Higher-Order Evidence as Undercutting Defeat

Alex Worsnip

To appear in Darren Bradley (ed.), *Bayesianism, Self-Doubt, and Humility*, Brill
Draft of April 2025

0. Introduction

Suppose, as often happens, that you get some evidence that some belief of yours is irrational. For example, suppose you believe that you have above-average teaching ability. And suppose you then learn (as is true) that people are generally prone to irrationally overestimate their own teaching abilities. Here's one thing that seems obvious: you should now at least somewhat increase your credence in the (higher-order) proposition *that your belief that you have above-average teaching ability is irrational*. So much is (mostly) uncontroversial in the contemporary epistemological literature on "higher-order evidence"—which includes, though is not exhausted by, evidence that your beliefs are irrational. More generally, evidence that some belief of yours is irrational should increase your credence in the (higher-order) proposition *that your belief is irrational*.[1] This is just a special case of the general principle that evidence for some proposition *p* should raise your credence for *p*, with a higher-order proposition (*that your belief is irrational*) substituted for *p* in both instances.[2]

The more controversial question is this: when you get evidence that some belief of yours is irrational, should this affect (specifically, decrease) your credence *in the (first-order) proposition that you believe*? (Or, to ask the question in terms of outright belief, does it provide at least *pro tanto* reason to give the belief up?) For example, when you get evidence that your belief that you have above-average teaching ability is irrational, should you decrease your confidence *that you have above-average teaching ability*?

One might think that it is almost equally obvious that it does. Intuitively, the thought that your belief may well be irrational provides some reason to doubt it. Yet philosophers on both sides of this

For helpful discussions on the topics of this paper, I'm grateful to Yuval Avnur and Jim Pryor. This paper picks up a theme that I was unable to develop with the space constraints of Worsnip (2023), but that was a part of some early presentations of that material; I echo my thanks to everyone who assisted me with that paper.

[1] Some (e.g. Titelbaum 2015) get close to denying this, holding that facts about evidential support are *a priori*, and that if your belief *is* supported by the evidence and hence is rational, it's always rational to you to believe it's rational, and (misleading) evidence that your belief is irrational can't affect this. Such views are, I think, highly implausible (Worsnip 2021: sec. 9.3). Moreover, even if they are correct, on a conception of evidence where evidence for *p* by definition affects the rational credence for *p* (at least under *some* envisigable circumstances), these views really deny the very possibility of evidence that your belief is irrational.

[2] Indeed, depending on the precise definition of 'higher-order evidence' that we adopt, it may be that evidence that your belief in *p* is irrational is not higher-order evidence *with respect to the (higher-order) proposition that your belief in p is irrational*. Specifically, if we define higher-order evidence *with respect to a proposition q* as evidence that bears on the rational status of one's belief in *q* (if at all) only in virtue of its bearing on some higher-order proposition about *p* or one's belief in it (Chen & Worsnip 2025), then evidence that your belief in *p* is irrational is not higher-order evidence *with respect to the (higher-order) proposition that your belief in p is irrational*, since it doesn't bear on one's belief in *q* only in virtue of bearing on a (still more) higher-order proposition about this proposition; rather, it bears on it directly. So the claim that is uncontroversial here arguably isn't a claim about higher-order evidence *qua* higher-order evidence.

debate have tended to agree that there is a puzzle about *why* evidence that your belief is irrational constitutes a reason to decrease credence in the proposition that you believe.

In brief, the alleged puzzle is this (it will be explained in more detail later). The fact that one person believes *p* irrationally doesn't on its own seem like very strong evidence, if it is any evidence at all, that *p* is false. After all, any true proposition can be believed irrationally (or for bad reasons). Now, it might be tempting to conclude immediately from this that evidence that your belief in some proposition is irrational doesn't constitute any reason to become less confident in that proposition. But this would be too quick. There's a well-known kind of evidence—that which puts into effect "undercutting defeat"—that (relatively uncontroversially) does not on its own indicate that not-*p*, but nevertheless makes it rational to be less confident in *p* than one was antecedently. Roughly speaking, undercutting defeaters make it rational to be less confident in *p* not by themselves supplying positive evidence that *not*-p, but by undercutting the rational import of one's existing reasons for believing *p*. So, if higher-order evidence could be assimilated to undercutting defeat, we'd have a clear explanation—one that can be subsumed under well-accepted general principles—of why it calls for a reduction in credence in the first-order proposition in play.

There's just one problem: everyone, opponents and proponents of the significance of higher-order evidence alike, seems to agree that higher-order evidence (or, at least, the kind of higher-order evidence that indicates irrationality) *can't* be assimilated to undercutting defeat.[3] Those who think that higher-order evidence isn't epistemologically significant (at the level of first-order propositions) take this to supply an argument for their view: higher-order evidence *neither* directly bears on the first-order proposition at issue *nor* constitutes an undercutting defeater, so it isn't epistemologically significant. Those who think that higher-order evidence *is* epistemologically significant, by contrast, take themselves to be in need of some *alternative* explanation of its significance beyond that of undercutting defeat. They reach for more complex explanations, for example in terms of a requirement to "bracket" evidence (Christensen 2010), non-evidential reasons to suspend judgment (Lord & Sylvan 2021), or a special kind of "dispossessing defeat" (González de Prado (2020).

