The Externalist's Guide to Fishing for Compliments

Bernhard Salow
Trinity College, Cambridge
bs416@cam.ac.uk

Suppose you'd like to believe that *p* (for example, 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*. This 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.

Sometimes, the truth is bleak. When this happens, we might prefer not to know; we might even prefer to have false beliefs. For the same reason, however, when a bleak claim is false, discovering that 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 that 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 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 whom 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, for instance, some predictable failure to assess my evidence correctly.

Mind, Vol. 127 . 507 . July 2018
doi:10.1093/mind/fzw029 Advance Access publication 20 February 2017

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, by precisifying and tentatively defending 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 do not work. Finally, in §3, I will explain the more abstract connection between the two topics, by formalizing both questions in a Bayesian framework and showing that the access principle and the impossibility of intentionally biased inquiry stand and fall together. This strengthens the objection against access deniers, by showing that the problems we discovered are unavoidable consequences of denying the access principle rather than irrelevant issues arising from the particular examples. And it strengthens the initial defence of the impossibility of intentionally biased inquiry, by showing that strategies which exploit alleged counterexamples to access, strategies we previously saw to be patently absurd, are 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.

[1] Titelbaum (2010) offers related, but less general, considerations. I should note that, in presenting the argument in this way, I am being a little disingenuous. I suspect that we can 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 be sure* 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 people suspect that it will fail for them? In other work, I try to show 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.

We have already seen one example with this structure: I want to know about my popularity only if I'm popular; and so I prefer receiving evidence that my colleagues like me whether or not they actually do. And other examples aren't hard to come by. I want to know whether I have such-and-such a fatal illness only if I don't; if I do have it, I'd rather live out my final days in blissful ignorance. Again, then, I prefer evidence reassuring me that I am in good health regardless of whether I am.

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? Some philosophers think I can. Parfit, 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. (Parfit 2011, p. 421)

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

1.1 What is intentionally biased inquiry?

I will think of inquiry as a process of evidence collection, to be characterized by the kind of evidence it yields rather than the opinions it results in.[4] This means that we can immediately set aside some otherwise plausible examples of intentionally biased inquiry. It is undeniable that I can exploit known biases in my perception or reasoning to

[3] A similar thought is implicit 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' (Kripke 2011, p. 49). This assumes that the actions described could be a way of intentionally favouring the known claim over its negation.

[4] One might worry that this precludes armchair inquiry, which (arguably) doesn't generate new evidence, but instead tries to determine what our old evidence supported all along. But this is not altogether unwelcome, since some of my arguments (especially the formal arguments of §3) do not straightforwardly apply to cases of armchair inquiry. I'm thus happy to leave open whether an intentionally biased armchair inquiry is possible; though the informal considerations of §1.2 suggest that it is at least much harder than one might have thought.

set up an inquiry that is more likely to leave me with one belief rather than another. For example, I might know that, if I were to print off the case for *p* in a clean, bold font on high-quality paper, and the case against *p* in Comic Sans, I would be more likely to believe *p* after reading both. But this is only because I would also be more likely to believe *p* even when this wasn't supported by the total evidence; this printing manoeuvre thus doesn't count as biasing my *inquiry*, since it doesn't bias the *evidence* I receive.[5] Similarly, if I exploit arational changes in my 'standards' or 'inductive policies' for assessing evidence (if these are possible[6]), I might succeed in biasing my future beliefs, this time without exploiting irrationality. But, again, this would not qualify as intentionally biasing my inquiry, since it doesn't involve biasing what evidence I receive.

This already makes clear that characterizing inquiry in terms of evidence, rather than beliefs, is a significant decision; and its significance will only grow in §1.2, when I argue that strategies such as reading selectively also bias at most one's verdicts, and not one's evidence. But it is the right decision. For if we characterize inquiry in terms of the attitudes it generates, it is unclear what distinguishes biasing one's inquiries from cruder methods of manipulating one's beliefs, such as hypnosis or brainwashing. Admittedly, most of us don't have ready access to these cruder strategies. But the appeal of biasing one's inquiries seems to be due less to greater practicality and more to the thought that it is less unsavoury, being quite consistent with only forming beliefs in response to evidence—or, to use Parfit's phrasing, with 'always respond[ing] rationally to any epistemic reason or apparent reason' (Parfit 2011, p. 421).

When is an inquiry, a process of evidence collection, biased towards *p*? Perhaps the most obvious case is when the inquiry is a *sure-win investigation*: the total evidence gathered in the investigation is guaranteed to be evidence for *p*. Slightly more generally, an inquiry still

[5] This is compatible with different accounts of the relevant biases. On the most straightforward account, they lead to belief without supplying any additional evidence. Even if they do supply evidence, however (for instance, by affecting what 'feels plausible', which may be a form of evidence), the import of this evidence is plausibly nullified if I know that it arises *only* because of the biased mechanism—cf. Kelly (2008), who offers a sophisticated account of how various biases can offer evidence, but 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 is not particularly likely to support *p* any more than it does at the outset, even if I am likely to form the belief.

[6] See Kelly (2013) and Schoenfield (2014) for discussion.

seems biased when it is what Titelbaum (2010) calls a *no-lose investigation*: it is designed so that the total evidence gathered might be evidence for p, but is guaranteed not to be evidence against p. Even this, however, is insufficiently general—as is clear from Parfit's discussion, it can be enough to reduce, rather than eliminate, the risk of evidence against p. In particular, when an inquiry is very likely to yield strong additional evidence for p (that is, evidence that would significantly raise p's evidential support) and has only a small chance of yielding at most quite evidence against p (that is, evidence that would lower p's evidential support at most by a little bit), it still looks like an investigation skewed in p's favour.

