it is the denier of Reflection who is overly optimistic, since he sees us as possessing tools for shaping our inquiries which we simply do not have. In my opinion, we should save our composure for facing up to this, much more unfortunate, realization.

Chapter 2

Lewis on Iterated Knowledge

Abstract

The status of the knowledge iteration principles in the account provided by Lewis in "Elusive Knowledge" is disputed. By distinguishing carefully between what in the account describes the contribution of the attributor's context and what describes the contribution of the subject's situation, we can resolve this dispute in favour of Holliday's (2015) claim that the iteration principles are rendered invalid. However, that is not the end of the story. For Lewis's account still predicts that counterexamples to the negative iteration principle ($\neg Kp \rightarrow K\neg Kp$) come out as elusive: such counterexamples can occur only in possibilities which the attributors of knowledge are ignoring. This consequence is more defensible than it might look at first sight.

One of the most influential versions of epistemic contextualism is the one Lewis develops in "Elusive Knowledge".[1] Despite its influence, this account is not always well understood. One place where matters are particularly unclear is the status of knowledge iteration principles in Lewis's account. Several authors (including Williamson (2001, 2009b), Holton (2003), and Greco (2014b), who all trace the claim to Lloyd Humberstone) maintain that Lewis's account validates an S5 epistemic logic, which would mean that it is committed to implausibly strong iteration principles for knowledge; by contrast, Holliday (2015) maintains that the knowledge iteration principles are invalid in Lewis's system.

[1] Lewis (1996). Blome-Tillman (2009, 2012, 2014) and Ichikawa (2011a,b, 2013) are recent defenders of (modified) versions of Lewis's account; commentators that pay close attention to Lewis's account in particular include Cohen (1998), Vogel (1999), Williams (2001), Williamson (2001), Schaffer (2004), Hawthorne (2004), Douven (2005), and Dutant (ms).

By distinguishing carefully between what is contributed by the conversational context of the agents attributing knowledge and what is contributed by the situation of the subject to whom knowledge is attributed, we can resolve this dispute in Holliday's favour: Lewis's system allows counterexamples to both the KK principle (that whenever someone knows something, they know that they know it) and what I will call the $K\neg K$ principle (that whenever someone doesn't know something, they know that they don't know it). However, we can also see that this is not the end of the story: counterexamples to the $K\neg K$ principle can only occur at worlds that the attributors of knowledge are ignoring. (No analogous result holds for the KK principle.) On the face of it, this surprising consequence of Lewis's account looks almost as implausible as the claim that the $K\neg K$ principle is valid. However, I will argue that there are ways of rendering the consequence acceptable.[2] Throughout the paper, I will try to draw more general lessons about the relationship between epistemic contextualism and the knowledge iteration principle, explaining why their interaction is both subtle and fruitful.

1 Lewis, Formalized

Discussions of epistemic logic standardly proceed in a possible worlds framework, in which an agent X is said to know p at w if and only if every world accessible from w (under the accessibility relation associated with X) is a p-world. Lewis seems to proceed similarly. Consider, for example, his well-known summary of the account:

> X knows that P iff X's evidence eliminates every possibility in which not-P – Psst! – except for those possibilities that we [attributors] are properly ignoring. (1996, p.554)

[2] I actually think that, in addition to it not being obviously false, there are positive reasons to want something like the Lewisian treatment of $K\neg K$ to be correct. For, as I argue in Salow (msa), it allows us to solve hard problems for the (thoroughly non-Lewisian) thesis, defended by Williamson (2000), that one's evidence consists of all and only the claims that one knows.

This seems to translate quite straightforwardly into the traditional framework: we simply say that a world is accessible if it is neither properly ignored nor ruled out by X's evidence.[3] One would thus expect it to be relatively straightforward to distil a logic from Lewis's account. However, as we will see shortly, there are some pitfalls here to be navigated.

To proceed with the approach just sketched, it is natural to look to 'frames' that consist of a set of worlds W, together with a specification of how Lewis's primitives behave at the various worlds; we can then see what happens when we define accessibility in terms of these primitives. Deciding on how to represent the primitives, however, requires some care. For Lewis's theory is, above all, a *contextualist* theory. This means that whether an attribution of knowledge correctly describes a situation depends on both features of the situation described and features of the context from which the attribution was made. However, only the features of the situation (the 'world of evaluation') will vary as we consider what an agent knows in different possible worlds; we are interested in the logic of 'knows' within a single context, and so whatever is supplied by context will remain fixed. Our frames thus need to represent the features of the situation as world-relative, but can represent the contributions of the context absolutely. Whether something is a feature of the situation described or of the context of ascription thus matters greatly to how our frames should represent it.

1.1 A Natural Mistake

How does this distinction between features of the context and features of the situation described apply to Lewis's account? The above summary of the account suggests that the correctness of knowledge attributions depends on two components: (i) what evidence the subject has, which we can represent by a relation E so that wEv iff v is compatible with the evidence X has in w, and

[3] This is not quite right as an interpretation of Lewis, since he uses 'possibilities' to mean something slightly different from possible worlds (1996, p.552). To keep the formalization of his account manageable, I ignore that complication here.

(ii) a set S of possibilities that are not being properly ignored. The first of these is clearly a feature of the situation described; the second looks, at least at first sight, like a feature of the context – that's why it seemed natural to represent it absolutely, i.e. as a set rather than a function from possibilities to the set of worlds ignored at that possibility.

We will see shortly that this approach isn't textually plausible. Nonetheless, it is worth briefly exploring it, since it helps explain the appeal of the idea that Lewis's account vindicates an S5 logic. For the current proposal would see Lewis vindicate the iteration principles. Lewis views a subject's 'evidence' as her total phenomenal state, so that wEv if and only if the subject is in the same total phenomenal state in w and v; this makes E an equivalence relation. The obvious definition of R_K, the accessibility relation for our subject's knowledge, holds that wR_Kv if and only if wEv and $v \in S$, so that an agent knows p only if her evidence eliminates all the unignored p-worlds. And on this definition, R_K will be transitive and Euclidean.[4] We thus validate both the KK principle ($Kp \rightarrow KKp$) and the $K\neg K$ principle ($\neg Kp \rightarrow K\neg Kp$).

However, we don't quite vindicate a full S5 logic. The missing principle is the most basic one: that what is known must be true. For note that no world outside of S will be accessible to *any* world under R_K, not even to itself. R_K thus isn't reflexive, and so we do not validate the T principle ($Kp \rightarrow p$); in worlds outside S, people can know things that aren't true there. This is a clear sign that something has gone wrong; the factivity of knowledge is not only epistemologically non-negotiable, but also a feature Lewis (1996, p.554) specifically intended his account to vindicate.

We run into this problem with factivity because our logic is sensitive to how knowledge behaves in possibilities that are properly ignored. Since Lewis (1996, p.555-559) explains that such possibilities are neither actual nor salient, this

[4] To see that it's transitive, note that from xR_Ky and yR_Kz it follows that $z \in S$ and xEy and yEz. So $z \in S$ and xEz (since E is transitive), and so xR_Kz. To see that it's Euclidean, note that if xR_Ky and xR_Kz, then $z \in S$ and xEy and xEz. So $z \in S$ and yEz (since E is euclidean), and hence yR_Kz.

sensitivity might seem excessive.[5] It can be avoided by redefining validity as truth at every not-properly-ignored-world in every model;[6] this would, in fact, allow us to vindicate a full S5 logic.[7] However, R_K won't be reflexive even on this revised approach, suggesting that the original problem has been hidden rather than solved. One way to bring this out is by considering what happens when we introduce other modal operators. For suppose we introduce an operator $\Box$ for metaphysical necessity. It seems plausible that some worlds outside S are metaphysically possible with respect to some worlds in S in at least some models. But then $\Box(Kp \rightarrow p)$ will not be a principle of the combined logic of knowledge and metaphysical necessity. This strikes me as no less serious than the original problem of allowing for *actual* factivity failures.

Our simple-minded approach, whilst hospitable to the iteration principles, thus has consequences which are both extremely unattractive and difficult to eliminate. The culprit seems to be the fact that the set of relevant possibilities that need to be eliminated is treated as something entirely supplied by context. For this means that the relevant possibilities cannot vary when we evaluate a knowledge attribution at different worlds; but this, in turn, implies that some possibilities aren't relevant to themselves, so that agents at those possibilities can eliminate all relevant ¬p-worlds (and thus know p) even though p is false.

[5] In assessing what he calls a 'rigid' interpretation of Lewis, Dutant (ms) first points out that this interpretation struggles with the factivity of knowledge, and then considers a response analogous to this one. He observes that, even once we acknowledge such a response, the interpretation still predicts that the sentence 'someone could have known something false' could be true, which is the inspiration for the objection I offer below.

[6] A variant of this is more familiar in modal logic. We could move to 'model structures' $< W, E, S, w >$ which designate world $w \in W$ as the actual world. Since the actual world is never properly ignored, we would then want to impose the structural requirement that $w \in S$. When working with model structures instead of frames, it's also natural to redefine validity as truth at the designated world of every model. The resulting system is very similar to the one discussed in the main text; in particular, it validates S5 for essentially the same reason.

[7] Why? Let us say that v can be reached from w if there are worlds $u_1, \ldots u_n$ such that $wR_K u_1, u_1 R_K u_2, \ldots u_n R_K v$. Then truth in a model depends only on what happens in worlds that are either in S or can be reached from a world in S. Moreover, the definition of R_K ensures that all such worlds are themselves in S. Finally, R_K is an equivalence relation when restricted to S (though not outside it). Together, these facts ensure that we validate an S5 logic.

We thus fail to capture the factivity of knowledge.

1.2 Doing Better

Fortunately, Lewis's discussion does not commit him to such an inadequate account. It is true that which possibilities are being *ignored* is settled by the context. But Lewis defines knowledge in terms of *proper ignoring*, and it is far from obvious that it is the context which settles which ignorings are proper. In fact, when Lewis, in introducing the 'Rule of Actuality', explicitly discusses this issue, he asserts that propriety is (at least partially) determined by the world of evaluation:

> The possibility that actually obtains is never properly ignored. ... Whose actuality? Ours, when we ascribe knowledge or ignorance to others? Or the subject's? ... [T]he right answer is that it is the subject's actuality, not the ascriber's, that never can be properly ignored. (1996, p.554f)

"The subject's actuality" seems to be the world of evaluation;[8] so what can be properly ignored depends on what the world of evaluation is. We therefore need to reinterpret *S* to represent only what is contributed by the context. Plausibly, that is the set of worlds that are not *in fact* ignored by the attributors; this set will thus leave out worlds that are ignored but only improperly so. This is how '*S*' will be interpreted from here on in. In addition to this reinterpretation, we need to enrich our frames to represent directly all the features of the worlds that constrain what can be properly ignored relative to each of them.

[8] Dutant (ms) argues that "the subject's actuality" might be construed instead as the (potentially counterfactual) world on which the conversation is focused; this would allow for context alone to determine propriety. I agree that such a reading is just about possible. But since it would leave us with the unsatisfactory account discussed in §1.1, and the context of the passage strongly suggests that Lewis is trying to rule out this variant account, I think it safe to assume that this is not how Lewis intended these remarks.

What features are these? Lewis articulates the limits of proper ignoring by appeal to the Rules of Actuality, Belief, and Resemblance.[9] The information relevant to the Rule of Actuality is trivially represented in the frame, since every world is actual relative to itself. So the first addition is the notion of the subject's beliefs,[10] which we will need to implement the 'Rule of Belief' stating that "a possibility that the subject believes to obtain is not properly ignored" (1996, p.555f). Following the standard formalization of belief, we can represent this by an accessibility relation R_B on worlds, where wR_Bv is understood as 'v is consistent with all of X's beliefs in w.'

The second addition required to constrain proper ignoring is that of relevant similarity, which we will need to implement the 'Rule of Resemblance':

> Suppose one possibility saliently resembles another. Then if one of them may not be properly ignored [in virtue of rules other than this rule], neither may the other. (1996, p.556)

Since it is context, rather than the world of evaluation, which determines which respects of similarity are salient, this can be represented by a binary relation 'C' (for 'closeness') with wCv read as 'w is close to/relevantly resembles v'. Crucially, we may not assume that C is transitive, since Lewis is at pains to distinguish between worlds resembling each other and worlds being connected by a chain of resembling worlds.

A full Lewisian frame is thus a 5-tuple $< W, E, S, R_B, C >$; such a frame does better at representing the information needed for an adequate formalization.

[9] What is the role of the 'permissive' rules, such as the Rules of Reliability, Method, and Conservatism (1996, pp.558-559)? I have to confess to finding these rather puzzling. As I understand Lewis, any world that isn't being attended to is automatically ignored, and thus properly ignored if no 'restrictive' rule prevents this from happening. But then what role could there be for the permissive rules to play? One hypothesis is that they aren't *rules* about the propriety of ignoring at all, but are rather *empirical generalizations* about what kind of worlds are in fact ignored in ordinary contexts. Another thought, suggested to me by Bob Stalnaker, is that they function as constraints on what 'restrictive' rules Lewis would be willing to add to his account: they had better be consistent with it being proper, except in very specific circumstances, to ignore worlds in which our faculties and methods are unreliable.

[10] Or what the agent should believe, but I will set that complication aside.

For we can now define proper ignoring in a way which ensures that different possibilities are properly ignored relative to different worlds of evaluation. According to Lewis, the worlds not properly ignored relative to w are (i) w itself (to respect the Rule of Actuality), (ii) the worlds consistent with X's beliefs at w (to respect the Rule of Belief)[11] (iii) the salient worlds S (to respect the Rule of Attention), and (iv) any world close to those mentioned in (i)-(iii) (to respect the Rule of Resemblance).

We formalize this thought by defining an 'alternatives' function $A : W \rightarrow \mathscr{P}(W)$, which takes each world w to its alternatives, i.e. the possibilities not properly ignored relative to w. We first implement (i)-(iii) to define an impoverished function A^-, and then 'fill it in' to define an A which also respects (iv):

$$A^-(w) =_{def} \{w\} \cup \{v : wR_B v\} \cup S$$
$$A(w) =_{def} \{u : \exists v \in A^-(w) \text{ s.t. } uCv\}$$

We then use A together with E to define the accessibility relation for knowledge R_K in the natural way: for all worlds u and v,

$$uR_K v \text{ if and only if } uEv \text{ and } v \in A(u).$$

The resulting system is essentially a special case of Holliday's (2015) formalization of Lewis.[12] Simplifying slightly, Holliday's frames are, in our notation,

[11] Given the above statement of the rule of belief, one might worry that this is much too strong: there, Lewis seems to say that a possibility believed to obtain isn't properly ignored, not that a possibility not believed not to obtain isn't properly ignored. But Lewis later clarifies that what he really means is that "a possibility may not be properly ignored if the subject gives it [...] a degree of belief that is sufficiently high," (1996, p.556) and context makes clear that "sufficiently high" is usually far below .5 (as it has to be, since otherwise almost no reasonable agent will have a "sufficiently high" degree of belief in any single possibility). So 'the worlds consistent with X's beliefs' is a better approximation of Lewis's rule than 'the world (if there is one) uniquely consistent with X's beliefs.' It is nonetheless merely an approximation of what Lewis was after; one consequence of this choice will be that, contrary to Lewis's (1996, p.556) explicit intentions, our formalization will not allow for knowledge without belief in cases like that of the reliable but underconfident examinee.

[12] Thanks to an anonymous referee for extremely helpful discussion on this point.

the triples $< W, E, A >$; the rule of actuality is built in by requiring that $w \in A(w)$. Our models are less general, because defining A in terms of S, R_B, and C imposes additional constraints.[13] Formally, this lesser generality will generate the surprising new result discussed §2; and at an informal level, I hope that building up A in the way I have done (and making explicit the rival approach discussed in §1.1) helps clarify why this really is the right way to formalize Lewis.

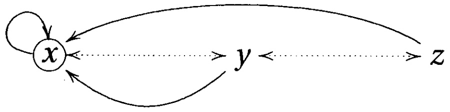
What, then, are the formal features of this system? Unlike the first attempt, it has no trouble accounting for the factivity of knowledge. And the way this account implements the rule of belief means that we *almost* validate the principle that everything known is believed.

(But only almost: as Ichikawa (2011a, p.386) points out, Lewis's account implies that if a proposition p is entailed by an agent's evidence, she automatically knows p, regardless of whether she believes it. In fact, she can know p whilst believing its negation: while $A(w)$ will contain the $\neg p$-worlds compatible with the subject's beliefs, those will then be ruled out by her evidence, and thus no longer accessible under R_K. This is a bad result even if, like Lewis (1996, p.556), we think that the connection between knowledge and belief is rather loose. But it seems to me an unavoidable feature of Lewis's thought that we know everything that is true in all the possibilities compatible with our evidence. Of course, we can reject this thought to preserve the link between belief and knowledge, e.g. by replacing E with $E \cup R_B$ in the definition of R_K. Alternatively, we can hold onto the Lewisian thought (and hence the original definition), and simply admit that, in so doing, we are restricting our attention to somewhat idealized agents who believe everything their evidence entails.[14] Since our models, as is standard, already build in a variety of similar idealizations, such as the assumption that agents always know and believe logical consequences of what they know and believe, I will opt for this simpler approach.)