The aim of this paper is to revive the possibility of—contrary to this broad consensus—simply assimilating higher-order evidence (of the sort that indicates irrationality) to undercutting defeat. The central idea is that once we accept a moderate kind of holism about evidential support, we see that the undercutting defeat explanation can be applied more widely than it first seems. If higher-order evidence can be seen as occasioning a kind of undercutting defeat, then one cannot reject its epistemic significance (for our first-order beliefs) without taking the very radical step of rejecting the phenomenon of undercutting defeat altogether.

1. Kinds of higher-order evidence

A persistent source of confusion in the literature is that there are numerous different phenomena that are grouped together under the label 'higher-order evidence', with importantly different epistemic properties. One such phenomenon is "evidence of evidence," where you learn (or get evidence that)

[3] See e.g. Feldman (2005: 111-3); Christensen (2010: 193-5); White (2010: 585); Lasonen-Aarnio (2014: 317-8); González de Prado (2020: 327-8); and Lord & Sylvan (2021: 124).

that there is some strong piece of evidence for (or against) p, without learning what that piece of evidence is.[4] The challenge about the epistemic significance of higher-order evidence that we're exploring in this paper simply doesn't apply to evidence of evidence. This is because, at least usually,[5] the fact that there is some strong piece of evidence for *p* itself provides some pro tanto evidence for *p*. So *this* kind of higher-order evidence clearly should affect your confidence, just because it is itself evidence for *p*. (And *mutatis mutandis* for evidence of evidence *against* p.) Thus, I will set aside evidence of evidence in what follows. Henceforth, when I talk of "higher-order evidence", I am implicitly restricting myself to the sort of evidence that doesn't itself provide evidence for or against p in this simple way.

The other main overarching category of higher-order evidence, besides evidence of evidence, is evidence that bears on one's epistemic performance in some way. This higher-order evidence might, for example, indicate that one's reasoning was flawed, that one was under the influence of a drug that impedes one's performance, that one employed an unreliable method, or that one succumbed to some sort of cognitive bias.[6] Even within this category though, there may be important divisions to be drawn. In particular, very plausibly, not all flawed epistemic performances or cognitive errors (in the broadest sense) are best understood as forms of *irrationality* per se. Most obviously, one might rely on a belief-forming method that is in fact reliable while having good reason to think it highly reliable. *Pace* a kind of unreconstructed reliability about rationality, beliefs formed by this method do not seem (automatically) irrational.[7]

This distinction turns out to be quite important for the challenge at issue in this paper, for (I will suggest) it is considerably easier to assimilate higher-order evidence of cognitive imperfection to undercutting defeat when it is *not* evidence of irrationality per se. Moreover, there are a lot of cases in which it is debatable whether a cognitive imperfection amounts to a form of irrationality. For example, it isn't clear to me whether we should say that making an error in mathematical calculation (automatically) constitutes irrationality. And (as I'll substantiate further in section 3), at least some of the "cognitive biases" that psychologists document may not involve irrationality per se, rather than some other kind of imperfection (at least in absence of knowledge on the fact of the agent that they are operating).

In the case with which I opened this paper, I simply stipulated that what you gain is evidence that people in general (and hence, insofar as you have no reason to think yourself special, you) are prone to *irrationally* overestimate their teaching abilities. And I do want to eventually defend the view that even evidence of *irrationality specifically*—the hardest kind of case to explain the epistemic significance of—can be assimilated to undercutting defeat. With that said, it's also worth noting that the more of the standard cases that we interpret as involving evidence of cognitive flaws that *don't* amount to irrationality, the narrower the range of cases of higher-order evidence to which the toughest challenge (with regard to explaining their epistemic significance) applies.

---

[4] Worsnip (2018) and Eder and Brössel (2019), for example, describe evidence of evidence as "higher-order evidence."
[5] See Fitelson (2012) for some putative exceptions.
[6] There can also be higher-order evidence indicating that one performed *well* epistemically. This is often neglected in the literature (see Matthews ms. for an honorable exception), and unfortunately this paper isn't going to buck that trend. How to generalize what I say here to these cases is well worthy of exploration.
[7] Notably, the arch-reliabilist Alvin Goldman stopped short of endorsing reliability about *rationality*, distinguishing it from *justification*, which he did want to construe along reliabilist lines. See Goldman (1986).

2. The case for Anti-Significance

Let's call the view that higher-order evidence of the sort that indicates irrationality is not evidentially significant (with respect to the first-order belief in question) **Anti-Significance**. On the standard way of developing Anti-Significance, whether you should hang on to some belief of yours turns not on what your (higher-order) evidence indicates about whether your belief is supported by your first-order evidence, but *solely* on whether it *is* supported by your first-order evidence.[8] I've already very roughly sketched the case for this view in the Introduction, but let's reconstruct it in a bit more detail. Let's again work with our opening case for concreteness. An initial argument for the Anti-Significance view about that case goes as follows:[9]

Initial Argument.

1. You should hold on to a belief iff it is supported by your total evidence.
2. But the fact that people tend to overestimate their teaching ability does not affect whether the belief that you have above-average teaching ability is supported by your total evidence.

Therefore,

3. The fact that people tend to overestimate their teaching ability does not affect whether you should hold onto your belief that you have above-average teaching ability.

Two clarifications, for the avoidance of all doubt.

First, this argument does not, of course, say that you should under no circumstances should give up the belief that your teaching ability is above-average. If you have *in fact* overestimated your teaching ability relative to your first-order evidence, such that your first-order evidence does *not* support your belief that your ability is above average, then you should give the belief up. The claim is simply that what matters is whether your first-order evidence *actually* supports this belief, not whether you have higher-order evidence (such as the fact that people tend to overestimate their teaching ability) that *indicates* that your first-order evidence doesn't support this belief.