The natural way of making this idea precise uses the idea of the *expected value* of a function V; this is the weighted average of all the values V might take, each weighted by the probability that it will take that value. Since we are interested in the rise and fall of p's evidential support, our choice of 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 (so V will take a negative value when the inquiry results in a decrease in the support for p). The obvious such V simply subtracts the initial evidential support from the later evidential support; the expected value can then be calculated by assigning probabilities to the various values (positive and negative) 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. No-lose investigations count as biased by this definition, since they might result in a positive difference (when they yield evidence for p), but are guaranteed not to result in a negative one (since they can't yield evidence against p), so that the expected difference must be positive. By contrast, an investigation that is simply guaranteed to conclusively settle whether p looks as though it will be unbiased.[7,8]

[7] We can prove this, on the assumption that the evidential support of a proposition is measured by its probability on the evidence. For let x be p's initial probability, and hence the degree to which the initial evidence supports p. Then the investigation has an x chance of boosting that support from x to 1, and a $(1 - x)$ chance of reducing that probability from x to 0, resulting in an expected difference of $x(1 - x) + (1 - x)(0 - x) = 0$. So the investigation is unbiased.

[8] A few comments on this definition:

(a) Strictly speaking, the definition doesn't say what it *is* for an inquiry to be biased, but rather, what it is for a body of information to *classify* an inquiry as biased; we need a body of information to determine the probability of the possible changes in

We now know what it is for an inquiry to be biased. Can an agent ensure that her own inquiries are biased in this way? We already saw that one obvious kind of approach – the kind where the agent exploits known irrationalities in how she forms beliefs—won't qualify on our account. There remain two other types of strategies I want to set aside. The first are cases in which the agent's current actions themselves provide 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

evidential support (or, in the less general versions, to determine what is 'guaranteed' to happen and what 'might' happen). Since I will be interested in inquiries that are biased from the investigator's point of view, I will suppress this qualification, and say that A's inquiry is biased if it is biased relative to A's (initial) evidence.

(b) The definition assumes that we can talk about the probability, given the agent's initial evidence, of V taking various values. I don't know how to characterize the intuitive idea that unlikely but large increases can be 'balanced out' by small but likely decreases without that assumption; but if there is a way to do it, the definition could be amended to use such an alternative characterization instead. It's also worth emphasizing that even if we can talk about the probability of a proposition on the agent's evidence, it doesn't follow that this is what measures the evidential support of that proposition; if we prefer a non-probabilistic theory of evidential support, we are free to use that theory in determining V's values instead. To that extent, the definition does not presuppose full-blown Bayesianism; and the considerations I offer against the possibility of intentionally biased inquiry in §1.2 are similarly ones that any theory should recognize.

(c) The definition allows that an inquiry can be biased towards or against p even if the inquiry is not an inquiry into whether p (or, more generally, into a question to which p is an answer). This is desirable: many investigations are unbiased about their 'official' topic, but biased about a related issue, on which the evidence gathered also seems to bear. It also means that our definition is not restricted to inquiries with yes/no answers, or even to inquiries with an antecedently understood range of possible answers. (Thanks to the associate editor for pushing me to clarify this.)

(d) There are investigations motivated by a desire for a certain kind of outcome that are unbiased by this definition. 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 0.9 that I am popular. Support of degree 0.3 is no worse than support of degree 0.7, and support of degree 0.99 is no better than support of degree 0.91. Then I could, as Kelly (2002, p. 170) points out, decide to keep inquiring until the first moment the evidential support exceeds 0.9, and stop immediately after. That decision increases the chances that my preferences will be satisfied; but this investigation may come out as unbiased.

(e) A different kind of unbiased, yet 'motivated', inquiry exploits the fact that we can desire evidential support for a specific opinion under a non-transparent mode of presentation. For example, I might not know where my friends stand on the issue of whether p, and care only about standing on the same side as them. As long as I think that their opinion is likely to be true, I could then simply inquire into whether p, and thereby increase my chances of obtaining evidence for the answer I want to believe. (Thanks to the associate editor for this example.)

second are cases in which the agent expects to lose relevant information in the course of the investigation, as when our agent asks a friend to tell her a year from now that she is popular, knowing that she will forget having given these instructions. These ways of biasing one's inquiry may succeed, but they seem intuitively very different from the one Parfit identifies.[9] 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). And when I exploit information loss, my biasing only succeeds because, at some point or another, I lack important information about what is going on. I will thus set these strategies aside, and restrict 'intentionally biased inquiry' to cases in which an agent succeeds in biasing her investigations in intuitively 'open-eyed' ways.

1.2 Intentionally biased inquiry is impossible

With this sharpening of the question in mind, we can revisit the issue 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. Thought about in 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 two lines of thought? 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

[9] In fact, Parfit (2011, p. 421) explicitly sets aside the cases where one's own actions are evidentially relevant to the question.