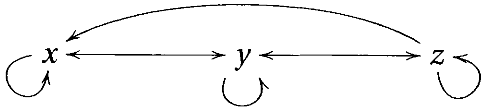
[13] Though Holliday (2013) considers imposing the constraint corresponding to the rule of belief.

[14] That is, we require that, in all our models, $wR_B v$ entails wEv. Cf Holliday (2013).

As Holliday points out, however, this system does not provide a hospitable environment for the iteration principles. For consider the three world model on which (a) x is the only salient world, (b) x is the only world compatible with our agent's beliefs in any of the three worlds, (c) x resembles y and y resembles z but x does not resemble z, and (d) our agent's evidence at each of the worlds is compatible with her inhabiting any of them. These facts can be visually represented as follows, with continuous lines standing for R_B, dotted lines standing for C, and worlds in S occurring inside the circle (information about E, being trivial, is omitted):



Then, under R_K, x will access only itself and y, whilst y and z both access all three worlds, as in the following diagram:



Now let p be a claim that is true in x and y, but false in z. Since p is true at both x and y, Kp will be true at x; but since p is false at z, Kp will be false at y. So KKp will fail at x even though Kp was true there, and so we have a counterexample to the KK principle. The same model also provides a counterexample to the $K\neg K$ principle, since $\neg Kp$ will be true at y and z but $K\neg Kp$ will be false at both.[15]

It is worth noting that these results are independently attractive. The $K\neg K$ principle in particular seems clearly invalid: someone who reasonably believes something false fails to know but doesn't know (and needn't be in a position to know) that he so fails. And Lewis seems to be trying to do justice to this thought. Thus he (1996, p.554) describes his account as "'externalist' – the subject himself may not be able to tell what is properly ignored." But this is inconsistent with

[15] It's worth noting that, while the counterexample to KK relies on the intransitivity of C, the counterexample to $K\neg K$ does not. For we can simply drop y from the example, rendering C irrelevant; the resulting model will validate KK, but $K\neg K$ will still fail at z.

the iteration principles, since the subject could use his knowledge of what he knows to work back to what is being properly ignored.[16]

2 Elusive $K\neg K$

Getting clear on whether the iteration principles are valid in Lewis's system matters if we are interested in what Lewis thought. It also matters if we want to appeal to their status in Lewis's system either to bolster the plausibility of a principle (as Greco (2014b) does in appealing to the claim that Lewis's system vindicates the KK principle) or to criticize Lewis's account (as Williamson (2001) does, in saying that Lewis's system vindicates the $K\neg K$ principle). But there is also a more surprising reason for noting that Lewis's account does not, in fact, validate the iteration principles: the $K\neg K$ principle turns out to have a different but still unusual status in this system.

For suppose we may assume that, on any interpretation of 'knows', the agent in question always knows what her beliefs are.[17] Then we can show that the $K\neg K$ principle has no counterexamples in any of the worlds that are in fact salient to the attributors:[18]

Elusive K¬K. *For any $w \in S$ and proposition p, $\neg Kp \rightarrow K\neg Kp$ is true at w.*

[16] Moreover, Bob Stalnaker tells me that, while Lewis initially thought that his theory should satisfy an S5 logic, he became convinced of the implausibility of the $K\neg K$ principle whilst presenting early versions of "Elusive Knowledge". This change of heart coincided with the introduction of his extended discussion of the Rule of Actuality, and we saw earlier that this is the crucial passage warning us against the iteration-friendly formalization of §1.1.

[17] Formally: $\forall x \forall y (xR_K y \rightarrow \forall z (xR_B z \leftrightarrow yR_B z))$. Given Lewis's account, this claim can be true on every interpretation of 'knows' only if a difference in beliefs always makes for a difference in phenomenal state; Smithies (2014) develops a notion of 'phenomenal state' designed to have this feature, and argues that one's justification supervenes on what phenomenal state (in this sense) one is in, so this might be a way of incorporating the introspection assumption into a broadly Lewisian account. It's also worth noting that, even if we deny that agents in general always know what they believe, it is still interesting and surprising that the Lewisian account predicts our result to hold of those that do.

[18] Recall that the actual world may not be salient to the attributors; the result thus doesn't entail that the $K\neg K$ principle will be true.

Proof. Suppose that $\neg Kp$ is true at $w \in S$. Then there must be some v at which p is false such that $wR_K v$, which implies $v \in A(w)$. Now let u be any world such that $wR_K u$. We will begin by showing that $v \in A(u)$: we argue that, since $v \in A(w)$, one of four conditions must hold, and that any of these are sufficient to ensure that $v \in A(u)$.

(i) $v \in A^-(w)$ because $v = w$. Since $w \in S$, this ensures that $v \in A^-(u)$.

(ii) $v \in A^-(w)$ because $wR_B v$. Since $wR_K u$, it follows from our introspection assumption that $uR_B v$ also. So $v \in A^-(u)$.

(iii) $v \in A^-(w)$ because $v \in S$. Then $v \in A^-(u)$ also.

(iv) $v \in A(w)$ but $v \notin A^-(w)$. Then there must be an $x \in A^-(w)$ such that vCx. But, then x must meet one of conditions (i)-(iii), and so $x \in A^-(u)$. So $v \in A(u)$ also.

So $v \in A(u)$. Since $wR_K v$ and $wR_K u$, we have wEv and wEu, which implies uEv since E is an equivalence relation. So $uR_K v$. So Kp is false at u also. Since u was an arbitrary world satisfying $wR_K u$, it follows that $K\neg Kp$ is true at w. Since w was an arbitrary member of S and p an arbitrary proposition, this establishes the result. □

This result is extremely surprising; it seems to say that we can never attend to agents who are unaware of the fact that they fail to know something, that counterexamples to the $K\neg K$ principle are elusive. That sounds obviously false: the $K\neg K$ principle isn't just invalid, but subject to clear counterexamples which we have no trouble thinking about. I will argue shortly that things may not be quite so straightforward; but first, we should attempt to understand why Lewis's account has this kind of consequence.

Counterexamples to $K\neg K$ seem easy to come by: just pick an agent who has a belief which, while it looks good 'from the inside', falls short of knowledge because of an uncooperative environment. To have a concrete example, consider someone whose belief that the wall in front of her is red falls short of knowledge because the lighting is unreliable. Since the belief 'looks good from the inside' our agent must have evidence that rules out the kind of $\neg p$–possibilities that any would-be knower has to rule out, such as possibilities in which the wall is and looks yellow. Since, nonetheless, her belief doesn't amount to knowledge, there must be some other, more idiosyncratic, $\neg p$–possibilities, that are relevant to her because her actual environment is uncooperative, and which her evidence doesn't eliminate; in our example, these would be possibilities in which the wall is white but the lighting is misleading. (These possibilities might be either actual, or relevantly similar to the one that is actual; it doesn't matter which.) But now suppose, for *reductio*, that our agent's actual circumstances are salient. Then, according to Lewis, we will use 'knowledge' in such a way that *anyone* has to rule out these supposedly 'idiosyncratic' possibilities to count as knowing by the standards of the current conversation; for, by the rules of attention and resemblance, any would-be knower has to rule every possibility which is either salient or relevantly similar to one that is salient. And so the error possibilities cannot be idiosyncratic to our subject after all, contradicting our assumption. So, if a case like that of misleading lighting is salient, it cannot, after all, be a case in which our agent fails to know without knowing that she fails.

What is generating the result is thus the feature of contextualism that was also responsible for the weird consequences of the naive formalization in §1.1: that something contributed by the conversational context (now: the set S of salient possibilities) is independent of the world of evaluation. This means that, once a possibility (such as the possibility of misleading lighting) is in S, *any* would-be knower has to eliminate it, regardless of what his or her world is like. Since rational $K\neg K$ failures intuitively arise from error possibilities that are specific to the subject who fails to know, this has the surprising consequence that the error possibilities generating the counterexample to $K\neg K$ must not themselves be salient (or relevantly similar to possibilities that are salient.) And that's just another way of saying that counterexamples to $K\neg K$ occur only in possibilities that aren't in S.

Interestingly, we get no analogue of **Elusive K¬K** for the KK-principle; in fact, the model described in §1.2 already showed that KK can fail even at a salient possibility. This reveals quite how different the counterexamples to these two principles are on the Lewisian treatment. KK fails because C isn't transitive: someone's evidence can rule out all the worlds resembling the actual one, without thereby ruling out all the worlds resembling some world that resembles the actual one. By contrast, $K\neg K$ fails because agents sometimes reasonably think they can ignore possibilities which, because of facts specific to their actual situation, turn out to be relevant. Making the actual world salient, and thereby forcing it to be relevant no matter what, prevents the second of these but leaves the first untouched.

Now that we understand a little better why Lewis's account entails **Elusive K¬K**, we can turn to examine whether this is problematic. At first sight, it seems terrible. We *can* describe clear and concrete counterexamples to the $K\neg K$ principle; and **Elusive K¬K** seems to predict that we can't. But matters are not quite so straightforward. In §2.1 and §2.2 I will describe two ways in which Lewisians can respond. The first yields no ground at all, and argues that we can still do justice to the clear examples; the second is more conciliatory, taking **Elusive K¬K** to motivate a different conception of what it is to 'ignore' a possibility.[19] Each, I think, has promise; so the fact that Lewis's account entails **Elusive K¬K** doesn't refute that account.

2.1 The Hard-Nosed Response

The Lewisian who wants to yield no ground has his work cut out for him. There are two natural ways of understanding Lewis's talk of 'ignoring'; and the prediction that $K\neg K$ failures happen only in ignored possibilities looks implausible on either one. The first way of understanding 'ignore' is more prominent in Lewis's discussion: a possibility isn't ignored if it is psychologically salient, if we

[19] A more radically conciliatory response would give up on the thought that worlds that aren't ignored always need to be eliminated. To preserve any of the Lewisian spirit, we would then have to offer a different account of the role S plays in defining A^- or A. Dutant (ms) discusses some interesting attempts along such lines, though he finds them all wanting.

are thinking or talking about it. But sometimes Lewis instead writes of which possibilities are compatible with our presuppositions; or, as I shall put it, which possibilities we take seriously. And, as Blome-Tillman (2009, 2014) emphasizes, what is salient and what is taken seriously need not coincide. I tell you that the wall in the seminar room is red. You raise the worry that the lighting might have been misleading. When I discover that you have no special reason to think so, I tell you to stop being so tedious. Even though you have made the possibility of misleading lighting salient, I refuse to take it seriously and continue to presuppose that it does not obtain.

There are a number of independent reasons why understanding 'ignore' in terms of presuppositions is more attractive than understanding it in terms of salience.[20] To these we can add that this way of understanding 'ignore' helps reconcile **Elusive K¬K** with the possibility of clear counterexamples to $K\neg K$ when these counterexamples are thought of hypothetically. I claim to know that the wall is red. I agree that it's not impossible for the lighting to be unreliable and that, if it had been unreliable, my belief that the wall is red would have fallen short of knowledge without my knowing that it did. Perhaps I even agree that if, contra everything I believe, the lighting was unreliable this time, my actual belief falls short of knowledge even though I do not know that it does. But I continue to presuppose that the antecedents of these conditionals are false. So that speech is no counterexample to **Elusive K¬K** (when 'ignore' is understood as 'don't take seriously') since the possibility in which I locate the counterexamples to $K\neg K$, being inconsistent with my presuppositions, isn't in S.

However, there are also clear counterexamples to $K\neg K$ that needn't be described hypothetically; these are most naturally described as cases in which the subject differs from the attributors. My friend Soraya says that the wall in the other room is red. But we know that the lighting in that room is unreliable. So it seems that we can rightly judge that Soraya fails to know but doesn't know that she so fails. After all, we know that (i) her belief, being formed in poor conditions,

[20] See Hawthorne (2004) and Blome-Tillman (2009, 2014) for discussion.

can't be knowledge, and (ii) she doesn't (and has no reason to) suspect, much less believe, that she doesn't know. In fact, she seems to think that she does know – otherwise she wouldn't have felt so confident in telling me the color of the wall. But her case is both salient to us and compatible with our presuppositions, since we believe it to be actual. Doesn't that refute **Elusive K¬K**?

Perhaps not. It does seem clear that we can judge that Soraya doesn't know but doesn't know she doesn't know. But it isn't clear that 'know' is interpreted relative to the possibilities salient to *us* throughout that judgement; and if it's not, the possibility of this judgement needn't conflict with the Lewisian result. For **Elusive K¬K** entails only that knowledge-relative-to-S behaves in line with $K\neg K$ throughout S; it makes no predictions about the behaviour of knowledge-relative-to-S', nor about principles which mix different interpretations of 'know'.

On Lewis's account, which relation is picked out by 'knows' depends on what possibilities are salient to, or taken seriously by, the speakers. In our example, Soraya is not, I assume, taking seriously the possibility that the lighting is odd – if she did take that possibility seriously, she wouldn't take herself to know that the wall is red. There are thus two senses of 'know' in play in the situation; since it takes more to know in our sense than in Soraya's, I will use 'know$_{hi}$' to name the relation 'know' refers to when the contextual parameter is filled with the possibilities *we* take seriously, and 'know$_{lo}$' for the relation it refers to when the contextual parameter is filled with the possibilities *Soraya* takes seriously.[21] **Elusive K¬K** then entails only that if Soraya doesn't know$_{hi}$, she knows$_{hi}$ that she doesn't know$_{hi}$; and I will show that the Lewisian has principled reason to deny that this conflicts with our intuitive judgement that Soraya fails to know without knowing that she does.

Let us begin by looking at what Soraya knows or believes about what she knows$_{lo}$ and knows$_{hi}$ about the wall. It seems pretty clear that she believes that she knows$_{lo}$ that the wall is red. That belief is why Soraya is inclined to say that

[21] This may be a little misleading, since, as I argue later, it's not very intuitive to think that our *standards* for knowledge are higher than Soraya's, which is what the notation suggests.

the wall is red, and that she knows this, when talking with people that she takes to share her epistemic standards.[22] The belief is false, since the fact that the lighting is actually unreliable means that Soraya has to rule out possibilities with misleading lighting even to know$_{lo}$. In spite of being false, however, the belief is perfectly reasonable: had the environment been more cooperative, Soraya wouldn't have had to rule out possibilities with misleading lighting to know$_{lo}$; and Soraya has no reason to suspect the lack of cooperation.

A belief that she knows$_{hi}$ that the wall is red is quite a different matter. After all, it's clear from the meaning of 'know$_{hi}$' that one doesn't know$_{hi}$ that the wall is red unless one can rule out the possibility of misleading lighting, no matter how dissimilar such worlds are from the actual situation. And Soraya can tell that she is in no position to rule out possibilities with misleading lighting. A belief that she knows$_{hi}$ that the wall is red would thus be a highly unreasonable belief for her to have; and since Soraya (like all subjects satisfying the idealizations implicit in our reconstruction of Lewis) is rational, she doesn't have such unreasonable beliefs.

This last point can be strengthened. Since it is clear to Soraya that she can't rule out possibilities in which the lighting is misleading, she is well aware that she *doesn't* know$_{hi}$ that the wall is red. Or, at least, she is aware of this if she has ever thought about what she knows$_{hi}$ at all; and, in keeping with our Lewisian idealizations, we shall assume that she has.[23] So we have that Soraya believes

[22] In saying this, we can be neutral on whether this is the belief expressed by her utterance, as it might not be if her conversational partners do not, in fact, take the same things seriously as she does. See DeRose (2004) for discussion.

[23] One might worry that this is in tension with our stipulation that Soraya is ignoring the possibility of misleading lighting; for if she is, how could she even articulate what it takes to know$_{hi}$? If 'ignoring' is understood in terms of presuppositions, the worry is easily dissolved, since Soraya can think about the possibilities of misleading lighting when determining what she knows$_{hi}$ without taking them seriously; that is, presumably, what most contextualists do when they agree that they know very little by sceptical standards. If 'ignoring' is understood in terms of salience, the worry has more bite; but we can still imagine that Soraya reflected earlier about what she would know$_{hi}$ in various situation, and that those earlier beliefs, which do not feature amongst her conscious thoughts when she is looking at the wall, are sufficient to constitute a belief that she does not know$_{hi}$ that the wall is red.

that she doesn't know$_{hi}$ that the wall is red, and that this belief (being based purely on introspection into her evidence and a priori reasoning) amounts to knowledge in every relevant sense.

(At this point, it might start to seem as though our idealizing assumption – that Soraya's beliefs are consistent and include everything entailed by her evidence – is pulling a lot of weight. But it would, I think, be a mistake to blame the surprising **Elusive K¬K** on the strength of these idealizations. For we also want to say that Soraya's case is one in which she fails to know but is in no position to know that she so fails. Yet, even if Soraya were less ideal than we have been assuming, the above considerations would still suggest that she is at least in a position to know that she doesn't know$_{hi}$ that the wall is red.)