Second, and relatedly, the argument does not beg the question in assuming premise (1). (1) says that whether you should hang on to a belief depends on whether it is supported by your *total* evidence (not: whether it is supported by your *first-order* evidence). This premise, in itself, leaves open whether your higher-order evidence affects what your total evidence supports. Premise (2) says that it doesn't, at least in this case. The proponent of the epistemic significance of higher-order evidence can deny either (1) or (2). While some of them may *end up* denying (1), this isn't a commitment of their view from the off—and denying (1) is undeniably costly, since (1) is orthodoxy.

But why accept premise (2)? One initial line of thought is something like this. Suppose (to simplify matters) that everyone believes they are an above-average teacher, and that in all cases this is produced by irrational, biased, motivated, cognition. Still, presumably, approximately (or just slightly

[8] See e.g. Titelbaum (2015) and Tal (2020). Lasonen-Aarnio (2014) and Weatherson (2019) hold related, but more complex, views.

[9] I formulate it here in terms of outright belief, but the same argument could be run in terms of credences.

fewer than) 50% of teachers *are* in fact above-average. With this in mind, the fact that I irrationally believe that I'm above-average doesn't seem to be any evidence that I'm in the group that isn't above-average, rather than the group that is. Generalizing: both true beliefs and false beliefs can be (and are often) held irrationally. Learning that a belief is irrational thus doesn't seem to be much, if any, evidence that it is false (or true).

To sharpen the point, consider the difference between someone's being *unreliable* (in the sense that their beliefs are not correlated with the truth) and their being *anti-reliable* (in the sense that their beliefs are *inversely* correlated with the truth). Even a random belief generator would, at least with respect to yes/no questions, be merely unreliable rather than anti-reliable. Now, in general, unless we (know that we) live in a world where evidence is misleading *more often than not*—like, say, a world in which we're all brains in vats—learning that someone's beliefs are rational will suggest that they are generally reliable. But learning that their beliefs are *ir*rational does not obviously indicate that they are *anti*-reliable, rather than merely *un*reliable. (After all, someone who forms beliefs on yes/no question by tossing a coin would be irrational, but only unreliable rather than anti-reliable.) Correspondingly, it seems like when we learn that some particular belief is irrational, we don't gain much reason to think that it's false.

However, this case for premise (2) is incomplete. As I've presented it so far, it looks like it tacitly assumes something like the following:

> **Only Direct Evidence Matters.** The only way for some evidence to affect whether your belief in p is supported by your total evidence is by *itself* constituting evidence that p is true or false.

But **Only Direct Evidence Matters** is false. In particular, it is shown to be false by the well-known phenomenon of "undercutting defeat."[10] Undercutting defeat occurs when some piece of evidence affects whether your evidence supports believing *p*, or lowers the rational credence for *p*, not by constituting positive evidence against *p* but rather by undercutting the force of your existing evidence for *p*. Here are two examples, one famous and another slightly less so:

> **Red Light.**[11] You believe that the wall appears red (R), on the basis of its appearing red (A). Normally appearances as of X justify believing that X obtains, so your belief in R is rational. But then you discover that the wall is being illuminated by a red light (I). I is not itself evidence for or against R. (After all, it's not like the fact that a wall is being illuminated by a red light suggests the wall is *not* red—or that it is.) Nevertheless, the discovery of I means that your total evidence no longer supports believing R (and that your credence for R should drop). This is because the discovery of R makes it the case that A no longer supports R, or at least drastically reduces the support that A gives to R.

[10] Pollock (1970) first introduced the notion of undercutting defeat, though not originally using that label.
[11] The example is also due to Pollock (1970).

Pathological Liar. You believe that your casual acquaintance Jeff was born in Dublin (D), on that basis of his testimony that he was born in Dublin (T). Normally, someone's testimony about their own life history justifies you in believing them – since people don't usually lie about mundane aspects of their life history – so your belief in D is rational. But then your extremely trustworthy friend Sara, who knows Jeff well, tells you that Jeff is a pathological liar who frequently (but not always) lies about random aspects of his life history (L). L is not itself evidence for or against D. (After all, it only says that Jeff *frequently* lies, not that he does always or the majority of the time.) Nevertheless, the discovery of L means that your total evidence no longer supports D [and that your credence for D should drop]. This is because the discovery of L makes it the case that T no longer supports D, or at least drastically reduces the support that T gives to D.

Pace some very extreme versions of epistemic externalism,[12] these examples show that **Only Direct Evidence Matters** is false. Note that nothing here involves denying the evidentialist orthodoxy that it's rational to believe what your total evidence supports (i.e., premise (1) of the argument for Anti-Significance). Rather, the point is that *whether your total evidence supports* p can be things other than evidence for or against *p* – namely, undercutting defeaters for other (would-be) evidence for or against *p*.

But this doesn't mean it's game over for the defender of premise (2)—and, more generally, Anti-Significance. Suppose it's conceded that cases of undercutting defeat show **Only Direct Evidence Matters** to be false. Still, the following principle might be true:

Undercutting Defeat is the Only Exception. The only way for some evidence to affect whether your belief in p is supported by your total evidence is by *either (a)* itself constituting evidence that p is true or false; *or* (b) constituting an undercutting defeater for (would-be) evidence for *p*.