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 incriminating facts that present the defendant in a particularly negative light. And the facts they did present were no more compelling than you would expect from such 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 were 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 exactly analogous. 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 a book will contain facts of a certain kind—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 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) would 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*.[10]

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 (rational) expectation that it is true, and that we would learn it, if *p* is true. But those expectations depend on what we know about the set-up. It is no surprise 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 innocent. Similar things apply to the creationist literature. When we try to bias our inquiry into *p* by affecting what evidence we might find, we are automatically changing what we should expect, and are thereby changing what counts as evidence for or against *p*. In other words, knowing of the bias (as we must if the biasing is intentional and 'open-eyed') undermines its effects.

This dynamic is quite familiar with a question far more important than the controversy over evolution: the question of what others think of us. We may try to nudge others into saying nice things; but we also know that, as the nudging becomes more obvious (or is targeted at a more receptive person), the nice words will become less meaningful and their absence more painful. It doesn't take much reflection to conclude that if we weren't so good at selectively forgetting how hard we tried, and how often we failed, the nudges would be pointless.

Despite this suggestive explanation, however, we don't yet have a general argument that intentionally biased inquiry must be impossible. But once we have disarmed the thought that it's obviously possible, this claim does look rather attractive. We experience questions like those concerning our popularity as putting us into a bind: we do not know how to inquire into them in a way that will get us what we want, and so we inquire, if at all, only reluctantly. The impossibility of intentionally biased inquiry would be an excellent explanation of this bind. Even if, say, you had a very loyal friend who knows you and your beliefs extremely well, and who also knows how popular you are, you

[10] Of course, it is also plausible that people generally don't give adequate weight to information about how the evidence they have was selected when they form their beliefs; see, for example, Kahneman (2011), who refers to this phenomenon as 'what you see is all there is'. The kinds of strategies Parfit describes might thus be quite effective in helping us form the desired beliefs; but they will achieve this only by exploiting our failure to conform our beliefs to our total evidence.

couldn't use him. Asking him to tell you if you're popular, and say nothing if you're not, is obviously pointless. And more complicated schemes (such as having him expose you only to carefully selected books) look as though they too would work only if you somehow failed to properly think through the information you would receive. Your efforts to bias your inquiries would keep undermining themselves.

This is not a knock-down argument. One might remain hopeful that, with enough ingenuity, we will think of better strategies—intentionally biased inquiry might not be impossible, just really difficult. In §2, I will elaborate some strategies which, according to certain 'externalist' accounts of evidence, should succeed. But it will, I think, be intuitively clear that they would fail. In §3, I will argue that, given a Bayesian theory of evidential support, these 'externalist' strategies are really the only ones that stand a chance. With these additional pieces in place, we will then have a kind of impossibility proof to bolster the less conclusive reflections just presented.[11]

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;[12] this is a significant oversight, since it turns out to be tightly

[11] This might be a good point at which to mention a very different kind of case that initially seems to raise doubts for my position (Thanks to Caspar Hare, Vann McGee, Jack Spencer and Roger White for discussion). 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 has 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.

[12] 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, pp. 543–9), Sober (2009) and Titelbaum (2010).

connected to the debate between epistemological internalism and externalism.[13]

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 (that is, 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 (namely, our evidence) had better tell us what our evidence is. In other words, the access principle had better be true.[14]

It will be helpful to separate 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.[15] 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' principles 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 externalist arguments of this kind that seem, on first sight, quite convincing: one is based on an alleged epistemic asymmetry between 'good' and 'bad'

[13] The issue also bears on whether we should postulate epistemic norms for how agents should structure their inquiries, or whether all apparent need for such norms is covered by the requirement that we conform our beliefs to the evidence. See Hedden (2015, ch. 10) for discussion.

[14] 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.

[15] Hintikka (1962) 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.

cases (§2.1), the other on our limited discriminatory capacities (§2.2). I will sketch how these considerations motivate concrete (though perhaps oversimplified) counterexamples to the access principle; I will then show that if these cases really were counterexamples to the access principle, someone could exploit them to intentionally bias her inquiry in ways that obviously don't work. This refutes the (alleged) counterexamples, and thereby casts some doubt on the considerations which motivated them. It thus constitutes an indirect defence of the access principle. More importantly, however, it 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, which, I argue, supports 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 specifically targets the negative access principle. We propose 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 (combined with the description of the cases) entails that I am in the bad case, 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 reflection on sceptical scenarios. 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.[16] 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.

[16] 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.

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 favourable 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.[17] To isolate the structural feature, it might help to consult this diagram of the situation:

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

Reflecting on intentionally biased inquiry reveals problems in this way of thinking about the case. 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 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 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 like me, I will see a red wall, and will thus receive conclusive evidence that they do. And 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

[17] There are other cases we might go to for examples of this kind. One might 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 hold 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.

have for my popularity seems a good way of achieving this goal,[18] it seems clear that I should initiate the strategy.[19]

This conclusion is obviously absurd: the procedure just outlined is no way out of the bind we have been discussing. It must, therefore, have been a mistake to think that if only the circumstances were favourable, looking at the wall would tell me more about its colour than that it *appears* red. But if that was a mistake, then so was the thought that this case is a counterexample to negative access. 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.