Here, then, are the natural predictions of the Lewisian account:

(a) Soraya doesn't know$_{lo}$ that the wall is red.

(b) Soraya believes that she knows$_{lo}$ that the wall is red.

(c) Soraya doesn't believe/know$_{lo}$/know$_{hi}$ that she doesn't know$_{lo}$ that the wall is red.

(d) Soraya doesn't know$_{hi}$ that the wall is red.

(e) Soraya does not believe that she knows$_{hi}$ that the wall is red.

(f) Soraya believes/knows$_{lo}$/knows$_{hi}$ that she doesn't know$_{hi}$ that the wall is red.

Do these allow us to recover the obvious natural language judgements, such as 'Soraya thinks she knows that the wall is red'? They do, if we combine them with a surprising claim about how the context-sensitivity of 'know' is resolved when the word occurs embedded in an attitude ascription. For in order to get the obvious judgement to come out true, we have to say that 'know', when embedded under 'Soraya thinks that', means know$_{lo}$ – even when said by us, with our high standards. More generally, we have to say that when 'know' is embedded in an attitude ascription, the contextual parameter relative to which it is interpreted is supplied not by the context of utterance, but by something like the private context of the subject of the attitude ascription.

I will revisit the plausibility of this linguistic claim shortly. For now, we should simply note that, if it is correct, it also reconciles our example with **Elusive K¬K**. It is natural for us to judge that, even though Soraya doesn't know that the wall is red, she doesn't know that she doesn't know this; this seems to be in tension with **Elusive K¬K** because we are attending to and taking seriously Soraya's situation. However, if the above linguistic claim is correct, the tension is illusory. For our judgement then amounts to the observation that Soraya doesn't know$_{hi}$ that she doesn't know$_{lo}$ that the wall is red. And the Lewisian description of the situation vindicates that judgement: Soraya has no reason to suspect that she doesn't know$_{lo}$ that the wall is red. **Elusive K¬K** entails only that Soraya knows$_{hi}$ that she doesn't know$_{hi}$ that the wall is red. And, as we saw above, that is actually a plausible thing to say about the situation.

This reconciliation relies on a linguistic hypothesis: that when 'know' is embedded in an attitude ascription, the contextual parameter relative to which it is interpreted is supplied not by the context of utterance, but by something like the private context of the subject of the attitude ascription. If this were a feature not shared by other context-sensitive vocabulary, this would be an implausible consequence of the Lewisian account. But, fortunately for the Lewisian, there is independent reason to think that this kind of behaviour is actually quite common. For consider two other expressions which are naturally treated as context-sensitive: 'fun' and 'might'. It looks as though, usually, the contextual parameter (a standard of taste or evaluation, a body of information) is provided by the context of utterance: when we say that something is fun, we mean that it is fun *for us*, and when we say that something might be true, we mean that its truth is compatible with the information available *to us*. However, when these expressions are embedded in belief attributions, this natural treatment seems to go wrong. Consider:

(1) Soraya thinks that roller-coasters are fun.

(2) Soraya thinks that it might be raining in Abidjan.

Intuitively, (1) is true whenever Soraya thinks that roller-coasters are fun *for* *her*; she might be well aware that we abhor them, so that 'Soraya thinks that roller-coasters are fun for us' is definitely false. Similarly, (2) is true even when Soraya knows that we are better informed about the weather in Abidjan than she is, and thus suspends judgement on whether, for all you and I know, it might be raining in Abidjan. This suggests that, when they occur embedded in attitude ascriptions, the parameter for these expressions is usually supplied not by the context of utterance but by a derived context which is particularly sensitive to the subject of the embedding verb. And that is exactly the same as what our Lewisian wants to say about 'know'.[24]

It's worth emphasizing that this line of reasoning cannot be used to defend the stronger claim that the $K\neg K$ principle is valid. Our reasoning shows that the example described needn't be a counterexample to the claim that, if someone doesn't know$_{hi}$ that p, they know$_{hi}$ that they don't know$_{hi}$ that p. But the case is a genuine counterexample to the claim that, if someone doesn't know$_{lo}$ that p, they know$_{lo}$ that they don't know$_{lo}$ that p. For, in the case described, Soraya doesn't know$_{lo}$ that the wall is red – there are worlds that relevantly resemble the actual

[24] The thought that context-sensitive expressions embedded in attitude ascriptions are not simply interpreted relative to the context of utterance is quite familiar; see e.g. Stalnaker (1988) for a classic articulation and defence. It is frequently applied by contextualists to handle embeddings under 'says that' or 'believes that'; see e.g. Cappelen and Hawthorne (2009).
This strategy does face an important challenge with embeddings under factive attitude verbs such as 'knows' (cf Weatherson (2008), Lasersohn (2009, p.369-372), and Yalcin (2012) for related discussion). For it seems to predict that we could say 'Soraya knows that roller-coasters are fun', even though we hate them (provided only that we think that Soraya loves them and knows that she does), which is clearly incorrect. We thus need to supplement the simple shifting story with a, perhaps pragmatic, account of why knowledge ascriptions seem to entail the proposition which their complement would have expressed had it not been embedded. (Silk (ms) sketches such an account.) But note that simply denying that embedding under 'knows' (unlike embedding under 'believes') shifts the parameter is also implausible. For we can say 'Soraya knows roller-coasters are fun' even if we know that she (falsely) believes that we hate them.
A less optimistic reaction to these problems is to conclude that they sink contextualism about such terms as 'fun' or 'might', and should push us towards relativism or expressivism instead. But then it seems like we could equally well rehabilitate a broadly Lewisian account of 'knows' in a relativist or expressivist framework. Abandoning the contextualist aspect of Lewis's account for relativism or expressivism seems to preserve all the applications Lewis makes of his contextualism; and it may have independent advantages, as claimed by MacFarlane (2005) for relativism and Chrisman (2007) for expressivism.

one in which it isn't (for all we've said, the actual world is such a world), and even knowing$_{lo}$ requires that one rule those out. But she (reasonably enough) thinks that she does know$_{lo}$ that the wall is red, and thus doesn't know$_{lo}$ that she doesn't know$_{lo}$ this. So the $K\neg K$ principle for know$_{lo}$ (and thus the general $K\neg K$ principle) is refuted by the example; it's just that, since the attributor's use of 'know' does not refer to know$_{lo}$, this does not refute the more modest claim that the $K\neg K$ principle for the relation attributors pick out with 'know' can fail only in cases which are ignored by those attributors.

How convincing is this hard-nosed response? I think that it is most attractive when the difference in what is presupposed by subject and attributors intuitively amounts to a difference in epistemic standards. By Soraya's standards, one does not, in general, have to verify that the lighting is good in order to use one's vision to know what colour an object is. By our standards, one does have to rule out such possibilities. Soraya knows that she doesn't know by our standards. But she reasonably (though falsely) believes that she knows by hers.

However, not all cases in which some attributors attend to a $K\neg K$ failure are intuitively described as cases in which their standards differ from the subject's. In fact, even the case of Soraya needn't be described as such. Perhaps we do not use 'know' in such a way that people need to, quite generally, rule out possibilities with misleading lighting before they can know the colour of an object. We think that many people know the colours of lots of things despite never performing such checks. We just also know about Soraya's specific situation, we know that the lighting in that specific room is unreliable, and thus want to deny knowledge to her in particular. If that is the situation, it doesn't seem as natural to describe us and Soraya as differing in standards; hence it also doesn't seem as natural to reconcile the case with **Elusive K¬K** by appeal to the fact that 'know' means something different for us than it does for Soraya.

(One might hope that such cases cannot arise: by the rule of resemblance, if the attributors attend to *any* possibilities in which the lighting is misleading, every subject has to rule out all of them before she can be said to know. But such a liberal application of the rule of resemblance would be disastrous, at least if 'ignoring' is understood in terms of presupposition.[25] When I was 10, someone stole my bicycle, so that it wasn't where I left it when I went to look for it. Since I know this, there are bike theft possibilities which are consistent with what I presuppose in almost any conversation. It had better not follow that 'know', in my mouth, is so stringent that I say something false whenever I claim of someone that she knows where her bike is.)

It should be noted that, even if it doesn't seem particularly natural, the hard-nosed strategy still applies in the cases where attributors and subject intuitively share standards. Since Soraya is ignoring the possibility that the lighting in this particular room is misleading, and we are not, the Lewisian theory predicts that we use the word 'know' differently – even if, in some intuitive sense, our epistemic standards are the same. We can thus still appeal to the different interpretations of 'know' to reconcile the case with **Elusive K¬K** along the lines indicated above. Doing so is not *ad hoc*, because the Lewisian theory predicts quite independently that these two different interpretations will both be in play. If there is something uncomfortable about the response, then, this is not because it is unnatural by the Lewisian's own lights. Rather, the response draws our attention to a feature of the Lewisian account, that the range of possible interpretations might not correspond to the range of epistemic standards, which some may find unattractive. In the next section, I explore what happens to **Elusive K¬K** when we try to revise the Lewisian account to avoid this feature. It turns out that this yields a different, but also quite attractive, way of learning to live with **Elusive K¬K**.

[25] If we understand 'ignoring' in terms of salience, we cannot handle the cases of hypothetical $K\neg K$ failures described above, since (i) a scenario is salient even if it is discussed only hypothetically, and (ii) subject and attributor attend to all the same possibilities in that case.

2.2 A Conciliatory Response

We attend to the possibility that the lighting next door is misleading; in fact, we positively affirm that possibility. Soraya ignores it. Yet, none of us are inclined to generally take seriously such misleading lighting; and all of us are inclined to do so when we have particular reason to be suspicious. There is thus a clear similarity between our standards and Soraya's, making it somewhat odd that the Lewisian theory predicts that 'know' means something different relative to our different contexts.

It will help to dig a little deeper into where, intuitively, the Lewisian theory goes wrong. I suspect that the problem is that there are really two very different reasons we have for taking possibilities seriously. Some we take seriously because our standards require us to: you just don't qualify for the kind of state we're interested in unless you have ruled these out. Others we take seriously just because we have particular reasons to think that they obtain. Only the former reflect our standards, and so only those who differ in what possibilities they take seriously for the former reason should be classified as using 'know' differently.

Interestingly, this is something like a converse to the *Problem of Known Presuppositions* discussed by Blome-Tillman (2012). Suppose that I'm in a 'high stakes' situation: it really matters to me whether the bank will be open this Saturday, because my paycheck needs to be paid in before Monday if I want to avoid disastrous results.[26] In fact, it matters so much that I'm initially inclined to take seriously that the bank has changed its weekend opening hours during the last month, which was the last time I checked. However, I am now looking at the bank's website, and can see that the opening hours haven't changed, so I stop taking that possibility seriously. Nonetheless, I am inclined to say 'Omar doesn't know that the bank will be open tomorrow' when all he has to go on is that it opened on Saturdays a month ago; and this is true even if Omar, being in a low-stakes situation, believes the bank to be open tomorrow. In this case, my standards seem to make relevant a possibility which, because of the particular

[26] Cf DeRose (1992)

evidence I have, I don't take seriously (in the sense that it is not compatible with my presuppositions); in the wall case, my particular evidence makes me take seriously a possibility (that the lighting next door is misleading) which my standards usually allow me to ignore.

We can solve both problems at once if we interpret 'ignoring' not in terms of which possibilities we take seriously (i.e. are compatible with our presuppositions), but rather in terms of which possibilities we consider *ordinary* or *normal.* When the stakes are high, I take possibilities in which the bank changes its opening hours to be sufficiently ordinary to be worth worrying about, regardless of whether I have evidence that allows *me* to rule it out. Conversely, I might think of all cases of misleading lighting as abnormal despite having evidence that a particular such case has actually occurred. So, in the wall case, we attributors can agree with Soraya that only possibilities with ordinary lighting are normal, so that 'knowledge' means the same relative to our context and hers.

We thus avoid the somewhat counterintuitive feature of the Lewisian account that the hard-nosed defence relied on. In doing so, we make room for a different way of responding to **Elusive K¬K**. For that principle says that counterexamples to $K\neg K$ can only occur in worlds that are 'ignored' by the attributors of knowledge, however that is spelled out. If 'ignoring' is understood in terms of presupposition or salience, that seems implausible, so that an extended reconciliation along the lines outlined in §2.1 is called for. But if 'ignored' is interpreted as meaning simply 'is considered abnormal', the result is not so surprising. When things are normal, rational beliefs amount to knowledge; it is only when the environment is abnormally uncooperative that they do not, leading to $K\neg K$ failures. **Elusive K¬K** thus no longer seems threatening.[27]

[27] Perhaps there will still be potential counterexamples in cases where attributors and subject do, intuitively, differ in their standards. Suppose that we are sceptics, refusing to dismiss any possibilities as abnormal. Should we describe ordinary people as failing to know without knowing that they fail? If so, such an ascription will have to be handled via the 'shifting' strategy developed in §2.1. But I actually have rather mixed feelings about this case; it strikes me as fairly natural to say that ordinary people, at least those that have encountered sceptical worries, do know that they don't *really* know, while a similar claim sounds absurd to me in the case of Soraya

The cost of responding in this way is that, unlike the notion of a presuppositions or of a possibility being salient, the notion of what attributors consider to be ordinary or normal remains somewhat unclear and does not feature elsewhere in our theories. But I do not here want to adjucate between the costs and benefits of the two responses I have suggested. The important point is that, between them, they show that **Elusive K¬K** is, initial appearances to the contrary, no *reductio* of a broadly Lewisian approach to 'knowledge'. The result is *prima facie* problematic if we interpret S so that attending to a world or treating it as a candidate for actuality automatically places it in S. Given such an understanding of S, however, the theory straightforwardly predicts that subject and attributors will often use 'know' differently, thus enabling the Lewisian to endorse the hard-nosed response without being *ad hoc*. If, on the other hand, we interpret S so that something more than salience or being a serious candidate for actuality is required to place a world in S, it is no longer clear that there is anything even *prima facie* implausible about **Elusive K¬K**. Either way then, the Lewisian needn't be worried.

3 Conclusion

The aim of this paper has been to investigate the status of the knowledge iteration principles according to the account of knowledge given by Lewis in "Elusive Knowledge". In §1 I showed how we could both (a) explain the wide-spread impression that Lewis's account vindicates the iteration principles and (b) confirm that, in fact, Holliday (2015) is right to maintain that the account invalidates them both; the key is to be careful to distinguish which parts of the account describe the dependence of knowledge attributions on the attributor's context and which parts describe the dependence of knowledge attributions on the subject's situation. In §2 I argued that, once this ground has been cleared, there

(provided we hold fixed that, in Soraya's case, the attributors don't *generally* take misleading lighting seriously). If that's right, it suggests that shifting, while perhaps possible, isn't obligatory, which would make trouble for the hard-nosed response.

is more to be said: while the $K\neg K$ principle is invalid, counterexamples to it are, in a certain sense, elusive, since they never occur in salient possibilities. I then argued that this consequence is, initial impressions to the contrary, quite defensible.

There are two novel lessons from this discussion that deserve to be highlighted, one general and one specific. The general lesson is that epistemic contextualism interacts in subtle and surprising ways with the knowledge iteration principles. The reason is that the contribution of context doesn't vary with the world of evaluation; it is therefore held fixed when we evaluate what is known at different worlds, and hence held fixed when we evaluate what is known at worlds compatible with the subject's actual knowledge. If we aren't careful, this can make iteration implausibly easy, as on the account discussed in §1.1. And even if we are careful, it leads to highly surprising theorems like **Elusive K¬K**. The connection is complicated somewhat by the fact that, as noted in §2.1, contextualists can cite precedents for holding that the contextual parameter with respect to which an embedded knowledge attribution is interpreted need not always be the one provided by the context of utterance. But this further complication doesn't show that there aren't interesting interactions between contextualism and iteration principles; only that the interaction may be quite complex.

These interactions are worth studying for their own sake, as I've done here. But they also highlight an under-explored difference between contextualist views and their *subject sensitive invariantist* cousins.[28] These two approaches diverge most obviously when we consider third-personal knowledge ascriptions, where ascriber and subject come apart, and those divergences have been discussed in some detail. They may also diverge when it comes to counterfactual or temporal embeddings, again because the contribution of context won't vary as we shift the world (or time) of evaluation, while the contribution of the subject's

[28] See Hawthorne (2004) and Stanley (2005) for subject sensitive invariantist views, and detailed discussion of their relation to contextualism.

situation will. To these known divergence we should now expect to add a third: the two approaches should make different predictions for iterated knowledge attributions. And this is exactly what we find here, since no analogue of **Elusive K¬K** would hold if, in a subject-sensitive invariantist spirit, we replaced the contextually supplied S with a relation R_S representing which possibilities are salient (to the subject, or the attributors, or anyone else) from each world. I have not attempted a systematic evaluation of which position does better with respect to this divergence; but I have argued that, initial impressions to the contrary, contextualists needn't be overly worried.