This principle is more plausible. Indeed, it's tacitly assumed in some of the literature on defeat, in which it's often taken for granted that new evidence that defeat the justification of a belief divides into (only) two categories: rebutting defeat (a above), and undercutting defeat (b above).[13]

If this principle is true, the question then becomes: can higher-order evidence (of the sort that indicates irrationality) be understood as giving rise to undercutting defeat? And here, the consensus in the literature—both among proponents and opponents of the significance of higher-order evidence—is that it can't be: there are deep differences between undercutting defeaters and higher-order evidence (of the sort that indicates irrationality) that make it impossible to assimilate the latter to the former.[14] I'll say what the alleged differences are in the next section, but for now, let's notice that if this claim

[12] Lasonen-Aarnio (2010, 2014) explores a view of this kind.
[13] Again see Pollock (1970) and the large literature that follows it.
[14] See the references in n. 3 above. Avnur & Scott-Kakures (2015) are a partial exception; see section 4 for a discussion of their view.

is correct, it puts us in a position to state an ancillary argument for premise (2) of the initial argument for Anti-Significance:

Ancillary Argument.

4. **Undercutting Defeat is the Only Exception**.
5. The fact that people tend to overestimate their teaching ability *neither* (a) itself constitutes evidence that you do or don't have above-average teaching ability *nor* (b) constitutes an undercutting defeater for your (would-be) evidence for this claim.

Therefore,

6. The fact that people tend to overestimate their teaching ability does not, in itself, affect whether the belief that you have above-average teaching ability is supported by your total evidence.

In light of this consensus that higher-order evidence is not assimilable to undercutting defeat, those who have wanted to affirm the epistemic significance of higher-order evidence (with respect to our first-order beliefs) have generally fallen into one of two categories. One group can be understood as denying **Undercutting Defeat is the Only Exception**, arguing that even though higher-order evidence of the kind we are considering does not give rise to undercutting defeat (or itself constitute evidence for or against the first-order proposition in play[15]), it nevertheless affects our what our total evidence supports in some other way.[16] A second group can be understood as accepting premise (2) of the initial argument, and the ancillary argument for it, but denying premise (1) of the initial argument, namely that you should hold on to a belief iff it's supported by your total evidence.[17]

I want to explore a third route to defending the significance of higher-order evidence of the sort that indicates irrationality—namely, to claim that such higher-order evidence *does*, after all, give rise to undercutting defeat. This view, then, denies premise (5) of the ancillary argument, and with it, premise (2) of the initial argument.

2.1 The case against assimilation

Before that, though, let's consider the reasons why so many philosophers have thought that higher-order evidence of the sort that involves evidence of irrationality cannot be assimilated to the phenomenon of undercutting defeat. The helps to complete my opponent's case for Anti-Significance (by shoring up premise (5) of the ancillary argument).

The main idea, I think, is this. Return to paradigmatic cases like Red Light. In this case, *before* you learn that the wall is being illuminated by a red light (I), the idea goes, it is completely rational for

[15] Again, the parenthetical here must be qualified to restrict us to higher-order evidence *that indicates one's irrationality*. As noted in the introduction, "evidence of evidence" *does* (often) constitute evidence for or against the first-order proposition in play.

[16] A paradigm example is Gonzalez De Prado (2020), who thinks that higher-order evidence gives rise to another, distinctive kind of defeat called "dispossessing defeat".

[17] E.g., Christensen (2010) argues that the result of higher-order evidence is that we're required to "bracket" some of our first-order evidence, which – it seems – will lead us to believe only what the non-bracketed *subset* of our evidence supports, not what our *total* evidence supports.

you to believe that the wall is red (R) on the basis of its appearing red (A). Indeed, for the person who hasn't yet learned I, A *really does* evidentially support R. By contrast, once you *have* learned I, it is no longer rational for you to believe R on the basis of A. The evidential support relation between A and R—which, again, really did hold prior to your learning I—is severed. The undercutting defeater—as the name suggests—*undercuts* an evidential support relation that was previously, or would otherwise, be present. Moreover—and this is really just a corollary of the point just made—learning I should not bring you to think that you were *already* being irrational (in believing R on the basis of A) before you learned I. It should only bring you to think that you would be irrational if you *continued* to believe R *now*.

But, the thought continues, cases of higher-order evidence of irrationality are by their nature not like this. When you get higher-order evidence that people are prone to *irrational* overconfidence in their own teaching abilities, for example, you get (pro tanto) reason to think that your belief that you are an above-average teacher was *already* irrational, even before you received the higher-order evidence in question.

This is already (indisputably) a difference with the Red Light case. But why does it (allegedly) make the framework of undercutting defeat inapt? To make things easier, let's call the first-order evidence you have relevant to your teaching abilities (your teaching evaluations, the success (or lack of) of your students, and so on) F; the proposition that you are an above-average teacher T, and the (true) proposition (or fact) that people tend to irrationally overestimate their teaching abilities O. Remember that uncontroversially, O provides at least some pro tanto evidence for the *higher-order* proposition that your belief in T is irrational (call that higher-order proposition H). What we are considering is whether it in doing so, it *undercuts* the justification you previously had, or would have had, for believing the *first-order proposition* T.