Someone convinced by the counterexample to negative access might respond instead that the example's structure changes when I initiate the strategy described above: while in an 'ordinary' case the fact that the wall is red can become part of my evidence when I look at it in favourable circumstances, that fact can never become part of my evidence when I look at the wall my friend prepared.[20] If there were such a change, we could say both that the attempt at setting up a no-lose investigation fails and that an ordinary case of looking at a wall is, nonetheless, a counterexample to negative access. But postulating such a change is theoretically unsatisfying. We can fill in the example so that it is highly unlikely, given my background information, that I am unpopular (and hence that the lighting would be misleading); and we can also postulate that, in fact, the possibilities in which I am unpopular are extremely 'remote' and 'much less normal' than the actual ones. It is then unclear what is supposed to distinguish my situation when I look at the wall my friend prepared from my situation in an ordinary case of judging the colour of a wall without verifying the

[18] One might challenge this step, since the cases in question are 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. In §2.3, I explain why a response along such lines doesn't allow us to avoid all versions of the problem.

[19] Titelbaum (2010) makes the same observation about cases like that of the wall. He presents this as a diagnosis of why the 'bootstrapping' reasoning apparently supported by the wall case (as observed by Vogel 2000 and Cohen 2002) is so bad. 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.

[20] The thought might be that my knowledge that I asked my friend to set up misleading lighting if I am unpopular gives me reason to think that circumstances are unfavourable; and externalists have recognized that reasons to think that the circumstances are unfavourable can prevent us from receiving evidence, even when the circumstances are actually good, since at least Goldman (1979, p. 20).

lighting.[21] And even if we waive this worry, it still seems to me that this style of response addresses only the symptoms, not the disease. For, according to this line of response, an ordinary case of looking at a wall is still a no-lose investigation; not, admittedly, into whether the wall is red (since it might look some other colour), but into whether the wall is the colour that it looks. After all, if the wall's colour is what it looks to be, we get conclusive evidence that it is; and if the wall's colour is not what it looks to be, we get no evidence either way. Using this (alleged) fact to set up a no-lose investigation first into whether the wall is red (by ensuring that I know beforehand that it will look red) and then into whether I am popular (by correlating redness with my popularity) dramatizes just how weird a claim this is. But it is, I think, primarily a dramatization of something whose weirdness we can also appreciate directly. For these two reasons, a response that focuses on the unusual circumstances of the inquiry described doesn't seem promising.

To see how robust our problem is for alleged examples of the good case/bad case structure, it's worth looking at how matters play out in the more intuitive cases recently proposed by Lasonen-Aarnio (2015).[22] 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' (Lasonen-Aarnio 2015, p. 160). Here's one such case (not quite the one used by Lasonen-Aarnio, but similar in spirit): 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 for you to 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

[21] Could the difference simply be that in this case, but not in an 'ordinary' one, the error possibility is salient to me? (Thanks to the associate editor for raising this question.) It seems implausible that, as the response considered in the main text would require, the salience of the error possibility should affect what evidence I receive (as opposed to, say, which beliefs I form based on that evidence). But it may be that, as an epistemic contextualist might hold, the salience of the error possibility can affect what I mean by 'evidence', and hence what 'evidence' I believe myself to have. This seems like a promising start for developing a way out along the lines suggested in footnote 1.

[22] Thanks to Jack Spencer for discussion.

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 account 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 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 the cases where the intuition she draws on is clearest, 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 my 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 counter-intuitive that I can reach this conclusion. But if we imagine my 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'.[23] This is well brought out by the attempt to use the case to set up a no-lose investigation. To do this, 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 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—or, at least, has no additional force over and above the anti-sceptical intuition some of us have even in cases like that of the red wall.

2.2 Limited discriminatory capacities

Our second class of examples targets the positive access principle as well as the negative one. Williamson (2000) has 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.[24]

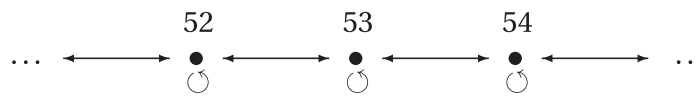
The basic case is as follows. Imagine that you are faced with an unmarked clock, with a single hand that can point in any one of 60 slightly different directions. Your ability to discriminate where it is pointing is good, but not unlimited. If you are to be reliable in your judgements, you need to leave 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.[25]

[23] 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.

[24] See, for example, Christensen (2010), Elga (2013) and Horowitz (2014). The 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).

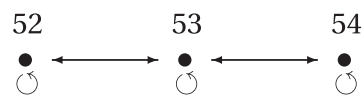
[25] This last claim is, of course, unlikely to be true of actual humans. It is also unrealistic to suppose that our evidence allows us only to rule out some positions; presumably it also

It is somewhat natural to identify your evidence with the strongest claim about the hand's position which you can reliably get right.[26] But this means that the total evidence propositions we obtain 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 the other, and yet they yield different evidence. Again, it might help to present this in a diagram:



It isn't hard to see why this case involves a violation of the positive access principle. Given the above description, there is a 1–1 correspondence between positions of the hand and what evidence I have.[27] So, if I know the set-up, the truth about what evidence I have 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, however, 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 construct an unmarked clock of the kind Williamson describes, and I ask my friend 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. Having given the instructions, I know that the clock will be set somewhere between 52 and 54, so the situation is represented by this simpler diagram:



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

[26] 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*. Plausibly, the point about limited discriminatory abilities applies here too.)

[27] 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.

Next, I take a look. If people actually like me, the hand will be set to 53, and so my evidence will only tell me that it is somewhere between 52 and 54, which I knew already. 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 is far enough away from the actual setting. But I knew that there was a fifty-fifty 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 its 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.[28]

This method may be 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, I don't care too much. Perhaps added evidence that I am popular is better when it's pointing me towards the truth; but given the desirability of self-confidence, I welcome it even when it is misleading.