This brings me to the more specific lesson of our discussion. I have shown that Lewis's account entails **Elusive K¬K**; very roughly, the claim that counterexamples to the $K\neg K$ principle can occur only in possibilities that are being ignored. Somewhat less roughly, rational subjects can fail to know, in the sense of 'knowledge' used by some attributors, without knowing that they fail to know *in this sense*, only if they inhabit possibilities which those attributors are ignoring. Whilst no doubt unexpected, I have argued that this consequence is not so surprising as to be a *reductio* of the Lewisian account. But it is still surprising enough, I think, to be epistemologically significant. Consider, for example, the Williamsonian E=K thesis that one's evidence consists of all and only the propositions one knows. Since $K\neg K$ is non-negotiably false, this will mean that the iteration principles for 'evidence' will fail; and this, in turn, leads to counterexamples to otherwise plausible 'reflection principles'.[29] By maintaining that counterexamples to the $K\neg K$ principle occur only in ignored possibilities, we may be able to ease this tension. Under-described as it is, such an application remains a promissory note. But it is one that we can only even think about writing as a result of the present discussion.[30]

[29] This includes both the standard diachronic reflection principles, as discussed in Williamson (2000, ch.10), Weisberg (2007), and Salow (msc), and synchronic 'rational reflection' principles, as discussed by Christensen (2010), Williamson (2011), Elga (2013), Horowitz (2014), and Lasonen-Aarnio (forthcoming).

[30] I take a first stab at making good on the promissory note in Salow (msa).

Chapter 3

Elusive Externalism

Abstract

Several epistemologists have recently noted a tension between (i) denying an extremely strong form of access internalism and (ii) maintaining that rational agents cannot be epistemically akratic, believing claims akin to 'p, but I shouldn't believe that p'. I bring out the tension, and then develop a way of resolving it. The basic strategy is to say that access internalism is false, but that rational agents always have to believe that the internalist principles happen to be true of them. I show that this would allow us to do justice to the motivations behind both (i) and (ii). I then sketch in some detail a view of evidence that implements this strategy and makes it independently plausible.

An agent is epistemically akratic if her attitude towards p conflicts with her views about what attitude she should take towards p; an extreme case is someone who believes p while taking herself to have overwhelming evidence for its negation. It is natural to think that such a combination of attitudes is a sure sign of irrationality. But, as I explain in §2, it turns out that, unless a strong form of access internalism is true, people can have evidence which supports such akratic conclusions, apparently making it rational for them to be akratic.[1] In particular, there will be such bodies of evidence unless we accept the following negative access principle:

[1] Whilst fairly widely appreciated, the tension between an anti-akrasia requirement and any departure from access internalism is perhaps most explicit in Bergmann (2005), Gibbons (2006), Smithies (2012c), Lasonen-Aarnio (forthcoming), Worsnip (ms), and Dorst (ms).

> **Negative Access**
> If p is not part of one's evidence, one's evidence entails that p is not part of one's evidence.

Access internalists are already committed to a principle of this kind, since they maintain that mere reflection lets us verify the facts that determine what we're justified in believing, and facts about which evidence we have or don't have are obviously amongst those facts. But, for reasons I will briefly explain in §1, the rest of us should find this principle unacceptable.

This leaves us in an uncomfortable position. Should we learn to make our peace with epistemic akrasia?[2] Should we accept an extremely demanding access internalism?[3] Or should we somehow find fault with the more general set up, perhaps by rejecting evidentialism and resigning ourselves to the conclusion that rationality and one's evidence will sometimes make incompatible demands?[4] On the face of it, none of those options seem particularly attractive.

Fortunately, there is another way out. We can reject negative access, avoid epistemic akrasia, and nonetheless accept evidentialism. As I show in §3, this combination of views is consistent if we maintain that one has akrasia supporting (and negative access violating) evidence only when one's beliefs do not perfectly match one's evidence. Moorean conjunctions of the kind 'p but I don't believe that p' demonstrate this possibility rather nicely. For it is clear that some agents have powerful evidence for such a conjunction (when they have powerful evidence for p but, not having reflected on this evidence, fail to believe that p). But it doesn't follow that there could be an agent whose beliefs in a Moorean conjunction conforms to her evidence; and hence it does not follow, even given evidentialism, that anyone could rationally believe it.

That the tension can be resolved in this way is only interesting if we can give

[2] This is the response suggested by Lasonen-Aarnio (2014, forthcoming); it seems to be endorsed, somewhat implicitly, 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] Smithies (2012c) draws this conclusion.

[4] Christensen (2010) floats such a response; Worsnip (ms) advocates it explicitly.

an independent explanation of why one might have akrasia supporting evidence (i.e. negative access violating evidence) only when one's beliefs fail to perfectly match one's evidence. I offer a first pass at such an explanation in §4. This first pass faces some serious problems; but these problems can be resolved if we appeal to epistemic contextualism to refine the proposal, along lines indicated in §5-6.[5]

The possibility of resolving the tension between the denial of negative access and epistemic akrasia has significant consequences for related epistemological debates. For, as others have noted, denials of the following two principles also seem to commit one to rational instances of epistemic akrasia:[6]

> **Positive Access**
> If *p* is part of one's evidence, one's evidence entails that *p* is part of one's evidence.
> **Restricted Fixed Point Thesis**
> If total evidence E isn't strong evidence for H, then E isn't strong evidence that total evidence E is strong evidence for H.

Both of those principles are highly controversial. But, unlike negative access, I do not think that they are indefensible; I will say a little more about this in §7. However, defending these principles does require some controversial commitments. When one thinks that avoiding akrasia in general requires accepting the implausible negative access principle, these commitments may seem undermotivated, since they still do not prevent akrasia in full generality. But, if my

[5] This makes the view we eventually end up with a version of what Greco (ms) calls 'Contextualist Foundationalism'. The details of how I propose to use contextualism are importantly different from the sketch offered by Greco; in particular, I will not use contextualism to defend the claim that Negative Access is true after all, but will instead use it to avoid some of the bad consequences that denying Negative Access seems to have. That being said, the general picture Greco sketches fits very nicely with the view developed here.

[6] Counterexamples to Positive Access feature prominently in Christensen (2010), Williamson (2011), Elga (2013), and Horowitz (2014, §6). Counterexamples to the Fixed Point Thesis are the focus of Christensen (2007), Elga (2007), Coates (2012), Horowitz (2014, §1-5), Titelbaum (forthcoming), Horowitz and Sliwa (forthcoming), and Weatherson (ms). Not all discussions are easy to classify into our three-way categorization.

reconciliation is accepted, the project may be worth a second look.

1 Against Negative Access

To see why the negative access principle ought to be rejected, it helps to see how it fares when combined with Williamson's (2000, ch.9) E=K thesis that one's evidence consists of all and only the claims one knows. Given E=K, negative access is true only if whenever we fail to know something we always know that we so fail. But that principle is clearly false, as has been recognized since the earliest days of epistemic logic.[7] For knowledge requires truth, and so I might fail to know something which I have every reason to think I know, simply because it is false. Faced with a white wall under red lighting, for example, I don't know that it's red, but might well have no reason to suspect this. So I can be ignorant of something without being in a position to know that I'm ignorant; thus, by E=K, the negative access principle is false.

E=K is a controversial thesis; and it might seem that we could easily escape this argument simply by rejecting it. However, following Weatherson (2011, p.451), we can adapt the argument into one which directly targets the negative access principle. The argument makes two substantial assumptions: (i) rational agents can make mistakes about the propositions which sometimes constitute evidence and (ii) only truths can be evidence. Now let p be a proposition of the kind that sometimes constitutes evidence. Then a rational agent might mistakenly believe p, even though it is false, and thus fail to realize that p isn't part of her evidence. But if negative introspection were true, her evidence would entail that her evidence doesn't contain p, and so our agent's failure to realize that p isn't part of her evidence would be a failure of rationality (at least if she considers the question), violating the stipulation that she is rational. Neither of the premises is undeniable,[8] but both are intuitively appealing.

[7] See e.g. Hintikka (1962).

[8] For example, Smithies (2012b) would deny (i) by maintaining that ideally rational agents would never be wrong about their phenomenal states, whilst Goldman (2009) seems to deny (ii).

If I am right that this is the reason we should reject negative access, a couple of things follow. For one, the primary reason why negative access is problematic is because of the possibility of (non-epistemological) error; this is important because it is already plausible (from, e.g. the preface paradox) that agents may face surprising limitations in reflecting on their own propensity for error. For another, it is not obvious that assuming E=K makes negative access more problematic than it would otherwise be, since the problem for negative access which is highlighted by E=K persists even if we reject that thesis. For this reason, and because the contextualist tools I will ultimately be employing are more familiar when applied to 'knowledge' than to 'evidence', I will assume E=K in the subsequent discussion.[9]

It will also help to have a toy example of a case that exhibits the failure of negative access. Our general argument indicates only what structure such examples will have, not what a concrete one will look like. If we wanted the case to be as intuitively compelling as possible, the proposition that's missing from our evidence should probably be a claim about our phenomenology or how things seem to us, which we believe even though it's false. But such cases are hard to think about, and so I will, for convenience, I will use the much more controversial example of the red wall. I will assume that looking at a red wall can yield conclusive evidence that the wall is red, even if one did not previously know that the wall wouldn't be a white wall illuminated by red lighting. We thus get a counterexample to negative access, since when one looks at the white wall under red lighting, one fails to know that it is red without knowing that one so fails: for all one knows, one is actually faced with the real thing and thus does know.

[9] Though the details of E=K won't matter to the account I find most promising; it could do equally well if one's evidence also included claims one is merely in a position to know (cf Williamson (2000, ch.10)), or was restricted to claims known non-inferentially (cf Bacon (2014)), both of which are modifications I find attractive.

2 The Akratic Paradox

In the previous section, I argued that we should reject the negative access principle. However, as I indicated in the introduction, there is a powerful argument that doing so commits us to the rationality of 'epistemic akrasia'. This section will lay out this problem; the rest of the paper will attempt to explain how it can be avoided in a principled manner.

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 epistemically akratic if her attitudes come apart from what she thinks they should be.

What does this come to when we move to a graded framework which recognizes multiple degrees of beliefs or credences? Consider someone who knows that *p* is either a tautology or a contradiction, but has no idea which. It seems reasonable for such a person to have mid-level confidence in *p*, even though she is sure that this is not the attitude her evidence supports. Or consider someone who thinks that it's most likely that her evidence supports *p* to degree .3, but who is also certain that her evidence supports *p* to no less than that degree and may well support it significantly more. It would be natural for such an agent to have more than .3 confidence in *p*, even though this is higher than her best guess at the evidential support.

These examples suggest that the natural generalization of the notion classifies as epistemically akratic anyone whose confidence comes apart from their *estimate* of the evidential support. Unlike guesses, estimates get credit for being close; hence why, for example, it makes sense to estimate above your best guess if you think that your best guess may be too low but definitely isn't too high. Probabilistic frameworks naturally interpret estimates of a quantity as the expected value of that quantity: the weighted average of the possible actual values, weighted by how likely the quantity is to have that value. In the case at hand, this means that an agent is akratic if her credence differs from the expected evidential support, as calculated using her degrees of confidence.[10]

Epistemic akrasia looks like a paradigm instance of the sort of internal conflict that is a sure sign of irrationality.[11] But then we have an inconsistent set of propositions, each one of which looks attractive:

(1) The negative access principle is false.
(2) If the negative access principle is false, someone could have evidence supporting an epistemically akratic state.
(3) If someone could have evidence supporting an akratic state, someone could rationally be epistemically akratic.
(4) Necessarily, anyone who is epistemically akratic is irrational.

The paradox is hard. (1) and (4) require no further motivation. (3) *seems* to merely reiterate a modest form of evidentialism, the claim that rational agents conform their beliefs to their evidence (though I will challenge this appearance shortly). And it is hard to see how one could deny (2). For consider someone who is a counterexample to the negative access principle. Her evidence will be different from what she has reason to think it is. But then her evidence will support one thing (support p to one degree) and what she rationally thinks her evidence is will support another (support 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

[10] Cf Christensen (2010) and Horowitz and Sliwa (forthcoming). I should note that I adopt a specific account of akrasia only to make the discussion more concrete; for as I suggest in Appendix A, it is hard to see how one could block the possibility of bodies of evidence that support states which count as radically akratic on any way of understanding that notion, once one gives up on negative access.

[11] For endorsement, defence, or sympathetic discussion of the claim that epistemic akrasia is irrational, see Adler (2002), Feldman (2005), Kolodny (2005), Gibbons (2006), Christensen (2007, 2010), Smithies (2012c), Elga (2013), Greco (2014b), Horowitz (2014), and Titelbaum (forthcoming). For criticism, see Williamson (2011), Coates (2012), Lasonen-Aarnio (2014, forthcoming), Weatherson (ms).

an akratic state.[12]

To make this abstract tension more compelling, it's worth seeing how it plays out in the case of the red wall.[13] To set up the case, 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 a 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, you know that it's red and hence have evidence which supports the claim that it is 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 at $.1 \times .9 + .9 \times 1 = .99$, whilst nonetheless being only .9 confident of the wall being red.[14] It thus looks as though, if you follow your evidence, you will be akratic.

[12] This intuitive argument is not quite watertight. As formulated, it suffers from presupposition failure: in many cases, there may be no proposition which our agent has reason to think is her total evidence. Moreover, it may be that the uncertainty in what our agent's evidence is 'cancels 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 seem to be an incredible coincidence if the negative access principle were false and yet no agent's evidence ever supported an akratic state. It thus makes (2) plausible, though not, perhaps, undeniable. However, Williamson (2011), Samet (forthcoming), and Dorst (ms) prove general theorems establishing that, if negative access fails, agents can have akrasia-supporting evidence.

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

[14] A divergence of .09 may strike some as sufficiently small as to not be worth worrying 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.

3 Outline of a New Solution

The paradox of epistemic akrasia is hard; however, it is not quite as irresistible as I, following others, have made it out to be. In particular, (3) is not as innocent as it looks. 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 highly non-trivial, even if we accept the evidentialist thesis that rational agents conform their beliefs to their evidence. For it might be that, by conforming her beliefs to the akrasia-supporting evidence, our agent would change what her evidence is in such a way that her new evidence would no longer support the akratic state she would end up in.[15]

It might be helpful to consider a different case, in which this is an intuitive diagnosis. Consider the Moorean conjunction 'q and I do not believe that q'. It is often noted that a claim of this form could be true. More important for our purposes is that it could be supported by my evidence. In fact, it seems obvious that *someone's* evidence supports a proposition of this form: just take anyone whose evidence supports q but who clearly fails to believe it. It doesn't follow that there is some possible agent who, in conforming her beliefs to the evidence, believes 'q and I do not believe that q'. In the Moorean case, we can give a principled (albeit controversial) explanation for why this tempting inference fails: agents always know what they believe, and belief distributes over conjunction; so your evidence can never support 'q but I do not believe that q' if you also believe that claim. So, in the Moorean case, we have a principled (and evidentialism-compatible) reason for insisting on the gap between 'someone's evidence supports p' and 'someone can rationally believe p.' If we can offer an account of evidence which predicts and explains why akrasia-supporting evidence should also be unstable in this way, then we could appeal to this same gap to reject (3) and thereby resolve the paradox.[16]

[15] Cf Smithies (2012c, p.288-292). The account I ultimately favour, as sketched in §5-6, would exploit changes in what 'evidence' refers to when uttered by the agent rather than changes in what evidence she has. But I will set that complication aside for now.

[16] Bergmann (2005) proposes a somewhat similar approach: he maintains that believing that

Most of the remaining paper will survey the prospects for such an account. Before we turn to that question, however, it may be worth explaining in more detail how rejecting (3) is consistent with evidentialism. *That* the two are compatible on any plausible way of articulating evidentialism is, I take it, obvious from the analogy with the Moorean conjunction. But it would still be helpful to dig a little deeper into *how* they are compatible.

For our purposes, we can focus on the following evidentialist thesis:

> **Evidentialism**
> If X has conclusive evidence for *p*, she should believe that *p*.

Like other norms of its kind, this can be understood either as a 'wide scope' norm stating that the agent should either lack this evidence or else believe that p, or as a 'narrow scope' norm stating that, since the agent in fact has conclusive evidence that p, rationality requires her to believe that p. How we explain the case of someone who has conclusive evidence for the Moorean conjunction 'q, but I don't believe that q', but who nevertheless shouldn't believe it, will depend on which interpretation of the norm we favour.

Suppose we like the 'wide scope' interpretation. Then we can maintain that the disjunction ought to be true because its first disjunct ought to be true: the agent ought not to have this evidence. But this is no strange mix of practical and epistemic rationality, as when we say that John mustn't find out about what happened or that Jenna ought to talk to a more reliable source before making up her mind. Rather, the agent ought not to have this evidence because she would have different evidence if she had responded to her evidence as she ought to, i.e. if she had come to believe that q. So the 'wide scope' norm is compatible with the claim that the agent shouldn't believe the Moorean conjunction.