Here's an argument by cases that it doesn't.[18] Either H is true (your belief in T actually is irrational), or H is false (your belief in T isn't irrational). In the former case, O (insofar as it supports H) is indicating a truth, whereas in the latter case, O (insofar as it supports H) is misleading.[19] Let's take these possibilities one by one. In the former case, F *doesn't* actually support (your belief in) T, and never did, even before you learned O. In this case, there is no prior evidential support relation between F and T to be undercut, so we can't have undercutting defeat. In the latter case, F *does* support your belief in T. But in this case, the argument goes, O does not *sever* or *undercut* this support relation (in the way that the discovery of the illumination of the wall severs the support relation between the appearance as of red and (belief in) the proposition that the wall is red). Notwithstanding your learning O, it continues to be true that your first-order evidence (your evaluations, student successes, etc.) support (believing) the proposition that you're an above average teacher.[20] Your higher-order evidence in this case *misleadingly indicates* the F does not support T; but that is not the same as *severing* or

[18] The argument bears some resemblance to that from Tal (2020), though it is not quite identical.

[19] Note that the latter case isn't one where O is false—where it's not true that *people in general* are prone to irrationally overestimate their teaching abilities. We're supposing as part of the set-up of both variants of the case that O is something you learn, such that it is true. Rather, the latter case is one where, though this is true of people in general, as a matter of fact you personally buck this trend, and your belief in T is not a product of irrational overconfidence.

[20] Cf. (e.g.) White 2010: 585; Weatherson 2019: 136.

undercutting this relation of support, such that F *really doesn't* support T. So here again, we don't have undercutting defeat.

Hence, the argument goes, in neither case does learning O sever or undercut an evidential support relation between F and T in the way that is characteristic of undercutting defeat. Either there *never was* such a relation, in which case there is nothing to be undercut (and O merely draws attention to this antecedent fact), or there *is* such a relation, but it persists despite your learning O (and O merely *misleadingly suggests* that there isn't such a relation). This, I think, is what makes many philosophers think that we'll *either* need to find a different mechanism (other than undercutting defeat) by which higher-order evidence (of this kind) is significant at the first-order level, or give up the idea that it is.

3. Evidence of imperfection without irrationality

Before I turn to defending the view that evidence of irrationality *can* after all be assimilated to undercutting defeat, I want to highlight the narrowness of applicability of the argument against this that I just presented. Recall from section 1 the distinction between higher-order evidence of a cognitive flaw that *does not* amount to irrationality and higher-order evidence of irrationality per se.[21] Let's consider whether the argument against assimilating the latter kind of higher-order evidence to undercutting defeat generalizes to the former kind of higher-order evidence. This is both interesting in its own right and will help to set up my eventual argument that even the kind of higher-order evidence that indicates irrationality can ultimately be assimilated to undercutting defeat after all.

The answer is that the argument does not generalize. The basic reason is simple. The argument that higher-order evidence of irrationality doesn't occasion undercutting defeat turned on the point that, when the higher-order evidence in question is not misleading, the belief is *already* irrational before you received the higher-order evidence—hence, there is no justification or evidential support relation to undercut. But in the case where the flaw that you get evidence of doesn't amount to a kind of irrationality, this isn't so. Hence, there's room for the higher-order evidence to *make* an antecedently rational (albeit flawed) belief irrational, just as the undercutting defeat account suggests.

To illustrate this further, let's consider some examples, drawing from psychological work on cognitive biases. Discussing these examples will enable me to do two things at once. First, it will enable me to make the case that a range of cases that are often taken to be cases of irrationality may be better diagnosed as cases of imperfection without irrationality. Second, it will enable me to illustrate why higher-order evidence of imperfection without irrationality doesn't resist assimilation to undercutting defeat.

The so-called hot hand "fallacy" involves assuming that someone is more likely to succeed in an attempt at some feat (for example, shooting a basketball) if their very recent attempts at the same feat succeeded (Gilivoch, Vallone & Tversky 1985). Whether this assumption is in fact empirically mistaken (in sports, at least) is now actually somewhat contested (Green & Zwiebel 2018). But let's just grant for the sake of argument that as a matter of empirical fact, it is mistaken, and that rate of

[21] I also discussed, and set aside, evidence of evidence. Evidence of evidence does not occasion undercutting defeat, but as I've already said, it evades the challenge about the epistemological significance of higher-order evidence in a different way—namely by itself constituting evidence for or against the first-order proposition in play.

success in very recent shot attempts is not a predictor of success in a player's next attempt.[22] Even if this is so, it's clearly a contingent empirical fact of which one could be unaware. Moreover, if one has no access to the empirical data, it seems rational (or at the very least, not obviously *ir*rational) to expect that success in very recent shot attempts *would* be a predictor of success. After all, a plausible folk-psychological theory is that hitting shots gives one confidence, and that confidence helps one to succeed in the next attempt—and conversely, that having missed recent shots gets one demoralized, and this leads to further failure.[23]

Suppose you now discover that this folk-psychological theory is empirically false, at least in the case of basketball. In that case you discover that a particular method for predicting the probability of shot-success is unreliable. But provided that you were *antecedently* rational in taking the method to be reliable, learning that this method is unreliable is *not*—or at least not obviously—an instance of learning that you were irrational all along. Rather, like the Red Light case, it is a case where given what you learn (viz. that the method is unreliable), it *becomes* irrational to continue relying on it by holding onto your belief. The difference with the Red Light case is that in the Red Light case, the method you are using (relying on appearances of redness as a guide to actually redness) is *usually* reliable, and you learn only that it is unreliable in your current (as it turns out, atypical) circumstances, in virtue of the unusual condition that the wall is being illuminated by a red light. By contrast, in the hot-hand case, we're supposing, you learn that (at least with respect to basketball) success in very recent attempts is more generally a poor guide to success on the next attempt. But this difference is inessential to the diagnosis of the cases in terms of undercutting defeat. What is crucial to that diagnosis is that the undercutting defeater makes it the case that you're no longer rational to rely on the method that you were previously rational to rely on, not whether the reason for this is that you learn that it's unreliable in your specific circumstances or more generally.