This evaluation, however, is 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 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. But what else could my evidence be in the scenario described above? One possibility, following Stalnaker (2009), might be to say that we need to distinguish scenarios, not simply by the position of the hand, but also by my 'best

[28] To design the right clock, I need to know something about my margin of error. But I don't need to know its exact value; I only need to know of some d that I'm not reliable at discriminating positions that are d millimetres away from each other, but am reliable at discriminating positions that are $2d$ millimetres away from each other. A little bit of research and experimentation should enable me to discover such a d.

guess' about its position.[29] Plausibly, my best guess won't always match the actual position (that's just what it is, we might say, 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. In fact, it is now natural to understand talk of 'margins for error' as talk about the maximum distance between my best guess and the actual hand (under ordinary conditions).

Reflections on limited discriminatory capacities give no obvious reason to think that I can't always tell what my best guess is—after all, these guesses are forced to take discrete values, and might be able to feature in 'transparent' inferences in a way in which details of my perceptual representations arguably do not. So if my evidence is determined by my best guess, rather than by the actual position of the hand, our previous reason for taking the case to violate the access principle disappears.[30] Moreover, for exactly the same reason, the strategy for setting up a no-lose investigation will no longer succeed. For when the clock is set to 52, there is a good chance that my best guess will be that it is set to 51. But, since I know that my margin for error is less than 2 (I designed the clock to make sure it would be), and I also know that conditions are ordinary (if I didn't know this, I could never draw any conclusions about the actual position of the hand, and hence about my popularity), this is compelling evidence that the clock isn't set to 53. Since I also know that my friend would have set the clock to 53 if I were popular, it is compelling evidence that I am unpopular. Far from being a no-lose investigation then, the inquiry described might well end up establishing my unpopularity.[31]

[29] For a different access-friendly account, see Smithies (2012).

[30] Of course, I might not make a guess; if this happens, my evidence is presumably determined by what my best guess *would have been*, and one might worry that I'm not always in a position to know what that is. But it is important to keep track of temporal indexes here. I clearly don't know now what my best guess would have been if I had made one back when I looked at the clock; but that only shows that I can't tell now what my evidence was then, and even the access principle doesn't require that I be able to do so. Was I in a position to know at the time what my best guess would have been then? Maybe I was. After all, I could have found out merely by taking a guess and registering its value.

[31] For further elaboration of the 'best guess' model, see Cohen and Comesaña (2013); for criticism, see Hawthorne and Magidor (2010), Goodman (2013, p. 34) and Williamson (2013, pp. 80–3). Since I can't discuss how to respond to such criticisms here, gesturing towards this alternative construal of the case remains a promissory note.

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 that obviously could not succeed. In discussing these cases, I have moved freely between a desire to be more confident that I am popular and a desire for evidence of my popularity, on the assumption that (since I'm fairly rational) these go together. Counterexamples to access, however, might be thought to put pressure on this assumption. Arguably, when one is in an access-violating scenario, one isn't rationally required to become more confident in p even when one receives evidence for p.[32] Alternatively, even if one is rationally required to become more confident in response to evidence, one's doing so might still be entirely *unreasonable*, manifesting a disposition not generally conducive to conforming one's confidence to the evidence.[33] Either of these claims might allow us to explain why agents like us are unlikely to raise our confidence in response to the evidence we receive in these cases. The externalist 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, yet 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 us.

Whatever the appeal of such a strategy in response to other objections to externalism, it will not solve the problem presented here. For I 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. Moreover, this knowledge need not itself be unreasonable, since it can be the result of perfectly ordinary belief-forming dispositions. This

[32] This claim is somewhat analogous to Hawthorne and Stanley's (2008, pp. 580–5) 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 (2009, pp. 360–1) 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 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'.

[33] See Lasonen-Aarnio (2010); see also Hawthorne and Srinivasan (2013) and Williamson (forthcoming), who connect this to the idea that raising one's confidence may be *blameworthy*.

makes it hard to see how failing to raise my credence could be reasonable or rationally permissible in such a situation. Yet the strategies for biasing one's inquiries seem equally absurd in these only slightly more complicated cases.

The simplest sure-win investigation combines the strategies discussed in the clock and wall cases: I ask my friend to arrange both a wall and a clock in line with the instructions described above, and then look at one and then later at the other. This 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 just received additional evidence of popularity. Since my knowledge about the set-up can be entirely reasonable, the same will be true about my knowledge at the end of inquiry that I just received additional evidence of popularity. But if I *know* that my evidence supports my popularity at least to a specific degree x, and that knowledge is not itself unreasonable, then I would surely be both irrational and unreasonable not to have at least the corresponding level of confidence in the claim that I am popular. If that is true, however, I would be irrational and unreasonable not to become more confident of my popularity after looking at both the clock and the wall.[34]

The combined case just described is effective against anyone who is sympathetic to both the wall and clock cases. But the epistemological motivations for these two cases are quite different, and one might be inclined to accept one without accepting the other. So it is worth noting that we only need one kind of case to set up a sure-win investigation.