Suppose we like the 'narrow scope' interpretation. Then we do, indeed, have conflicting oughts: the agent ought to believe the Moorean conjunction (since it's

one shouldn't believe *p* defeats one's (doxastic) justification for believing *p*. However, his account does not predict that, or explain why, believing that one shouldn't believe *p* would change one's evidence for *p*; this leaves it open to Smithies' (2012c, p.288-292) objection that it is insufficiently explanatory.

supported by her evidence), but also ought not believe the Moorean conjunction (since doing so is incoherent or self-defeating). But we should deny that this is a problem. In particular, we should deny that this leaves the agent in a dilemma. For, as Jackson (1985) points out, an agent is in a dilemma only if she violates a norm (a true ought claim) no matter what she does. And this is not true of our agent: if she believes that q, and refrains from believing that she doesn't believe it, she will both respect her evidence and satisfy the anti-Moorean constraint. So the 'narrow scope' norm can also be combined with the claim that the agent shouldn't believe the Moorean conjunction, without thereby leaving our agent in a dilemma.

I introduced my favoured solution by saying that we could reject (3), provided we find a way to maintain that agents violate negative access only when they already fail to conform their beliefs to their evidence. That is an adequate solution if all we want to do is explain why agents who perfectly conform their beliefs to their evidence are never akratic. But we might want to go further. For conforming one's beliefs to the evidence might be extremely hard – it might, for example, involve assigning different credences to the wall being red depending on whether one *sees* it to be red or merely *seems to see* it to be red. If that is right, a truly satisfactory treatment of akrasia should aim for more. For we want to explain not just why agents who perfectly conform their belief to the evidence aren't akratic, but also why agents who are reasonable in some less demanding sense aren't akratic either.

Fortunately, the kind of solution I will be presenting can be generalized to cover these slightly harder cases as well. For another way of presenting the strategy is as an attempt to show that reasonable agents can never suspect that they themselves might be counterexamples to the negative access principle. To put it a little more fancifully, counter-examples to negative access, whilst genuine, will always fall into a 'blindspot' of the relevant agent: for structural reasons, they must always escape her notice.[17] If we assume that an all-out

[17] The 'blindspot' terminology is from Sorensen (1988). My usage of it is a little idiosyncratic

belief that p conforms to one's evidence only if one's evidence entails p, and that evidence is true, it follows that agents who conform to their evidence perfectly always do satisfy the negative access principle.[18] But the more basic point about 'blindspots', that one always has to believe that one isn't a counterexample to negative access, will hold even of agents who are only reasonable in the weaker sense of being coherent and epistemologically well-informed.

This is important because no agent of whom the blindspot claim is true will be akratic because of negative access failures – even if that agent is in fact a counterexample to negative access. To see why, let's imagine an agent who knows that she satisfies positive access and knows exactly what the evidential support relation is, but who might not satisfy negative access. It's natural to model the credences of such an agent by having her 'priors' match the a priori evidential probabilities, and having her credences at any point be obtained by conditionalizing her priors on the set of possibilities consistent with what she takes herself to know: the set of possibilities she suspects might obtain. Her estimate for the evidential support for p can then be calculated by summing across worlds w the product of our agent's credence that w is actual with the evidential support for p provided by the evidence our agent would have if w were actual. If our agent satisfies the blindspot constraint, the various evidential propositions that she suspects she might have form a partition of the worlds she suspects might obtain. And it follows from this that her credence will always match her estimate of the evidential support, so that she will not be akratic.[19]

The remainder of this paper will try to sketch how the blindspot thesis, which

since I say that something falls in a blindspot only if the relevant agent still has to believe it even when it's false; it's more common to classify something as a blindspot as long as the relevant agent can't disbelieve it even when it's false, which is consistent with the agent at least suspecting that it's false. (Moorean propositions are blindspots only in this weaker sense.)

[18] This assumption is compatible with everything I say; but I do not mean to endorse it. If it is rejected, the statement of evidentialism becomes somewhat more complicated (since multiple attitudes may now count as conforming perfectly to one's evidence); for that reason, we would also have to be somewhat more careful than I was above when explaining how exactly our strategy is consistent with evidentalism.

[19] See Appendix B for a more careful, and more formal, discussion.

this resolution relies on, could be independently motivated in its full generality.

In closing this section, it is worth noting a non-obvious way in which the strategy I am suggesting is particularly attractive. For, while it's intuitively clear that akratic states are irrational, it's also puzzling why they would be. After all, 'p' and 'I should believe that p' are usually about two different subject matters. One is about politics, the weather, or what-have-you. The other is about epistemology. So why would there be a rational tension in accepting one without the other?[20]

By allowing that one's evidence can support an akratic conjunction, we respect this worry where it is most powerful. The two components of an akratic state are independent, there are situations in which both components are true, and so one's evidence could indicate that one is in one of these situations. But we already know from Moore's paradox that it doesn't follow from these observations that there is no tension in actually believing both components to obtain. By offering an analogous explanation for why this doesn't follow in the case of akratic conjunctions, we reject this worry where it is at its weakest. For it is also clearly absurd to try defending ones akratic belief by saying 'What's the problem? I happen to both believe that it's raining and that my evidence tells against that belief. Meteorology and epistemology are different topics, so where's the tension here?'

4 A First Pass Implementation

So far in this paper, I have reconstructed a paradox and suggested a new strategy for resolving it. The next task is to make plausible that this strategy can be implemented in a systematic and independently motivated way. This, however, marks a change in the dialectic. In setting up the paradox and sketching my way out, I have tried not to make too many controversial assumptions – after all, that part of the paper requires independent motivation. In showing how this sketch

[20] Greco (2014b) and Lasonen-Aarnio (forthcoming) both press this worry.

might be filled in, by contrast, I will often help myself to highly contested theses, without necessarily stopping to defend them. Defending each one here would have been impossible; moreover, the dialectic does not call for it since (a) I am primarily arguing for the *possibility* of implementing the strategy, rather than for a particular implementation, and (b) I am in fact giving an indirect argument for the combination of theses, by showing that they together allow us to resolve the akratic paradox in an attractive manner.

To implement our strategy, we need an account of evidence according to which failures of the access principles always fall into the relevant agent's blindspot. In investigating how such an account might go, we should start by looking more carefully at the case of the wall. I know that if the wall is red and I believe it to be red, I will thereby know that it's red, and so my evidence will entail that it's red. So it seems plausible that if I all-out believe that the wall is red, I cannot be akratic, since it follows from this belief that my evidence entails that the wall is red. This will be true even if, in fact, the wall isn't red, so that my evidence actually supports the akratic state. So this case won't yield rational akrasia.

In order to leave room for akrasia, we thus need to assume that I don't all-out believe that the wall is red; the worry is then that, being uncertain about whether the wall is red, I'll be uncertain about what my evidence is. To block this result, we would need to find a way for the fact that I take seriously the possibility that the wall might be white to prevent my evidence from containing the proposition that the wall is red, even if, in fact, it is.

Since we are assuming $E = K$, a candidate explanation naturally suggests itself: evidence consists of knowledge, and knowledge requires belief; my failure to all-out believe that the wall is red thus ensures that it's not part of my evidence that it is red. When I am uncertain about the wall colour, then, my evidence will be restricted to the claim that the wall looks red, irrespective of its colour. So, by noticing that I don't all-out believe that the wall is red, I can reassure myself that my lack of confidence is definitely in line with my evidence; and knowing this, I won't be akratic.[21]

These reflections on the particular case can be generalized to give a principled account of why agents can *never* suspect that they might currently be a counterexample to negative access, and hence cannot be akratic. The intuitive thought is as follows. Our reason for denying negative access relies on the possibility that we might mistakenly believe a claim and thus fail to notice that, due to its falsity, it isn't evidence. But it is unclear that consistent agents can have a stable worry that they have made a mistake about a particular claim p. For if I worry that I might have made a mistake with regards to p, I need to think that p might be false. But in order to think that p might be false, I have to cease all-out believing that p. And if I don't believe that p, then I definitely don't mistakenly believe it. In particular then, I cannot take seriously the possibility that I have made a mistake about what my evidence is because of a false belief; even though it is of course possible that this is exactly what has happened.[22]

This intuitive generalization can be formalized by showing that $B(\neg Kp \rightarrow K\neg Kp)$ follows from three principles, endorsed, for example, by Lenzen (1978) and Stalnaker (2006):[23]

[21] It should be clear that this approach has any plausibility only if we accept that our evidence is what we know, rather than what we are in a position to know. I think this is a serious problem; but I will focus on a different, though related, one below.

[22] This last step brings us into difficult territory related to the preface paradox. For even if I can't think of any particular belief that it might be mistaken, it is tempting to think that I should nonetheless accept that, in general, I might (and almost certainly do) have mistaken beliefs. It is worth noting that the strategy doesn't immediately require us to reject this more tempting thought – though I have no new suggestions for how to reconcile this belief in an existential claim with the rejection of each of its instances.

[23] For $B(\neg Kp \rightarrow K\neg Kp)$, we argue by cases. Either BKp or $\neg BKp$. If the former, we obviously have $B(\neg Kp \rightarrow K\neg Kp)$. If the latter, we have $\neg Bp$ by the contrapositive of the 'strong belief' principle. So by negative iteration, we have $B\neg Bp$. Since knowledge requires belief, and our agent realizes this, we get $B\neg Kp$ and so $BK\neg Kp$ by strong belief. So again we have $B(\neg Kp \rightarrow K\neg Kp)$.
Interestingly, the same principles also entail $B(Kp \rightarrow KKp)$, the claim that agents always have to believe they satisfy the *positive* access principle. We also argue by cases. Either BKp or $\neg BKp$. If the former, we have $BKKp$ by a simple application of strong belief. If the latter, we get $B\neg Kp$ by the same argument as above. Putting the cases together, we have as a theorem $BKKp \vee B\neg Kp$ and hence $B(KKp \vee \neg Kp)$ and hence $B(Kp \rightarrow KKp)$. But I'm inclined to think that this second

$\neg Bp \rightarrow B\neg Bp$	Negative Belief Iteration
$Kp \rightarrow Bp$	Knowledge Entails Belief
$Bp \rightarrow BKp$	Belief is 'Strong Belief'

Of course, these principles are controversial. Some of that controversy can be side-stepped by noting that it doesn't matter much that the attitude in question be the one picked out by natural language attributions of *belief.* It may well be, for example, that the attitude we ordinarily talk about requires relatively little conviction, so that there is nothing wrong with believing something without believing that one knows it. We can grant this claim and still insist on our point if there is a different attitude (call it *all out believing, being sure, being positive,* or *taking oneself to know*) that satisfies all three principles.

But not all reservations about these principles are merely verbal. Perhaps the most pressing issue is that Negative Belief Iteration seems to sit rather poorly with the general fallibility thesis that we appealed to in arguing against negative access in §1, since it suggests a picture on which rational and reflective agents always know what they believe. I'm inclined to think that we can adopt such a picture and still run the fallibility argument. For it is plausible that, where there are very close constitutive connections between our beliefs and the facts they are about, infallibility is not so puzzling. And, while I think we should be open to such tight constitutive connections in the case of our beliefs, it would be problematic if they also held for the kinds of propositions that comprise our evidence. For our evidence is supposed to *constrain* our beliefs; what is true about the subject matter from which our evidence is drawn thus needs to be fairly independent of what we believe about it.[24] That is a mere sketch of a response; but I hope that it is nonetheless enough to show that we shouldn't dismiss the response because of its reliance on these principles.

result is less interesting, since those inclined to reject $Kp \rightarrow KKp$ should be sceptical about the Strong Belief principle. For repeated application of that principle immediately yields $Bp \rightarrow BK^n p$ for any n; and if one thinks that we have only finitely many iterations of knowledge for most propositions, that result entails that most beliefs, no matter how good their credentials, will be accompanied by infinitely many false ones.

[24] See Srinivasan (2013) and Stalnaker (forthcoming) for related discussion.

Even if we grant the controversial principles, however, and are thus able to generalize the explanation for why agents always think they satisfy negative access, the account remains an unsatisfying way to implement our strategy. For the account predicts an odd kind of circularity in how an agent's low confidence can end up justified. Consider again the wall case discussed above: if I am unsure that the wall is red, and am worried that I am being less confident than I ought to be, it is not reassuring to learn that my lack of belief entails that I do not have the evidence that the wall is red, and that I am therefore conforming my beliefs to the evidence in remaining unsure. Such 'reassurance' makes my lack of confidence self-justifying in an intuitively problematic way.[25]

What generates this problem? Our attempted solution works by noting that an agent's worries about mistakes about p can prevent her from knowing p (and can thus prevent p from being part of her evidence) whether or not she actually makes such a mistake. However, the mechanism we have considered offers no way for our agent's worries about mistakes to prevent *someone else* from knowing p. As a special case, the agent's actual worries do not prevent her less cautious counterfactual self from knowing p; this is, in turn, what makes her lack of confidence problematically self-justifying. For she should now think it very likely that, if she hadn't lacked confidence, her actual lack of confidence would have been unjustified; and that means that her lack of confidence justifies itself.[26]

Once we realize that this is the problem, it becomes hard to see how one could do better. For what mechanism could there possibly be that would allow

[25] The problem feels quite similar to the 'bootstrapping' problem discussed in the literature on intentions. That problem is an objection, originating in Bratman (1987, p.24-27), to views on which forming an intention to ϕ gives an agent a reason to ϕ; the worry is that it allows that an agent might end up being right to ϕ (and thus, presumably, to intend to ϕ) only because he intends to ϕ in the first place.

[26] The same feature also makes this first pass solution unpromising for addressing problems, such as violations of the reflection principle, that arise when an agent worries that she might be a counterexample to the access principle in the future. For an agent's current worries needn't prevent her future self from believing, and thus knowing, p on the basis of a method which fails in possibilities she is now concerned about.

my (actual) doubts to affect *someone else's* (or my own counterfactual) epistemic situation?

5 Contextualism to the Rescue

I can't imagine such a mechanism. But the literature on epistemic contextualism suggests a close surrogate. My doubts cannot change the facts about your epistemic situation. However, they can change my *standards*, and thus, if epistemic contextualism is true, how I should describe your situation. Take, again, the example of the wall. It's natural to think that whether I take seriously the possibility of funny lighting affects my standards for 'knowledge' in the following way: when I don't take the error possibility seriously, I use 'knowledge' so that someone can come to know the colour of a wall just by looking at it; but when I do take them seriously, my usage becomes more stringent so that subjects must now perform an independent check on the lighting to count as 'knowing' (in my stringent sense) what colour it is.[27] Like the previous implementation, this lets us escape the akratic conclusion in this case: when I take the error possibility seriously, I can tell that 'the wall is red' won't be part of my 'evidence' (as I now use that term) regardless of which possibility I'm in, and so I needn't be worried that I should be more confident that the wall is red. But unlike the previous strategy, this reasoning doesn't seem so obviously self-justifying. For if I'd been more confident that the wall is red, I still wouldn't have had more evidence that it is red – I would have just used the word 'evidence' differently.

(I have, and will continue to, put the point in contextualist terms, since that is the most familiar version of the kind of view required. But one could say exactly the same things as a relativist or expressivist about knowledge attributions.[28] What's crucial is that the assessor's standards, which possibilities she takes seriously, influence who she can legitimately describe as knowing,

[27] See e.g. Hawthorne (2004, p.73-77) and Neta (2005).

[28] For relativism about knowledge attributions, see MacFarlane (2005); for expressivism, see Chrisman (2007).

and contextualism, relativism, and expressivism are different frameworks for explaining how this might be possible.)

To assess this contextualist surrogate, we need a schematic version of contextualism to focus on. Start with a (much too) simple reliabilist account of knowledge:

> **Simple Reliabilism**
> X's belief that p at *w* is knowledge iff X doesn't falsely believe that p in possibilities R-related to *w*.

Different versions of this view will offer different candidates for the relation *R*: being close to, being relevantly similar or a relevant alternative to, being at least as normal as.[29] This simple account is easily modified 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:[30]

> **Relativized Reliabilism**
> X's belief that p at *w* is knowledge relative to *S* iff X doesn't falsely believe that p in possibilities R-related to *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: it is the set of possibilities taken seriously by the speakers.

[29] 'Being close to' will result in a safety-theoretic account of the kind endorsed by Sosa (1999) and Williamson (2000); 'being a relevant alternative to' will result in a relevant alternatives theory of the kind advocated by Dretske (1970) and Goldman (1976); 'being at least as normal as' will result in a normal conditions account of the kind favoured by Dretske (1981) and Stalnaker (2006). I don't mean to imply that our intuitive grasp on these different notions guarantees that they are different, though when they are embedded in particular theories such differences may emerge. (One important structural difference is that some of these relations are more plausibly construed as transitive than others; I discuss the importance of this in §7.) Another option is to endorse this kind of account in a non-reductivist spirit, refusing to say anything non-circular about the R-relation, as suggested by Williamson (2009a).

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

Underspecified though it is, this account plausibly generates the judgements about knowledge outlined above. Every world is R-related 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 being taken seriously. Similarly, if X believes that the wall is red merely based on its looks, and the lighting is unreliable, she won't know in any sense of 'knows'; for there will be worlds R-related 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 R-related to the actual 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 correctly describe a belief that the wall is red formed just by looking at it as falling short of knowledge. If, on the other hand, they don't take such possibilities seriously, they would have to say that this belief does qualify as knowledge.