Thus, I think, the hot hand case can be diagnosed in terms of undercutting defeat. On this diagnosis, provided that you're rational in taking very recent success to be reliable a guide to success on the next attempt (prior to learning that it isn't), the fact that a player succeeded on recent attempts

---

[22] Or at least, that insofar as it is a predictor, it's "screened off" by the player's overall rate of success over a longer period of time. For example, perhaps a player who hits 40% of their overall shots is 40% likely to make their next shot, regardless of their rate of success in their most recent shots. (The complication is that a player who makes 40% of their shots is more likely to have made their most recent shots than a player who makes 20% of their recent shots. Thus, insofar as the rate of success in recent attempts is correlated with overall rate of success, success in recent attempts will also (to some degree) predict future success in the absence of data about the overall rate of success. But once we have that data, it "screens off", or renders probabilistically irrelevant, the rate of success in recent attempts as an *additional* predictor.)

[23] I suspect that one thing that encourages thinking that the hot hand reasoning really is a "fallacy" is the temptation to interpret it as based on some kind of supernatural thinking, according to which what underlies the assumed hot hand effect is something like it's being one's "lucky day". But there's just no reason why the assumed effect would need to be explained this way. It could instead be explained by the entirely naturalistically respectable psychological hypothesis that we often perform better when we're feeling confident than when we're feeling demoralized, and that success breeds confidence and failure demoralization. However, there are contexts other than basketball in which this alternative explanation doesn't work, namely those where success is entirely a matter of luck and hence couldn't be affected by the person's psychological state of feeling confident or demoralized. Certain forms of gambling might be like this. Here, the only way to explain a "hot hand" effect would be via the kind of supernatural "lucky day" kind of explanation, and hence the reasoning seems closer to a genuine fallacy.

really does, for you, evidentially support believing that they'll succeed on their next attempt.[24] Once you learn that it isn't a reliable guide, however, this evidential support relation is severed or undercut.

I think that a similar diagnosis may hold with respect to numerous other alleged "cognitive biases". For example, the "Barnum effect" occurs when you take the fact that (for example) an astrological descriptor describes you well to be evidence that astrology is accurate, when in fact the descriptor could be construed as describing *many* people well (Meehl 1956). Again, though, it seems like you could be rationally ignorant of the empirical fact that the descriptors can be construed as describing many people well, and in the absence of this knowledge, I don't think it's obviously irrational to take the fact that the descriptor fits you well to be (some, pro tanto) evidence of astrology's accuracy.[25] Once you learn that the descriptors fit many people well, though, this evidential support relation is undercut.

More arguably (and perhaps controversially), something like this may be true with at least some forms of implicit bias. Consider the widely-known result that people are inclined to evaluate CVs differently based on the racial associations of the name at the top of the CV (Bertrand & Mullainathan 2004). It's part of the nature of this bias that people generally aren't aware that it is affecting them, when it is. Now, does the fact that people are affected by this bias show that the beliefs that result (about how good or hireable the candidate is) are irrational, even if they don't know they are affected by the bias (or, to make things cleaner, even if they don't know that thifs is a common bias in general)? That is a tricky question. It depends whether facts like *this seems like a strong (weak) CV to me* themselves serve as evidence for the proposition that it is a strong (weak) CV.[26] If they do, it's not a given that the belief in question is irrational when the bias is unknown. But—and this is the crucial point—once the person knows that they are or might be affected by the bias—this evidential support relation is undercut. Indeed, this seems like a paradigm case of undercutting. Just as in the Red Light case you get reason to think that determining whether the wall appears red to you isn't a reliable method for determining whether it is red, here you get reason to think determining which CV seems better to you isn't a reliable method for determining which is better. Appearances aren't a good guide to reality in either case.[27]

---

[24] Objection: this conflates whether you're rational to believe that E evidentially supports P with whether E really does, for you, evidentially support P. Reply: no; what I'm suggesting is subtly different. I'm suggesting that (at least usually), when you're rational to believe that E *is a reliable indicator* that P, E really does, for you, evidentially support P. This only amounts to the alleged conflation if we presume a reliabilist conception of evidential support according to which whether E evidentially supports P is a matter of whether E is a reliable indicator that P.

[25] Of course, this is compatible with there being stronger countervailing reasons not to believe in astrology's accuracy (such as the lack of a plausible causal mechanism by which the date of a person's birth determines their personality, or by which the position of the stars affects what happens to them), such that the belief in astrology's accuracy is still all-things-considered irrational. But the Barnum effect can be generalized to less supernatural contexts where there may not be such strong countervailing reasons.

[26] The claim need not be that such seemings are the *only* evidence for such propositions. Hence, we're not committed to the tendentious (and, I agree, implausible) thesis that Brian Weatherson (2019: 188) calls "Judgments Screen Evidence" here.