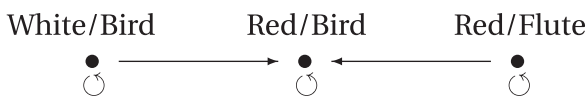
Let us look first at the cases motivated by our limited discriminatory capacities. The original instructions in the clock case were designed to generate no evidence either way when I'm popular and

[34] This style of argument also applies to responses, perhaps inspired by Gallow (2014), Bronfman (2014) and Schoenfield (forthcoming), which maintain that agents should 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 sometimes forbidding agents from raising 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 J. Dmitri Gallow for discussion.)

evidence that I'm popular when I'm not. But they are easily adapted to provide 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 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 I'm popular, and no evidence at all if I'm not. Alternating this strategy with the original one, 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 won't know what exactly that evidence was).

Let us next look 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 sophisticated bird flutes. And, plausibly, my abilities 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:[35]



What will happen to my evidence about my popularity? If I am popular, I will get conclusive evidence that I'm in Red/Bird, and thus that I am popular. If I am not popular, there are two possibilities. One

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

is that I am in White/Bird; in that case, my evidence will rule out Red/Flute and nothing else. The other is that I am in Red/Flute; my evidence will then rule out White/Bird and nothing else. So if I'm unpopular, my evidence will eliminate exactly one possibility in which I am unpopular and nothing else, meaning I 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.

Explaining the absurdity of the strategies described in §§2.1–2.2 by forcing a gap between evidence and rational (or reasonable) belief has some initial appeal. It is not entirely unnatural to think that, in some sense, 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 motivations driving the initial cases can be used to generate cases in which one not only receives evidence for one's popularity but knows that one does. Yet the strategies described in these cases seem equally absurd. The attempted alternative explanation of the absurdity thus fails.

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 (if somewhat oversimplified) counterexamples to the access principle by showing that, if genuine, they would enable us to intentionally bias our inquiries in highly counter-intuitive 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 counter-intuitive 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 Bayesian 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 Bayesian theory), and (ii) that if the access principle has any counterexamples, people should be able to exploit these to bias their 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 'reflection principles' widely discussed in formal epistemology.

3.1 Formalizing biased inquiry

Recall the discussion of §1.1. We were wondering whether you could embark on a course of action (for instance, 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 goal 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 claim that one cannot do this, 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[36]) 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 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.

[36] 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.

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.[37] It seems possible to imagine uncertainty about either of those factors: I can be unsure both about the colour of the thirty 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^+.[38] 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 the future and the initial evidential support if your total initial evidence is E_i and your total later evidence is E_j^+.

[37] Or, perhaps, what they support relative to the agent's 'standards' or 'inductive policies', if one is sceptical about an objective evidential support relation. Recall that, since changes in these standards or policies seem to be based on something other than the evidence one receives, ways of biasing one's future opinions which exploit such changes don't qualify as 'intentionally biased inquiry'; we thus lose no generality by ignoring such relativity in the support relation.

[38] 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.

Then we can write the claim that, conditional on any A_k, the expected difference is 0 as[39]

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

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.

Assuming that the E=Ei and E+=E+j are independent, we can rearrange and simplify (¬IBI) to yield

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

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

From this equation, we can make our way towards something more recognizable. 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

[39] 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 what impact those things would have on the proposition in question. Moreover, the connection between the rewritten principles and the access principle is a bit less straightforward.

(that is, 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,[40]

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

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

And *this* claim, stating that the expected future probability is equal to the expected current probability, is the obvious consequence of two instances of van Fraassen's (1984) reflection principle, one synchronic and one future-directed:[41,42]

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

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

[40] Proof:

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

[41] 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, for example, van Fraassen (1995, p. 19) and Weisberg (2007, p. 180)—though Williamson (2000, pp. 230–7) 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).

[42] Kadane, Schervish and Seidenfeld (1996) take 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.

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

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

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 sidestep objections to reflection arising from the fact that future credences might be formed a- or ir-rationally.[43] And, crucially, it means that (S-REF) is a version of Christensen's (2010) rational reflection principle, stating that the probability of a proposition matches the expected current evidential support. But it's well known that it's hard to reconcile rational reflection with denials of the access principle, and rational reflection is widely rejected for this reason.[44] 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 me, 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.[45]

[43] 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 are likewise not relevant for our purposes.

[44] See, for example, Christensen (2010), Williamson (2011), Elga (2013) and Lasonen-Aarnio (2015).

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

This is not to say that (EQEXP) is or should be neutral on whether the access principle is true. Williamson (2000, ch. 10) and Weisberg (2007) highlight a serious tension between (F-REF) and denials of the access principle;[46] 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 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.[47] 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 $E = E_j$.[48] 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 complex. 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

[46] See also Hawthorne (2004, pp. 75–7) and Weatherson (2011) for briefer discussion.

[47] We also need that the current evidence entails that all evidence is true and that there will be no information loss; I will suppress these assumptions henceforth.

[48] 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 $E = E_i$.