The contextualist surrogate of the first pass proposal departs a little from how we first presented the strategy, since the contextualist view denies that an agent's beliefs literally affect what her evidence is, and maintains instead that they affect how she uses the word 'evidence'. This makes the treatment of epistemically akratic evidence a little less analogous with the treatment of Moore paradoxical evidence discussed in §3; and we should look in some detail at how the solution applies to the case of the red wall to determine whether the plausibility of the solution is affected by this difference.

According to our simple contextualism, what 'knowledge' or 'evidence' means in a context depends on which possibilities the speakers take seriously. In thinking about akrasia, we are interested in the first-personal perspective of the agent facing the wall, the relevant possibilities are the ones the agent himself is taking seriously. We thus need to distinguish two possibilities, according to whether I (the subject/attributer) am taking seriously the possibility that the wall is white with a red light shining on it. If I am not, 'if I am currently facing a red wall, it is part of my evidence that the wall is red' expresses a truth in my mouth and hence 'my evidence supports an akratic state' does too. But it seems plausible that this is something that I cannot recognize, since it is true only because I am faced with a white wall, and that is not a possibility I am taking seriously. More precisely, suppose we understand 'taking seriously' so that it satisfies the following Seriousness-Belief Connection:[31]

> **Seriousness-Belief Connection**
> If w is consistent with X's all-out beliefs, then X takes w seriously.

Then it follows that the possibility in which what I call 'evidence' supports an akratic state, the possibility in which the wall is actually white, is inconsistent with my all-out beliefs. It's thus clear that I won't be akratic in this case.

Now suppose that instead I am taking seriously the possibility that the wall is white with a red light shining on it. Then 'if I am currently facing a red wall, it is part of my evidence that the wall is red' will not express a truth in my mouth, since the belief that the wall is red is one that I sometimes hold mistakenly even in relevant conditions (namely when I fall for the illusion). So 'my evidence supports an akratic state' expresses a falsehood in my mouth as well: I shouldn't be worried that I am being overly cautious in being only .9 confident that the wall is red, since that is the confidence I see as warranted by my evidence in any of the possibilities. So, again, I won't be akratic.

It is tempting to infer from the fact that X has evidence supporting an akratic conjunction that, if X came to believe that conjunction, X would be both rational and akratic. Where does this inference go wrong, on the contextualist account? Suppose that X starts off, at time t, in the bad case: she's faced with a white wall, but doesn't take that possibility seriously. Then her evidence$_{X_t}$ supports the claim she would express with the akratic conjunction 'the wall is .9 likely

[31] Not all contextualist accounts will involve such a notion of 'taking seriously'. One that comes close (close enough for our purposes) is the one developed and defended by Blome-Tillman (2009, 2014), since one pragmatically presupposes something only if one accepts it. Blome-Tillman (2012) discusses a compelling reason for thinking that the converse of the principle is false: one can take a possibility seriously even if, because of one's particular situation, one was able to rule it out.

to be red, but I estimate that my evidence supports it being red to degree .99.' Now, in order to come to recognize this without being inconsistent, she would have to start taking seriously the possibility that the wall is white (we'll call the new set of possibilities this generates X_{t+1}). Once she does so, she can 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$_{X_t}$ supports it being red to degree .99. But this claim no longer has the form of an akratic conjunction: believing or asserting this is no more akratic than doing ϕ whilst believing or asserting that, according to rule R *which you don't endorse*, you shouldn't ϕ. Moreover if our agent were to utter the akratic sentence, she would be saying that the wall is .9 likely to be red, but that she estimates that her evidence$_{X_{t+1}}$ supports it being red to degree .99; and that conjunction is not something she believes, since she knows full well that her evidence$_{X_{t+1}}$ definitely supports the wall being red to degree .9.

From the above discussion, it may seem as if this solution is a case of linguistic trickery: the only thing we predict is that agents will never sincerely utter akratic conjunctions. But I don't think that's right. The important point is that what kinds of propositions we think it makes sense to conform one's beliefs to (call them 'the special propositions') depends on our standards, i.e. on what possibilities we take seriously. Genuine akrasia would be a state in which we have one attitude, whilst also thinking that the special propositions 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 p, such that believing p whilst having her current standards would be such a state. For in trying to follow her evidence, our agent would change her standards and hence which propositions she thinks are special; so if she then comes to believe p, doing so would no longer put her into a state where her attitudes and her views about what the special propositions support come apart.

There are clear affinities between this treatment of akrasia and the one we discussed in §4; but there are also some crucial differences. Most importantly, the cautious agent's failure to believe that the wall is red affects the extension of what she means by 'evidence' via the content of that term rather than in a more direct manner. In particular, the earlier strategy made the cautious agent's lack of confidence look like it justifies itself, since our agent should agree that, had she been more adventurous, she would have had more evidence so that her actual caution would likely have been unreasonable. By contrast, the current strategy faces no such problem, since the cautious agent should think that a bolder version of herself who doesn't take seriously the white wall possibility would have exactly the same evidence available to her, and would hence be overconfident regardless of whether the wall is red or not. Of course, our cautious agent can see that this bolder version of herself might well mean something different when she uses 'evidence'; this bolder version would thus be unimpressed by this charge of overconfidence made on the basis of a notion she isn't interested in. But our cautious agent shouldn't envy her bolder self for that. After all, she cares about proportioning her beliefs to the thing she refers to when *she* uses 'evidence,' and the bolder version of herself simply fails at that task.[32]

6 The Scope of the Contextualist Solution

We have seen that a simple contextualist account allows us to implement our strategy in the case of the red wall, and to do so in a rather attractive way. But it is not enough that we be able to handle that particular case. For this strategy to solve the akratic paradox, we need to show that it can be suitably generalized. In

[32] Note that this response to the 'self-justifying' worry would break down if we allowed for a 'subject sensitive invariantist' or 'subject relativist' interpretation of our framework, on which it is the possibilities which the *subject* (as opposed to the attributor) takes seriously which matter to the truth of 'p is part of S's evidence.' (For subject sensitive invariantism, see Hawthorne (2004) and Stanley (2005).) For on such an interpretation of the framework, our cautious agent should think that the bolder version of herself might well have 'the wall is red' as part of her evidence. In fact, if we identify 'takes w seriously' with 'has beliefs consistent with w', the subject sensitive invariantist view is just a version of the 'solution' considered and rejected in §4. By contrast, relativist or expressivist variants of our view would seem to work just fine.

particular, we need to show that agents always have to believe that they satisfy the negative access principle: quite generally, not just in the case of the red wall.

Fortunately, our contextualist account allows for a straightforward generalization.[33] Recall the bare-bones account:

> **Relativized Reliabilism**
> X's belief that p at w is knowledge relative to S iff X doesn't falsely believe that p in possibilities R-related to w or any world in S.

It follows from this account that there are no counterexamples to the principle $\neg K_S \rightarrow K_S \neg K_S$ in worlds that are themselves in S. For suppose that X's belief that p in world u doesn't amount to knowledge$_S$, where u is in S.[34] Then there is a world v, in which X believes p even though it's false, that is R-related to either u or some world in S. Since u is in S, v is R-related to some world in S. But then it's utterly obvious from the meaning of 'knows$_S$' that the belief that p cannot be knowledge$_S$, since it's false in a world R-related to a world in S. So anyone remotely thoughtful, including X herself, is in a position to know that X's belief doesn't amount to knowledge$_S$. So that belief won't be a counterexample to negative access.

(To see the point, it may help to work through the example of the red wall. When an agent is faced with a white wall, we're supposed to have a counterexample to $\neg K \rightarrow K \neg K$. For, given that the wall isn't red, she doesn't know that it's red; but, since she doesn't know that the lighting is misleading, and if it weren't misleading she could know the colour just by looking, she can't know that she doesn't know this. But these judgements rely on us using 'knows' in such a way

[33] In Salow (mse), I show in some detail that Lewis's (1996) theory of knowledge has a similar consequence, and defend the consequence against some immediate objections. That discussion is a little more detailed than the current one.

[34] This way of arguing assumes that the only case of an agent not knowing p worth considering is one in which she believes p; the natural way to motivate this is via the thought that if she failed to know p because she didn't believe it, she could discover simply by introspection that she fails to know p. This assumes that agents always know (in any sense of 'know') what their beliefs are. As mentioned in §4, I think that assumption is defensible, or, at any rate, somewhat orthogonal to the current reasons for being sceptical about negative access.

that our agent could know the colour of the wall just by looking – i.e. on us using 'knows' in such a way that worlds in which the lighting is misleading are not in S. If they were in S, it would simply false to say that "if [the lighting] weren't misleading [our agent] could know the colour just by looking". To put it slightly differently: our agent can surely tell that she doesn't know the colour of the wall by standards that require her to rule out misleading lighting, since knowing this doesn't require knowing anything about whether the lighting is actually misleading. So the case fails as a counterexample to $\neg K \rightarrow K\neg K$ as soon as the case which happens to be actual, the one in which the lighting is misleading, becomes a serious possibility.)

So we have that there can be no counterexamples to the principle $\neg K_S \rightarrow K_S\neg K_S$ in worlds that are themselves in S. And we also have the

> **Seriousness-Belief Connection**
> If w is consistent with X's all-out beliefs, then X takes w seriously.

Together, these two claims entail that agents must always think that the relation they pick out with 'knowledge' behaves in line with the negative access principle. In this way, we have a perfectly general way of predicting that agents must believe that the information they would describe as 'evidence' behaves in line with negative access, and hence of solving the akratic paradox.

In a sense, that is the result we wanted. Still, one might reasonably worry that it's a *reductio* of the conjunction of **Relativized Reliabilism** and the **Seriousness-Belief Connection**. For note that, unlike the generalization of the first pass implementation in §4, this generalization predicts no first-person/third-person asymmetry. I have to believe that the relation I call 'knowledge' obeys negative access, even when it is you who is the subject of the relation. But that agents have to believe that other people have to obey negative access is even harder to swallow than that they have to believe it of themselves. There may be logical limits to the extent to which I can appreciate my own fallibility; but it is quite mysterious what would limit the extent to which I can appreciate yours.

The point can be made more pressing by focusing on a particular example. Suppose I recreate the wall case, to fool my friend Xumei. She falls for it, taking the wall to be red when it is really white. In talking about her, you and I take the possibility that the wall is white very seriously; in fact, we think it's actual. Still, it seems that we can say that, while Xumei doesn't know that the wall is red, she doesn't know that she doesn't know this. This seems a straightforward counterexample to the prediction derived above by combining **Relativized Reliabilism** and the **Seriousness-Belief Connection**.

There are two ways to respond to this worry, both of which allow us to hold on to the kind of general result we wanted whilst capturing the non-negotiable idea that Xumei doesn't know that she doesn't know that the wall is red. The first, hard-nosed, response holds on to everything we have said so far, and wields in some more sophisticated philosophy of language to reconcile it with our intuitive judgements about Xumei. The second, more conciliatory, response revises the **Seriousness-Belief Connection** in such a way that Xumei's case is no counterexample to the thought that negative access failures cannot happen in possibilities we take seriously.

6.1 Hard-nosed Response: Embedded Attitudes

It seems clear that we can describe Xumei as unaware of her own ignorance about the colour of the wall; yet we are taking her case seriously. Isn't that straight-out inconsistent with the combination of **Relativized Reliabilism** and the **Seriousness-Belief Connection**?

Not quite. For the two these together imply only that Xumei knows that she doesn't know that the wall is red *in our sense of knows.* But she herself is using 'knows' differently. So that consequence is compatible with Xumei being unaware that she doesn't know that the wall is red *in her sense of knows.* And it's not unnatural to think that it is this lack of awareness that we are reporting when we say that she's unaware of her own ignorance.

Let's go through this response more carefully. The first point to note is that

Xumei is not taking seriously all of the possibilities we are; if she were, she would take seriously the possibility that the lighting is misleading, and hence wouldn't believe that the wall is red. It thus follows immediately from our toy contextualist account that 'know' means something different when the relevant S is the set of possibilities Xumei takes seriously from what it does when S is the set of possibilities we take seriously.[35] To have context-invariant labels for these meanings, let's call the first relation know$_X$ and the second know$_{us}$.

What does Xumei believe or know about what she knows$_X$ and knows$_{us}$? It seems pretty clear that she takes herself to know$_X$. That belief is false, since actual reliability is required even to know$_X$. But it is false only because of her particular circumstances: if the lighting hadn't been misleading, she would have known$_X$ what colour the wall was just by looking at it. Since she is unaware of the particular circumstances, her belief that she knows$_X$ is perfectly reasonable despite being false.

By contrast, it is not clear that Xumei takes herself to know$_{us}$ that the wall is red. It's built into the meaning of 'know$_{us}$' that one needs to rule out the possibility in which the wall is white but misleadingly lit before one can know$_{us}$ its colour; and that one needs to do so regardless of whether it's R-related to the actual circumstances. Ignorance of the actual circumstances is thus no excuse for thinking that one knows$_{us}$ what colour the wall is when one hasn't checked on the lighting. Since Xumei can obviously tell that she hasn't checked on the lighting, she is clearly in a position to see that she doesn't know$_{us}$ what colour the wall is. And since, we shall assume, she is rational and has considered the question, that means that she plausibly knows (in either sense) that she doesn't know$_{us}$ what colour the wall is.

Do these predictions of the theory allow us to recover our intuitive judgements about what Xumei knows and believes about what she knows? They

[35] A minor wrinkle: it doesn't quite follow merely from the schematic accout that we get different intensions for 'know', since it could be that S and S' are R-related to the same possibilities, despite having different members. But I take it that, in the particular case at hand, this isn't plausible.

do, if they are combined with a linguistic hypothesis: that, when embedded in an attitude ascription, the contextual parameter for 'knows' is provided by something like the private context of the subject of the embedding attitude.[36] For if that's right, then 'Xumei thinks she knows that the wall is red' says that Xumei thinks that she knows$_X$ that the wall is red; and, as we just saw, that is a correct description of the situation. Similarly, 'Xumei doesn't know that she doesn't know that the wall is red' says that Xumei doesn't know$_{us}$ that she doesn't know$_X$ that the wall is red – and again, that is something predicted by the above discussion. Moreover, it is consistent with the surprising general result which Xumei's case was supposed to be a counterexample against; for that general result entails only that Xumei knows$_{us}$ that she doesn't know$_{us}$ that the wall is red, and that is not only consistent with her being unaware of her ignorance$_X$ but actually quite a plausible thing to say about her case.

6.2 Conciliatory Response: Being Self-Centred

Assessed from the perspective of the theory itself, the hard-nosed response is rather attractive. The two interpretations of 'knows' that it postulates are predicted in a non-*ad hoc* manner by the theory; the judgements about what Xumei thinks she knows$_X$ and knows$_{us}$ are extremely plausible; and the hypothesis that attitude ascriptions shift the context to that of the subject of the embedding attitude is independently supported. Each of the key ingredients is thus well-motivated, and none are wheeled in only to deal with the apparent problem we encountered.

Assessed externally, however, one might still be dissatisfied. This is easiest to see by focusing on the claim that Xumei uses 'know' differently from the way

[36] Note that 'know' would not be unique in this respect. If they are context-sensitive, 'fun' or 'might' work similarly: 'Xumei thinks that rollercoasters are fun' says not that Xumei takes rollercoasters to be fun *for us*, but that she takes them to be fun *for her*. Of course, observations like these might just be taken to support the somewhat popular view that 'fun' and 'might' shouldn't be given a standard contextualist treatment, but a relativist or expressivist one. So let me just re-iterate that nothing essential to the current discussion depends on us adopting epistemic contextualism rather than relativism or expressivism.

we do. This claim is independently predicted by the theory. But it is not terribly intuitive, even from a contextualist perspective. For, on a natural way of filling in the case, we (the attributors) do not *generally* take seriously possibilities with misleading lighting. We describe all kinds of people as knowing the colour of objects despite not having paused to verify the lighting. We just happen to take seriously this otherwise exotic possibility in Xumei's case, because we both know that I set up the lighting to be misleading. If we are using 'know' in a special sense, it is one that makes it very hard for Xumei to know the colour of walls without making it equally hard for other people. But it is not very intuitive to think that senses of 'know' can be extremely fine-grained in this way.[37]

(The worry is not that it shouldn't be harder for Xumei to know than for other people. It's just a fact of life that it *is* harder for her to know, since her environment is uncooperative. But it shouldn't be possible to build into our meaning of 'know' that it be difficult for her and easy for others. One way to tease these apart is by noting that it doesn't actually matter that we're right to think that the lighting is misleading; even if it isn't, Xumei won't know$_{us}$ that the wall is red.)