[27] Indeed, one might even say that most real-world versions of our opening case, where someone gets evidence that they're overestimated their teaching abilities, are ones where—contrary to the stipulation I made in introducing the case—they don't get evidence that they've *irrationally* overestimated their teaching abilities, precisely because similarly, the fact that you seem to yourself to be a good teacher is some evidence that you are; and this evidential support relation is undercut. Still, it's surely at least *possible* that your evaluation of your own teaching ability is an irrational response to your total evidence

It's worth noting that a similar treatment can generalize to any case where a kind of seeming serves as evidence for *p*. For example, in moral or otherwise philosophical cases where the relevant evidence for one's belief consists partly of intuitions,[28] evidence that one's (apparent) intuitions are not truth-tracking (say, because they are the product of one's upbringing or interests, as in "irrelevant influence" cases[29]) will be an undercutting defeater that severs the evidential support relation between one's intuitions and one's beliefs that would otherwise have obtained and hence would have (ceteris paribus) justified one's belief. Again, this seems quite analogous to the Red Light case. Of course, it's *possible* that (apparent) intuitions produced by one's upbringing or interests just so happen to coincide with the truth, just as it's *possible* that the wall happens to actually be red even though its appearance as red is overdetermined by the red light illuminating it. Nevertheless, just as the discovery that the appearance of redness is due to a red light undermines the evidential support relation between the appearance of redness and the proposition that the wall is really red, the discovery that an (apparent) intuition of an act's moral wrongness is due to one's upbringing or interests undermines the evidential support relation between this intuition and the proposition that the act is really wrong.

To repeat, we've learned two things in this section. First, we've learned that a number of cases where you gain higher-order evidence of cognitive imperfection plausibly *aren't* cases where you gain higher-order evidence of irrationality. Second, we've learned that cases where you gain higher-order evidence of cognitive imperfection without gaining evidence of irrationality do not resist assimilation to undercutting defeat. Putting these two points together, we get the result that the argument against assimilating higher-order evidence to undercutting defeat applies in fewer cases than we might have first thought.

Still, there are surely some cases where we do get higher-order evidence of irrationality specifically, and we still need to know whether *these* can (contra the argument surveyed in section 2.1) be assimilated to undercutting defeat. It's to this that I turn next.

4. Evidence of irrationality

Before I get to my own view about why evidence of irrationality effects undercutting defeat, I want to consider a different view due to Avnur & Scott-Kakures (2015). Avnur & Scott-Kakures hold that evidence that one's belief that *p* is irrational defeats one's *doxastic* justification for believing *p* (if one has such doxastic justification to start with). Classically, to say that one is *propositionally* (or *ex ante*, or *prospectively*) justified in believing *p* is roughly to say that one has evidence that, *were* one to base one's belief on it properly, would result in a token justified belief; to say that one is *doxastically* (or *ex post*, or *retrospectively*) justified in believing *p* is to say that one actually has a token belief in *p* that is properly based on that evidence in such a way. Avnur & Scott-Kakures hold that even when one's total evidence in fact supports (believing) *p* (and hence belief in *p* is propositionally justified), evidence that one's

(and the same holds for the CV case). So it will still be important to consider, in the next section, whether that version of the case can be assimilated to undercutting defeat.

[28] Cf. e.g. Pust (2000).

[29] See e.g. White (2010), Avnur & Scott-Kakures (2015), and Vavova (2018) for discussion.

belief in *p* is irrational precludes one from being able to base one's belief in *p* on that evidence in the "proper" way and hence precludes one from being doxastically justified.

I think this falls short of what those of us sympathetic to the epistemological relevance of higher-order evidence would like to say. In at least some cases where one has powerful evidence that one's belief is irrational, I think we want to say that one *should give up* that belief. But it's not clear that Avnur & Scott-Kakures can vindicate this verdict. It's orthodoxy that a doxastic attitude is doxastically justified only if it's propositionally justified: doxastic justification requires propositional justification *plus more* (viz., proper basing on what propositionally justifies). But now suppose that belief in *p* is the *only* propositionally justified attitude on one's evidence.[30] In that case, since doxastic justification requires propositional justification, no other attitude can be doxastically justified either. So it doesn't seem like we can categorically say that one ought to give up one's belief: though one's belief in *p* is not doxastically justified, no other coarse-grained attitude that one could take toward *p* (suspension of judgment, for example) is doxastically justified either.[31] Rather, the result seems to be that one is in a sort of dilemma.

More boldly, I want to explore the idea that evidence of irrationality undercuts propositional justification (just as it does in standard cases of undercutting defeat, such as Red Light). Let's briefly recall the case against this contention that I set out in section 2.1. The thought was this: either your first-order evidence doesn't support your belief that you're an above average teacher, or it does. If it doesn't, there's no evidential support relation to be severed; if it does, the discovery that people tend to overestimate their teaching abilities doesn't sever this relation; the evidence of your superior skills still supports the proposition that you're above-average. Let's start with the latter case. My contention is that the argument just given rests upon an insufficiently *holistic* conception of evidential support.

To bring out the relevant kind of holism, let's consider a different case. Take a controversial and complex political issue such as whether an increase in the minimum wage would increase unemployment. Let's suppose just for the sake of argument that there is some empirical data that can be seen by a qualified expert to conclusively establish that an increase in the minimum wage would *not* increase unemployment. Now suppose you put that (raw) data in front of a total non-expert who lacks the background knowledge to competently understand it and draw conclusions from it. Does this person's total evidence support believing that an increase in the minimum wage would increase unemployment? I think not—at least not in any sense of evidential support that is tightly tied to rationality, such that this person would be irrational for failing to believe what her total evidence supports.