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).[49]

If the access principle holds in general, then, we would expect both (S-REF) and (F-REF) to be true. (EQEXP) follows trivially 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 of 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.[50]

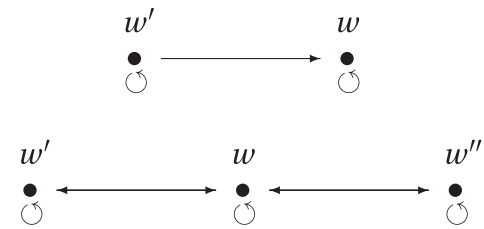
To settle our second question, and generalize the argument from the impossibility of intentionally biased inquiry to the access principle, we need something like the converse entailment: that (EQEXP) fails unless the access principle is true. Unfortunately, the connection is not quite so straightforward. For (EQEXP) holds trivially of an agent who is certain not to receive any evidence, regardless of whether she satisfies the access principle.[51]

[49] Note that, in guaranteeing the equivalence of E_j^+ and $E^+ = E_j^+$, the access principle ensures 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, for example, van Fraassen (1995, p. 17) and Briggs (2009, p. 69). Weisberg (2007, pp. 183–4) discusses the partitionality assumption in such proofs in detail, and concludes that it can only be motivated by something like the access principle.

[50] Kadane, Schervish and Seidenfeld (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.

[51] Elsewhere, 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.

There is, however, a less straightforward connection that is strong enough for our purposes. Suppose that there are possibilities in which the access principle 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 'doesn't rule out' relation between the possibilities will either fail to be symmetric or fail to be transitive.[52] 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 that 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 that, and only that, he is in one of w, w' and 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 that, 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 higher than its initial probability in both cases, regardless of which (non-zero) initial probabilities are assigned to each world. That's precisely why, in discussing the wall and clock cases, we thought it desirable to somehow associate w with the possibilities in which I am popular.

[52] 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 S_5 logic. And it's a well-known theorem of modal logic that a modal operator obeys an S_5 logic if and only if the corresponding accessibility relation is reflexive, symmetric and transitive.

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 argument for the access principle requires both the general recipe and its application to specific cases. If we have only the general recipe, it may not be obvious why we should 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 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. (So radical, perhaps, that some will be inclined to *modus tollens* my argument, and conclude that intentionally biased inquiry is possible after all. That would still be an interesting discovery.) We have already encountered several reasons to reject the access principle when motivating the alleged counterexamples 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 premisses: (i) rational agents can be wrong 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, by (i), a rational agent might believe p even though it is false, and thus fail to realize that, by (ii), p isn't part of her evidence. But if negative access 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 premisses is undeniable,[53] but both are intuitively appealing.[54]

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 be clearest if we contrast our discussion of reflection with Williamson's. For Williamson, in addition to being perhaps 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.[55] 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:

[53] Smithies (2012) would deny (i) by maintaining that ideally rational agents would never be wrong about their phenomenal states; Goldman (2009) seems to deny (ii).

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

[55] See also Hawthorne and Srinivasan (2013) for a similar picture of the access principle.

> 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. (Williamson 2000, p. 237)

The connection I've drawn between intentionally biased inquiry and reflection paints a less rosy picture of the principle: it imposes a frustrating limitation on our pursuit of desirable beliefs. The denier of access is the true optimist, promising us wonderful tools for shaping the outcomes of our inquiries, tools that we would love to employ. But we do not have these tools. We should save our composure for facing up to this, far more unfortunate, realization.[56]

References

Briggs, Rachael 2009: 'Distorted Reflection'. *Philosophical Review*, 118, pp. 59–38.

Bronfman, Aaron 2014: 'Conditionalization and Not Knowing that One Knows'. *Erkenntnis*, 79, pp. 871–92.

Christensen, David 2010: 'Rational Reflection'. *Philosophical Perspectives*, 24, pp. 121–40.

Cohen, Stewart 2002: 'Basic Knowledge and the Problem of the Problem of Easy Knowledge'. *Philosophy and Phenomenological Research*, 65, pp. 309–29.

Cohen, Stewart and Juan Comesaña 2013: 'Williamson on Gettier Cases and Epistemic Logic'. *Inquiry*, 56, pp. 15–29.

Elga, Adam 2013: 'The Puzzle of the Unmarked Clock and the New Rational Reflection Principle'. *Philosophical Studies*, 164, pp. 127–39.

Gallow, J. Dimitri 2014: 'How to Learn from Theory-Dependent Evidence; or Commutativity and Holism: A Solution for Conditionalizers'. *British Journal for the Philosophy of Science*, 65, pp. 493–519.

[56] Ideas from this paper have been presented at the 2012 ARCHE/CSMN graduate conference, the MIT Dissertation Workshop, the 2014 MITing of the Minds, the 2014 Formal Epistemology Workshop, and the 2014 Joint Session of the Aristotelian Society and the Mind Association. Thanks to the audiences and my commentators Stew Cohen, Declan Smithies and Jeff Dunn on these occasions. In addition, the paper has been improved greatly as a result of comments from Andrew Bacon, Dan Barras, Alex Byrne, Nilanjan Das, Kevin Dorst, Jane Friedman, J. Dmitri Gallow, Jeremy Goodman, Dan Greco, John Hawthorne, Brian Hedden, Sophie Horowitz, Abby Jacques, Brendan de Kenessey, Harvey Lederman, Jack Marley-Payne, Damien Rochford, Ginger Schultheis, Kieran Setiya, Jack Spencer, Bob Stalnaker, Josh Thorpe, Jonathan Vogel, Ian Wells, Steve Yablo, four anonymous referees, an associate editor for *Mind*, and, especially, Roger White.