If we want to avoid this result, we need to think of the attitude that places worlds in S differently. In particular, the attitude has to be such that the S generated is not quite so sensitive to the details of the beliefs of the bearer of the attitude. A somewhat intuitive gloss on such an attitude might be *taking w to be sufficiently ordinary to be generally worth worrying about.* People sometimes differ in which possibilities they bear that attitude to: being cautious, I think that possibilities in which roads have been blocked overnight because of roadworks

[37] Nor can we reply that, as soon as we take seriously the possibility of misleading lighting in Xumei's case, ruling out such error possibilities becomes a universal pre-requisite for knowledge. That would lead to scepticism. For I do take seriously that certain unusual things have happened to particularly unfortunate people: a few years ago, my brother's bike was stolen and so wasn't where he left it; a few weeks ago, the gym changed its opening hours without notifying me; at one point in the past, a friend's uncle died of an unexpected heart-attack. It had better not follow that I use 'know' in such a way that no one can know where their bike is, when the gym shuts, or what they will be lecturing next year.

are ordinary enough to be worth worrying about; thinking me tedious, you disagree. But people can differ a little in their view of the world without differing in what they bear this attitude to: both Xumei and I might agree that possibilities in which the lighting is misleading are not sufficiently ordinary to be worth worrying about in general, even though I (but not Xumei) happen to think that such a possibility obtains in this case.

As that example shows, the attitude of taking w to be sufficiently ordinary to be worth worrying about does not obey an analogue of the **Seriousness-Belief Connection**. This means that the case of Xumei and the misleading lighting doesn't even raise *prima facie* trouble for a version of Relativized Reliabilism that takes S to be generated by this alternative attitude. But the converse of this attractive consequence is that it's no longer clear how we can justify the judgements required for our akrasia-free treatment of the wall case; and it's even harder to see why we should be confident that this treatment can be generalized.

Fortunately, there is a much weaker version of the **Seriousness-Belief Connection** that would do for our purposes. We can call it the

> **Self-Centred Ordinary-Belief Connection**
> If w is consistent with X's beliefs, and w is or is R-related to a possibility in which X currently has a false belief, then X takes w to be sufficiently ordinary to be worth worrying about.

Unlike the original principle, this one gives a special role to the agent's beliefs about herself. In this way, it allows us to interpret the argument at the beginning of this section as an argument that agents always have to believe that they themselves satisfy negative access, without thereby allowing it to be an argument that agents always have to believe that everyone satisfies negative access.

Moreover, assigning such a special role to beliefs speakers have about themselves makes sense from a contextualist perspective. On the contextualist picture, how agents use 'know' or 'evidence' depends on their particular projects. It is thus unsurprising if there ends up being something mildly parochial about its extension. It is important to speakers that their 'knowledge' is something they can rely on, a belief that they needn't worry might be false. It is less important that the same be true of other people's knowledge; so it makes sense to use knowledge in such a way that 'knowing' requires reliably ruling out every possibility in which *we* are wrong, without necessarily ruling out every possibility in which *anyone* is wrong.

Each of the responses is attractive in a different way. Both can recover the judgements we appealed to in §5. And both allow us to give a *general* argument why agent's always have to think that negative access holds for their notion of 'evidence', so that akrasia is avoided quite generally. For that reason, and because my aim in this paper is less to develop a particular view than to show that a general strategy can be implemented in an attractive way, I will not choose between them.

7 Upshots and Extensions

This ends my explanation of how one can reconcile the denial of negative access and a commitment to evidentialism with the irrationality of epistemic akrasia. In §3, I sketched and motivated the strategy at a fairly general level. In §4 I considered an invariantist implementation which, while helpful for fixing ideas, turned out to be unsatisfying. Finally, in §5-6, I explained in some detail how a natural contextualist view could do better. Having put the details of the reconciliation on the table, it makes sense to give a brief overview of how the current proposal (a) interacts with other discussions of epistemic akrasia and (b) could potentially be extended to help with other, formally similar but philosophically quite different, problems facing externalism.

7.1 Akrasia Revisited

As mentioned in the introduction, a ban on epistemic akrasia is not only in tension with denying negative access, but also in tension with denying one of the following two principles:

> **Positive Access**
> For all p, if p is part of one's evidence, one's evidence entails that p is part of one's evidence.
> **Restricted Fixed Point Thesis**
> If total evidence E doesn't support doxastic state D, then E isn't strong evidence that E does support doxastic state D.

The reason is essentially the same as in the case of negative access. Agents who fail Positive Access can have evidence that their evidence is something different from what it actually is; hence their evidence can support an akratic state. Agents who are counterexamples to the Restricted Fixed Point Thesis will have reason to think that their evidence supports something different from what it actually supports; if they follow their evidence, they will thus believe one thing (the thing that their evidence actually supports) whilst believing that their evidence supports another.

So the first question we should ask in assessing the scope of our strategy is whether the kind of strategy developed here could also be used to reconcile an anti-akrasia constraint with denials of the positive access principle or the fixed point thesis. (This should also help in connecting the current proposal to the existing literature since, somewhat surprisingly, that literature has mainly focused on akrasia as it would seem to arise from counterexamples to positive access or the restricted fixed point thesis.) Let us look first at the positive access principle. Here the answer must surely be *in principle, yes.* If one could somehow find a way of defending the view that, while the positive access principle is false, agents can never think that they themselves might be a counterexample to it, that would allow for an equally satisfying resolution of the tension.

In practice, however, I am less inclined to be optimistic. For I can't see any reason to think that counterexamples to the positive access principle, if real, will always fall into the relevant agent's blindspot.[38] The basic problem

[38] It might be worth saying something about why the particular proposals we have discussed cannot be naturally extended to predict blindspots for positive access. The contextualist account

is that doubts about positive access are of a very different kind from doubts about negative access. Doubts about negative access, as explained in §1, arise from worries about our epistemic methods being essentially fallible, and thus sometimes failing to yield evidence in ways that we cannot guard against. It is not so surprising, I think, that such failures fall into our blindspots; the whole point is that they are problems we cannot guard against. By contrast, doubts about positive access, which I will discuss a little more explicitly further down, usually rely on the thought that knowing that one has some evidence requires running an independent check on the reliability of that evidence, and that we may not always be able to run such an independent check. But, if that is right, it is unclear why we couldn't be open-eyed about it, and make use of a belief that we hope to be reliable, while being uncertain (given the lack of an independent check) about whether it really is.

Things are even more problematic when it comes to the fixed point thesis. For our Moorean analogy does not provide any useful guidance here. Why would it be that agents can only get misleading evidence about what a body of evidence supports if they are already failing to conform their beliefs to the evidence? How could conforming one's beliefs to one's evidence change what evidence one has about the evidential support relation? Nothing like the current strategy seems to apply to the fixed point thesis, even in principle.

discussed in §5-6 predicts that positive access will fail if the R-relation is intransitive. If it is, we should get counterexamples to positive access even in possibilities that are taken seriously. For suppose that p is true in w and v, but false in u; moreover, it's believed in all three worlds, and only w is taken seriously. Then, at w, the belief will amount to knowledge (assuming there aren't any other relevant possibilities). But at v it will not. But then the subject's belief, at w, that she knows p would not itself be knowledge, since that belief would have been false at the R-related v. So, at w, the subject knows without being in position to know that she knows, even though w is a serious possibility. See also Salow (mse) for more detailed discussion of this point in the context of Lewis's contextualist theory.

By contrast, the principles appealed to in §4 do allow us to prove that agents always have to believe that they satisfy the positive access principle. So the first pass strategy can, at least in principle, be extended to cover positive access. The problem is that one of the principles, the one stating that when one believes something one believes that one knows it, is hard to reconcile with the denial of positive access. Cf footnote 23.

It thus seems that our strategy can't work as a general response to the problem of akrasia. Is that a problem for our strategy? It certainly would be, if the lesson to draw from these other tensions were that epistemic akrasia is rational after all.[39] For if there are cases where it's rational to be akratic, why not just accept that counterexamples to negative access are simply more such cases? My response is to deny the antecedent; for, while the negative access principle is definitely false, there are interesting (and externalist-friendly) ways of defending both positive access and the fixed point thesis. I cannot here give a detailed survey or evaluation of the possible defences. But I will give a very brief and selective overview of some approaches that complement our strategy rather nicely.

Positive Access. The most influential challenge to positive access arises from considerations about the reliability of judgements made in situations in which our discriminatory capacities are imperfect.[40] When faced with a tree that is 50 inches tall, I can reliably judge that it is between 45 and 55 inches tall. But I cannot reliably judge that this judgement is reliable; after all, I cannot reliably judge that the tree is taller than 46 inches, and, (so the objector claims) if it were not taller than 46 inches my judgement that it is between 45 and 55, while true, would have been to close to being false to still be reliable. So, if reliability is the only relevant consideration for what is part of my evidence in this scenario, my evidence includes the claim that the tree is between 45 and 55 inches tall but does not include the claim that this is part of my evidence.

But, as Greco (2014a) and Stalnaker (forthcoming) have pointed out, this objection relies very much on understanding 'reliability' as something like 'no error in nearby possibilities'. But such an understanding of 'reliability' isn't obligatory; we could instead think of a judgement as reliable if, and only if, it would not have been falsely made in circumstances that are at least as 'normal' as the actual ones. That notion covers the clear cases of cases unreliability as

[39] Or that, as suggested by Christensen (2007, 2010) and Worsnip (ms), there is an irreconcilable conflict between evidential norms and norms of coherence.

[40] For classic articulation, see Williamson (2000, ch.4-5).

well as any other: a case in which the lighting is misleading and the wall is white is just as 'normal' (as far as our perceptual capacities are concerned) as one in which the lighting is misleading and the wall is actually red; hence a judgement that the wall is red, made when the lighting is misleading, is unreliable even on this conception. But this way of conceiving of reliability blocks the above argument. For if the strongest thing I can reliably judge in the actual situation is that the tree is between 45 and 55 inches tall, it must be normal for me to judge this when the tree is 46 inches tall. So a judgement that things are normal (and hence that my first-order judgements are reliable) need not rule out that the tree is 46 inches tall; hence I can reliably make this 'higher-order' judgement without being reliably able to judge that the tree is taller than 46 inches.[41]

The point generalizes. A simple way to see this is that 'is at least as normal as' is a transitive relation. But if the R-relation in our schematic reliabilism is transitive, the resulting account will plausibly vindicate the positive access principle that anyone who knows p is in a position to know that they know p.[42] Yet the resulting account is clearly 'externalist', being a species of reliabilism. In particular, the resulting account is perfectly compatible with the points about fallibility that motivated our rejection of negative access; and it is perfectly consistent with holding that there are no interesting conditions of which we can

[41] For more specific discussion of what to say about cases of limited discriminatory capacities, see Stalnaker (2009) and Cohen and Comesaña (2013). For an argument that treating these cases as counterexamples to positive access has highly counterintuitive consequences, see Salow (msd). For a different response to the reliability argument against positive access, see Das and Salow (ms).

[42] Suppose that, at w, X knows that p. Then X doesn't believe that p falsely in any worlds that are R-related to w or to any world in S. Let v be some world R-related to w or some world in S in which X believes p. Then any world R-related to v or any world in S is also R-related to w or some world in S. So X doesn't believe that p falsely at any world R-related to v or any world in S. So X knows that p at v. So X knows that p at any world R-related to w or some world in S in which X believes p.

Now suppose that, at w, X also believes that she knows that p. We can reasonably assume that she won't have that belief in worlds in which she doesn't even believe p. So we need only look at whether that belief is true in worlds R-related to w or any world in S in which she believes p. But the above paragraph shows that it is. So her belief, at w, that she knows that p is itself knowledge. So the positive access principle is true.

always reliably judge whether they obtain.

Restricted Fixed Point Thesis.[43] It is less clear that there is a similarly salient single argument against the fixed point thesis. On the one hand, there are apparent counterexamples: for example, agents who have evidence that entails p, but, because of reasonable doubts about their own deductive capacities, have good reason to be sceptical about that entailment. On the other hand, there is simply the challenge of why the Restricted Fixed Point Thesis *should* be true: why couldn't a body of evidence be misleading about what it itself supports?

There are many things to be said about both challenges, and plausible answers to one often pull in the opposite direction of plausible answers to the other. Instead of attempting a fair an general assessment, I will just focus on two ways of meeting the general explanatory challenge, to show that it can be met in a way that complements the strategy I have developed here rather nicely.

One way of meeting the explanatory challenge could be called the rationalist route. Facts about what's evidence for what are necessary and knowable *a priori*; they are not the kind of thing we learn by weighing the (empirical) evidence for or against them.[44] So the Restricted Fixed Point Thesis is trivally true: since nothing is evidence for or against claims of the kind 'E supports D', E isn't evidence for or against such a claim either. Note that there is no analogue of this response for defending negative access, since what evidence a subject has is obviously neither necessary nor *a priori*.

One way to dramatize that this explanation addresses only the Fixed Point Thesis is by considering two different kinds of mistakes about what other people should believe. There is clearly nothing wrong with mistakenly believing that Xumei shouldn't believe that p, if that belief is based on a mistaken (but reasonable) belief about Xumei's evidential situation – e.g. if I am (falsely) told that Xumei hasn't yet seen the latest studies, and thus mistakenly conclude that she shouldn't believe the drug to be effective. By contrast, it does seem that one

[43] Thanks to Ginger Schultheis for discussion.

[44] This kind of response is advocated by Titelbaum (forthcoming) and Smithies (forthcoming).

is making some kind of rational mistake if one mistakenly believes that Xumei shouldn't believe that p because one has misevaluated what her evidence supports – e.g. if one knows that Xumei has read the latest studies but mistakenly judges that they do not support the drug's effectiveness. This highlights the disanalogy between negative access and the fixed point thesis: while one can easily have misleading evidence about what someone's evidential situation is like, it is less clear that one can really have misleading evidence about what a body of evidence supports. Unless we can give a principled reason for why one cannot have misleading evidence about one's own evidential situation, the rationalist strategy for defending the Fixed Point Thesis cannot be extended to defend negative access; it thus requires something like the strategy we've been developing to complement it.

Another way of meeting the explanatory challenge is the expressivist route suggested by Greco (2014b). Very roughly, the expressivist holds that a belief that E is evidence for H just is a conditional belief that H, on the hypothesis that E. Suppose then, that E doesn't support H, so that believing H when one has E isn't rational. Now, someone who has the conditional belief that H, on the hypothesis that E, will believe H when they have total evidence E. So they will be irrational. It follows that if E is not in fact evidence for H, one cannot rationally believe that E is evidence for H whilst having evidence E.

Again, it's worth noting that the expressivist response does not help prevent akrasia that results from failures of negative access (a point that isn't immediately obvious from Greco's presentation). For all things considered judgements that someone's evidence supports p are not purely first-order judgement about p, but a mix of conditional beliefs about p and categorical opinions about the subject's evidential situation; hence why it makes sense to say 'p, but Xumei's evidence doesn't support that p.' If one can be misled about one's own evidential situations as one can be about that of others, as failures of the negative access principle would suggest, one should be able to rationally believe 'p, but my evidence doesn't support that p' for the same reason. So the expressivist strategy, like the rationalist one, only helps with defending the Restricted Fixed Point Thesis, and needs to be supplemented with an account of the kind developed here to handle potential cases of epistemic akrasia that arise from failures of negative access.

This discussion of the connection between akrasia, positive access, and the fixed point thesis, has been cursory and selective. I hope that it can nonetheless serve to illustrate two lessons. Firstly, there are promising general strategies for defending positive access and the fixed point thesis, thus avoiding the akratic states that counterexamples to these theses would give rise to. Secondly, those general strategies are themselves limited in scope: while they allow us to avoid akrasia arising from the particular source they discuss, they are not naturally generalized to explain why negative access failures are impossible or why they do not lead to akrasia. This makes it reasonable to think that a general account of akrasia will incorporate the strategy developed in this paper, as well as some of the strategies discussed in this section.

On the resulting picture, there is no perfectly general explanation for why conforming one's beliefs to one's evidence will never require one to be akratic; the explanation will instead be piecemeal. An outstanding question is whether that is in itself a strike against the resulting picture. I myself am inclined to think that it's not, since many of the piecemeal explanations strike me as both (i) correct, in the cases they cover, and (ii) not entirely general. But I have no further argument for this view, other than by looking at the details. Since I can't look more closely at the details here, I will thus have to leave this general worry unanswered.

7.2 Looking Ahead

The primary aim of this paper has been to reconcile denial of the negative access principle with a ban on akrasia (while nonetheless subscribing to evidentialism). But the reconciliation I have proposed has the potential to resolve some other puzzles for access-deniers in a similar manner, problem that arise when agents look ahead to their future evidence. Akrasia occurs when an agent fails to satisfy a 'synchronic' reflection principle, telling her to match her credences to her expectation of the *current* evidential support. But we also get odd consequences when an agent fail to satisfy a 'diachronic' reflection principle, telling her to match her credences to her expectation of the *future* evidential support; such an agent will, for instance, be subject to predictable exploitation and will be able to bias her own investigations to manufacture evidence for propositions she would like to believe. And when an agent anticipates finding herself in a situation in which the access principle might fail, she will generally violate such diachronic reflection principles.[45]

A closely related puzzle is that agents who anticipate violating the negative access principle generally won't expect that conditionalizing on the evidence they will receive will be the response most conducive to forming accurate opinions about the world.[46] In fact, there will even be cases in which they will expect conditionalizing on the additional evidence to actively mislead them, leaving them less accurate than they would have been if they had never encountered it.[47] This seems to require deniers of negative access to choose between denying that conditionalizing on new evidence is the rational way of reacting to it (thus denying evidentialism) or drawing the surprising conclusion that rational agents should sometimes regard further evidence as misleading (which suggest that, if they care about the truth, they should dogmatically avoid it).