Goldman, Alvin 1979: 'What is Justified Belief?' In George Pappas (ed.), *Justification and Knowledge*. Dordrecht: D. Reidel.
——2009: 'Williamson on Knowledge and Evidence'. In Patrick Greenough and Duncan Pritchard (eds.), *Williamson on Knowledge*. Oxford: Oxford University Press.
Goodman, Jeremy 2013: 'Inexact Knowledge without Improbable Knowing'. *Inquiry*, 56, pp. 30–53.
Hawthorne, John 2004: *Knowledge and Lotteries*. Oxford: Oxford University Press.
Hawthorne, John and Ofra Magidor 2010: 'Assertion and Epistemic Opacity'. *Mind*, 119, pp. 1087–105.
Hawthorne, John and Amia Srinivasan 2013: 'Disagreement without Transparency: Some Bleak Thoughts'. In David Christensen and Jennifer Lackey (eds.), *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press.
Hawthorne, John and Jason Stanley 2008: 'Knowledge and Action'. *Journal of Philosophy*, 105, pp. 571–90.
Hedden, Brian 2015: *Reasons Without Persons: Rationality, Identity, and Time*. Oxford: Oxford University Press.
Hintikka, Jaakko 1962*: Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.
Horowitz, Sophie 2014: 'Epistemic Akrasia'. *Noûs*, 48, pp. 718–44.
Kadane, Joseph, Mark Schervish, and Teddy Seidenfeld 1996: 'Reasoning to a Foregone Conclusion'. *Journal of the American Statistical Association*, 91, pp. 1228–35.
Kahneman, Daniel 2011: *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
Kelly, Thomas 2002: 'The Rationality of Belief and Some Other Propositional Attitudes'. *Philosophical Studies*, 110, pp. 163–96.
——2008 'Disagreement, Dogmatism, and Belief Polarization'. *Journal of Philosophy*, 105, pp. 611–33.
——2013: 'How to Be an Epistemic Permissivist'. In Matthias Steup, John Turri and Ernest Sosa (eds.), *Contemporary Debates in Epistemology*, 2nd edition. Malden, MA: Wiley Blackwell.
Kripke, Saul 2011: 'Two Paradoxes of Knowledge'. In his *Philosophical Troubles*. Oxford: Oxford University Press.
Lasonen-Aarnio, Maria 2010: 'Unreasonable Knowledge'. *Philosophical Perspectives*, 24, pp. 1–21.
——2015: 'New Rational Reflection and Internalism about Rationality'. In Tamar Szabo Gendler and John Hawthorne (eds.), *Oxford Studies in Epistemology*, Vol. 5. Oxford: Oxford University Press.

Littlejohn, Clayton 2013: 'No Evidence is False'. *Acta Analytica*, 28, pp. 145–59.
Parfit, Derek 2011: *On What Matters*, *Volume* Vol. 1. Oxford: Oxford University Press.
Popper, Karl 1961: *The Logic of Scientific Discovery*. New York: Science Editions.
Schoenfield, Miriam 2014: 'Permission to Believe: Why Permissivism is True and What It Tells Us about Irrelevant Influences on Belief'. *Noûs*, 48, pp. 193–218.
—— forthcoming: 'Conditionalization Does Not (in General) Maximize Expected Accuracy'. To appear in *Mind*.
Smithies, Declan 2012: 'Mentalism and Epistemic Transparency'. *Australasian Journal of Philosophy*, 90, pp. 723–41.
Sober, Elliott 2009: 'Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads'. *Philosophical Studies*, 143, pp. 63–90.
Stalnaker, Robert 2009: 'On Hawthorne and Magidor on Assertion, Context, and Epistemic Accessibility'. *Mind*, 118, pp. 399–409.
Titelbaum, Michael 2010: 'Tell Me You Love Me: Bootstrapping, Externalism, and No-Lose Epistemology'. *Philosophical Studies*, 149, pp. 119–34.
van Fraassen, Bas C. 1984: 'Belief and the Will'. *Journal of Philosophy*, 81, pp. 235–56.
—— 1995: 'Belief and the Problem of Ulysses and the Sirens'. *Philosophical Studies*, 77, pp. 7–37.
Vogel, Jonathan 2000: 'Reliabilism Leveled'. *Journal of Philosophy*, 97, pp. 602–23.
Weatherson, Brian 2011: 'Stalnaker on Sleeping Beauty'. *Philosophical Studies*, 155, pp. 445–56.
Wedgwood, Ralph 2002 'Internalism Explained'. *Philosophy and Phenomenological Research*, 65, pp. 349–69.
Weisberg, Jonathan 2007: 'Conditionalization, Reflection, and Self-Knowledge'. *Philosophical Studies*, 135, pp. 179–97.
White, Roger 2006: 'Problems for Dogmatism'. *Philosophical Studies*, 131, pp. 525–57.
Williamson, Timothy 2000: *Knowledge and Its Limits*. Oxford: Oxford University Press.
—— 2009: 'Reply to Stephen Schiffer'. In Patrick Greenough and Duncan Pritchard (eds.), *Williamson on Knowledge*. Oxford: Oxford University Press.

—— 2011: 'Improbable Knowing'. In Trent Dougherty (ed.), *Evidentialism and Its Discontents*. Oxford: Oxford University Press.
—— 2013: 'Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier Cases in Epistemic Logic'. *Inquiry*, 56, pp. 77–96.
—— forthcoming: 'Justifications, Excuses, and Sceptical Scenarios'. In Julien Dutant and Fabian Dorsch (eds.), *The New Evil Demon*. Oxford: Oxford University Press.