I do not here want to argue that these problems for denying negative access are insurmountable. But they do seem somewhat serious. And the strategy developed here has the potential to be extended so that it address these issues

[45] See Williamson (2000) and Salow (msd) for discussion. Hawthorne (2004), Weisberg (2007), and Weatherson (2011) also discuss the connection between the access principles and the reflection principle, though they do not discuss in much detail whether this is a serious problem for access deniers. Salow (msf) argues that this problem is genuinely distinct from the problem of epistemic akrasia.

[46] See Bronfman (2014) and Schoenfield (ms).

[47] See Horowitz (2014) and Salow (msb).

as well. For we may be able to extend our contextualist explanation for why agents must always believe that they are not currently counterexamples to the negative access principle to predict that agents also have to believe that they won't be counterexamples to the negative access principle in the future.[48] If so, we can explain why agents cannot anticipate violations of negative access, and hence can't engage in the weird behaviours just described. To assess the plausibility of such an extension, and to determine how satisfying a solution it offers to the puzzles just cited, we will, of course, have to look at the details. But there is at least reason to hope that the strategy developed here is not a 'one-trick wonder', but a general way for externalists to escape some of the least attractive consequences of their views.

8 Conclusion

We began with a paradox. On the one hand, considerations stemming from our universal fallibility suggest that the negative access principle must be mistaken. On the other hand, rational agents can't be akratic: they can't hold one opinion, while estimating that a different one better reflects their evidence. The two seemed incompatible because, when negative access fails, agent's have evidence that supports an akratic conclusion. I then proposed a way out: agents only have evidence that supports an akratic conclusion if they are already failing to conform perfectly to their evidence; the fact that some agents have akrasia-supporting evidence thus 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, we need a theory of evidence that predicts that, while negative access can fail, agents always have to believe that it holds true of them. I showed that a schematic reliabilist-

[48] Note that there is no hope of extending the invariantist strategy discussed in §4 in this way; I take this to be another reason to think that this strategy is not particularly attractive. Cf footnote 26.

contextualist account of evidence naturally makes such a prediction, and argued that this prediction is both defensible and makes for an attractive akrasia-free treatment of counterexamples 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 hope that the proposed implementation at least passes the threshold of minimal plausibility required to show that something like the general strategy I have suggested might be workable.

If the strategy succeeds, that fact has significant ramifications for other live issues in epistemology. The first is that it makes the project of finding externalist-friendly way of defending the positive access principle or the restricted fixed point thesis more appealing; for, once we have solved the problem with negative access in the way I suggested, such defences would suffice to vindicate the thought that akrasia is irrational. The second arises from the fact that the strategy relies on controversial epistemological theses, such as contextualism about 'evidence'. If there is no equally attractive way of escaping the akratic paradox without these theses, that is a powerful argument in their favour.

At a more general level, the strategy also promises to offer a somewhat deeper insight into the debate between internalism and externalism in epistemology. It is easy to get the sense that how one comes down in this debate depends on how one approaches epistemology.[49] 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, we 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. Our view offers a new way of vindicating this, somewhat vague, intuition. For our view predicts the access principle, which I have, perhaps tendentiously, identified with internalism, to be 'true from the first person perspective,' since any rational agent must believe

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

that it holds of her. But our 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 further progress in the debate will, however, also have to remain a project for future work.

Appendix A 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.[50] And it suggests that a different understanding of what it is for a graded state to be akratic might not classify that state as akratic at all.

These lines of thought aren't promising, because we can construct structurally similar cases which do yield radical akrasia by any measure I can imagine. 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.[51] The state supported by such evidence is clearly *radically* akratic on any remotely reasonable account of akrasia.

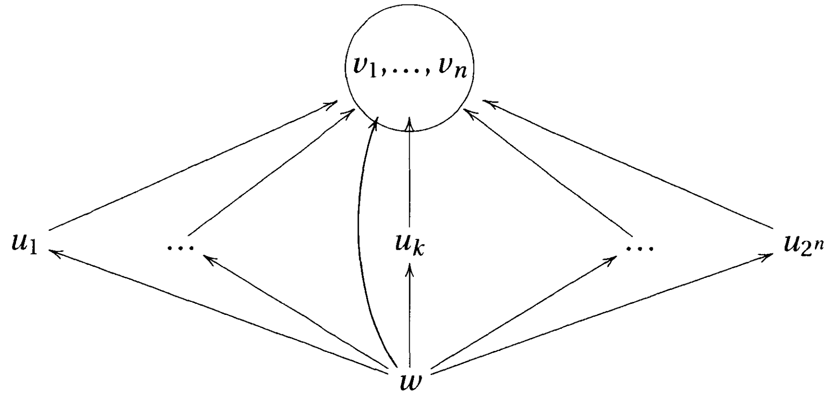
I'll first describe the case abstractly, using epistemic logic. We have a set of possible worlds W and an accessibility relation R between worlds such that wRw' only if the evidence had by the relevant agent at w does not entail that she is not in w'. To simplify the discussion, I will use only finite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

The abstract model goes as follows. For some finite n, $W = \{w, v_1, v_2, \ldots, v_n, u_1, u_2, \ldots u_{2^n}\}$ 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:[52]

[50] Cf Horowitz and Sliwa (forthcoming), who give some arguments to bolster this appearance. Whether this impression is right, however, depends on what exactly one takes to be problematic about akratic agents. As I see it, the problem is that they take themselves to be smarter than their evidence – which doesn't make sense, since they, like everyone else, have nothing other than their evidence to go on. If that is the problem, it does not seem to be alleviated by maintaining that rational agents can only take themselves to be *a little bit* smarter than their evidence.

[51] Thanks to Kevin Dorst for his help with this construction.

[52] 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. I have omitted the reflexive arrows indicating that each world can see itself to avoid clutter.



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 $\frac{2^n}{2^n+n+1}$ which approaches 1 as n increases. Finally, the probability at w of p is $\frac{n}{2^n+n+1}$, 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.

In our model, R is transitive and reflexive; 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. For imagine a creature with 2^n independent mechanisms for learning about the external world; 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.[53] Then we can think of w as the ultimate sceptical scenario, in which all the mechanisms are led astray; of each v_j as a possibility in which all mechanisms are reliable; and of each u_i as a possibility in which exactly one of the mechanisms delivers a falsehood. This will yield the desired accessibility relations, since every working mechanism will deliver information which allows us to rule out the possibility in which that mechanism is malfunctioning.

[53] 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

In the main text, I claimed that an agent who (i) knows that she knows exactly what the evidential support relation is, (ii) knows that her evidence obeys the positive access principle and (iii) takes herself to know that her evidence satisfies the negative access principle will never be akratic.

To prove this rigorously, we need a little bit of 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 takes herself to know. As discussed in the main text, (i) means that we can plausibly 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 given by $Pr_w(p) =_{def} P(p|R_K(w))$.

By positive access, we have that R_K is transitive and reflexive. By the blindspot thesis, we have that, for each w, R_K as restricted to $R_B(w)$ is symmetric. Finally, since the agent knows that she satisfies the positive access principle, we can assume that if our agent doesn't take herself to know that the she knows p, she also won't take herself to know p. Since R_B is supposed to represent what the agent takes herself to know, this yields the constraint that $R_K(u) \subseteq R_B(w)$ whenever $u \in R_B(w)$; i.e. $R_K(u) = R_K(u) \cap R_B(w)$ whenever $u \in R_B(w)$.

Transitivity, reflexivity, and restricted symmetry guarantee that, for each w, R_K as restricted to $R_B(w)$ is a partition of $R_B(w)$. Call the cells of the partition $c_1, \ldots, c_n$. Note that, for any $u \in c_i$, $R_K(u) = c_i$.

Then the law of total probability gives us that

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

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

Adler, J. (2002). *Belief's Own Ethics.* MIT Press.

Bacon, A. (2014). Giving your knowledge half a chance. *Philosophical Studies,* 171:373–397.

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

Blome-Tillman, M. (2009). Knowledge and presuppositions. *Mind,* 118:241–295.

Blome-Tillman, M. (2012). Contextualism and the problem of known presuppositions. In Brown, J. and Gerken, M., editors, *Knowledge Ascriptions.* Oxford UP.

Blome-Tillman, M. (2014). *Knowledge and Presuppositions.* Oxford UP.

Bratman, M. (1987). *Intention, Plans, and Practical Reason.* Harvard UP.

Briggs, R. (2009). Distorted reflection. *Philosophical Review,* 118:59–85.

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

Cappelen, H. and Hawthorne, J. (2009). *Relativism and Monadic Truth.* Oxford UP.

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

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

Christensen, D. (2010). Rational reflection. *Philosophical Perspectives,* 24:121–140.

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

Cohen, S. (1998). Contextualist solutions to epistemological problems: Scepticism, Gettier and the lottery. *Australasian Journal of Philosophy,* 76:289–306.

Cohen, S. (2002). Basic knowledge and the problem of the problem of easy knowledge. *Philosophy and Phenomenological Research*, 65:309–329.

Cohen, S. and Comesaña, J. (2013). Williamson on Gettier cases and epistemic logic. *Inquiry*, 56:15–29.

Das, N. and Salow, B. (ms). Access and transparency.

DeRose, K. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research*, 52:913–929.

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

DeRose, K. (2004). Single scoreboard semantics. *Philosophical Studies*, 119:1–21.

Dorst, K. (ms). Either akrasia is rational or externalism is false.

Douven, I. (2005). Lewis on fallible knowledge. *Australasian Journal of Philosophy*, 83:573–580.

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

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press.

Dutant, J. (ms). Three Lewisian semantics for 'knows'.

Elga, A. (2007). Reflection and disagreement. *Nous*, 41:478–502.

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

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

Gallow, J. (2014). How to learn from theory-dependent evidence; or commutativity and holism: A solution for conditionalizers. *British Journal for the Philosophy of Science*, 65:493–519.

Gibbons, J. (2006). Access externalism. *Mind*, 115:19–39.

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

Goldman, A. (2009). Williamson on knowledge and evidence. In Greenough, P. and Pritchard, D., editors, *Williamson on Knowledge*. Oxford UP.

Goodman, J. (2013). Inexact knowledge without improbable knowing. *Inquiry*, 56:30–53.

Greco, D. (2014a). Could KK be OK? *Journal of Philosophy*, 111:169–197.

Greco, D. (2014b). A puzzle about epistemic akrasia. *Philosophical Studies*, 167:201–219.

Greco, D. (ms). Cognitive mobile homes.

Hawthorne, J. (2004). *Knowledge and Lotteries*. Oxford UP.

Hawthorne, J. and Magidor, O. (2010). Assertion and epistemic opacity. *Mind*, 119:1087–1105.

Hawthorne, J. and Stanley, J. (2008). Knowledge and action. *Journal of Philosophy*, 105:571–90.

Hintikka, J. (1962). *Knowledge and Belief: an Introduction to the Logic of the Two Notions*. Cornell UP.

Holliday, W. (2013). Response to Egré and Xu. In van Benthem, J. and Liu, F., editors, *Logic Across the University: Foundations and Applications*. College Publications.

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

Holton, R. (2003). David Lewis's philosophy of language. *Mind and Language*, 18:286–95.

Horowitz, S. (2014). Epistemic akrasia. *Nous*, 48:718–744.

Horowitz, S. and Sliwa, P. (forthcoming). Respecting *all* the evidence. *Philosophical Studies*.

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

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

Ichikawa, J. (2013). Basic knowledge and contextualist "E=K". *Thought*, 2:282–292.

Jackson, F. (1985). On the semantics and logic of obligation. *Mind*, 94:177–195.

Kadane, J., Schervish, M., and Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, 91:1228–1235.

Kelly, T. (2002). The rationality of belief and some other propositional attitudes. *Philosophical Studies*, 110:163–196.

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *Journal of Philosophy*, 105:611–633.

Kelly, T. (2013). How to be an epistemic permissivist. In Turri, J., editor, *Contemporary Debates in Epistemology*. Blackwell, 2 edition.

Kolodny, N. (2005). Why be rational? *Mind*, 114:509–563.

Kornblith, H. (2001). *Epistemology: Internalism and Externalism*. Wiley Blackwell.

Kripke, S. (2011). Two paradoxes of knowledge. In *Philosophical Troubles*. Oxford UP.

Lasersohn, P. (2009). Relative truth, speaker commitment, and control of implicit arguments. *Synthese*, 166:359–374.

Lasonen-Aarnio, M. (2010). Unreasonable knowledge. *Philosophical Perspectives*, 24:1–21.

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

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

Lenzen, W. (1978). *Recent Work in Epistemic Logic*. Acta Philosophica Fennica.

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

Littlejohn, C. (2013). No evidence is false. *Acta Analytica*, 28:145–59.

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

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

Parfit, D. (2011). *On What Matters.* Oxford UP.

Popper, K. (1961). *The logic of scientific discovery.* Science Editions.

Salow, B. (msa). Elusive externalism.

Salow, B. (msb). Expecting misleading evidence.

Salow, B. (msc). The externalist's guide to fishing for compliments.

Salow, B. (msd). The externalist's guide to fishing for compliments.

Salow, B. (mse). Lewis on iterated knowledge.

Salow, B. (msf). Reflection without access?

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

Schaffer, J. (2004). Scepticism, contextualism, and discrimination. *Philosophy and Phenomenological Research*, 69:138–155.

Schoenfield, M. (2014). Permission to believe: Why permissivism is true and what it tells us about irrelevant influences on beliefâĂŹ. *Nous*, 48:193–218.

Schoenfield, M. (ms). Conditionalization does not (in general) maximize expected accuracy.

Smithies, D. (2012a). Mentalism and epistemic transparency. *Australasian Journal of Philosophy*, 90:723–741.

Smithies, D. (2012b). Mentalism and epistemic transparency. *Australasian Journal of Philosophy*, 90:723–741.

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

Smithies, D. (2014). The phenomenal basis of epistemic justification. In Kallestrup, J. and Sprevak, M., editors, *New Waves in Philosophy of Mind.* Palgrave Macmillan.

Smithies, D. (forthcoming). Ideal rationality and logical omniscience. *Synthese.*

Sober, E. (2009). Absence of evidence and evidence of absence: Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies*, 143:63–90.

Sorensen, R. (1988). *Blindspots.* Oxford UP.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, 13:141–154.

Srinivasan, A. (2013). Are we luminous? *Philosophy and Phenomenological Research*, 90:294–319.

Stalnaker, R. (1988). Belief attribution and context. In Grimm, R. and Merill, D., editors, *Contents of Thought.* University of Arizona Press.

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

Stalnaker, R. (2009). On Hawthorne and Magidor on assertion, context, and epistemic accessibility. *Mind*, 118:399–409.

Stalnaker, R. (forthcoming). Luminosity and the KK thesis. In Goldberg, S., editor, *Externalism, Self-Knowledge, and Skepticism.* Cambridge UP.

Stanley, J. (2005). *Knowledge and Practical Interests.* Oxford UP.

Titelbaum, M. (2010). Tell me you love me: Bootstrapping, externalism, and no-lose epistemology. *Philosophical Studies*, 149:119–134.

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

van Fraassen, B. (1984). Belief and the will. *Journal of Philosophy*, 81:235–56.

van Fraassen, B. (1995). Belief and the problem of Ulysses and the sirens. *Philosophical Studies*, 77:7–37.

Vogel, J. (1999). The new relevant alternatives theory. *Nous*, 33:155–80.

Vogel, J. (2000). Reliabilism leveled. *The Journal of Philosophy*, 97:602–623.

Weatherson, B. (2008). Attitudes and relativism. *Philosophical Perspectives*, 22:527–544.

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

Weatherson, B. (ms). Do judgements screen evidence?

Wedgwood, R. (2002). Internalism explained. *Philosophy and Phenomenological Research*, 65:349–369.

Weisberg, J. (2007). Conditionalization, reflection, and self-knowledge. *Philosophical Studies*, 135:179–197.

White, R. (2006). Problems for dogmatism. *Philosophical Studies*, 131:525–557.

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

Williams, M. (2001). Contextualism, externalism, and epistemic standards. *Philosophical Studies*, 103:1–23.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford UP.

Williamson, T. (2001). Comments on Michael Williams 'scepticism, contextualism, and discrimination'. *Philosophical Studies*, 103:25–33.

Williamson, T. (2009a). Probability and danger. *The Amherst Lecture in Philosophy*, 4:1–35.

Williamson, T. (2009b). Reply to Stephen Schiffer. In Greenough, P. and Pritchard, D., editors, *Williamson on Knowledge*. Oxford UP.

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

Williamson, T. (2013). Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier cases in epistemic logic. *Inquiry*, 56:77–96.

Worsnip, A. (ms). The conflict of evidence and coherence.

Yalcin, S. (2012). A counterexample to modus tollens. *Journal of Philosophical Logic*, 41:1001–1024.