What this shows, I think, is that merely having access to the empirical data that supports a conclusion *in the hands of an expert* doesn't suffice for *your* total evidence supporting that conclusion (for you). You also need to have the background knowledge and competence to evaluate the data and correctly figure out what it supports. This is the respect in which evidential support is holistic: the very

[30] If the "uniqueness thesis" (White 2005) holds, this will be true *whenever* one's total evidence supports believing *p*.

[31] It might be replied that in such a case, one should take no attitude whatsoever toward *p* (where this is distinct from suspension; cf. Friedman 2013). But while I agree that taking no attitude whatsoever toward *p* is possible, and that it is distinct from suspension, I don't think that taking no attitude whatsoever is a deliberative option when one is actively deliberating about whether *p* (as opposed to when one has never considered whether *p*). I think that if one actively deliberates about whether *p* and can't make up one's mind, one thereby counts as suspending judgment.

same slice of evidence can have evidential significance for one person but not another, depending on their other background knowledge and capacities.[32] Moreover, evidence that is what we might call an "objective indicator" of the truth of some proposition *p*—in the sense that it supports *p* from the position of somewhat who has *all* the relevant background knowledge and capacities—may fail to support *p* for someone who lacks that relevant background knowledge or those capacities.[33]

Now let us return to our central case, where you have evidence that you have irrationally overestimated your teaching abilities. Recall that in the variant of the case we're currently considering, we've stipulated that this evidence is misleading, and that you in fact *haven't* irrationally overestimated your teaching abilities. For this reason (and perhaps others), the case is not perfectly analogous to the case in which you have the data about the minimum wage but aren't competent to evaluate it. By hypothesis you *have* correctly evaluated your first-order evidence concerning your teaching abilities; it's merely that you have *reason to think* that you haven't, and more generally that you have a bias that interferes with your capacity to do so.

Nevertheless, with the points about the holistic nature of evidential support that we've garnered from the minimum wage case in mind, let's reconsider the contention—from the opponent of assimilating evidence of irrationality to undercutting defeat—that just because you gain reason to think you're (irrationally) biased, it doesn't cease to be the case that your first-order evidence supports the proposition that you're an above-average teacher. What is certainly true is that your first-order evidence (your teaching evaluations, emails of thanks from students, concrete evidence of their progress in your classes, and so on) don't cease to be an *objective indicator* that you're an above-average teacher. But as the minimum wage case shows, that doesn't suffice for it supporting (still less, supporting *to a degree sufficient to justify outright belief*) the proposition that you are an above-average teacher *for you*. This opens up the conceptual space to say that the evidence that you have irrationally overestimated your abilities is an undercutting defeater: namely, if it obviates or diminishes the force of the first-order evidence *for you* in justifying the belief that you're an above-average teacher.[34]

And, indeed, this is precisely what I think we should say, if we are trying to assimilate evidence of irrationality to undercutting defeat. The view, then, is that it's not just your *actual* incompetence in handling the first-order evidence that prevents it from having the epistemic significance it would otherwise have (as in the minimum wage case); *evidence* of incompetence can also have this same effect. That is, propositions about your own competence or incompetence in assessing some kind of first-order evidence are among the background base of information that holistically determines whether that first-order evidence supports a belief for you, or not.

---

[32] Cf. Weatherson (2019: 133-137). Weatherson argues we can explain everything with reference to background evidence (such that capacities do not need to be invoked, or are reducible to background evidence), such that this point doesn't violate any kind of evidentialist principle.

[33] Cf. Kelly's (2014) distinction between "normative evidence" and "indicator evidence"

[34] There is something similar in spirit about the view I'm developing here and that of González de Prado (2020), in that we both say that the higher-order evidence can prevent the first-order evidence having its justifying force *for you*. However, González de Prado theorizes this by saying that the evidence of irrationality makes it the case that you no longer *possess* the first-order evidence in question. This seems to me an unnatural thing to say. What happens is not that you cease to be in possession of the relevant facts that constitute the first-order evidence (viz., that you got particular scores in your teaching evaluations, or whatever), but that, for you, they cease to *support* the belief that you're an above-average teacher.

Let me try to give a slightly deeper explanation of why this would be so. Once we appreciate the holistic nature of evidential support, we'll see that often, some piece of evidence is an objective indicator that *p*, but isn't (strong) evidence *for you* that *p* because you don't have the background information that puts you in a position to know that *p* is an objective indicator that *p*. That's roughly what happens in the case where you have the raw data about the minimum wage but don't have the competence to interpret it: the problem is precisely that you don't know enough to see what the data indicates about the minimum wage's effects. But (strong) higher-order evidence of irrationality—even misleading higher-order evidence of irrationality—very plausibly *does* rob you of your knowledge about what is an objective indicator of what.[35] For example, when you get reason to think that you've irrationally overestimated your teaching abilities, you *ipso facto* get reason to think that your first-order evidence doesn't in fact objectively indicate that you're an above-average teacher. But, at least, if this reason is strong,[36] it means you're no longer in a position to know that your first-order evidence objectively indicates that you're an above-average teacher.[37] And while that doesn't mean that the first-order evidence stops objectively indicating that you're an above-average teacher, it does mean that the first-order evidence no longer (strongly) supports this conclusion for you, precisely because you're no longer in a position to be aware of this objective indication.

I've said that the evidence that you irrationally overestimated your teaching abilities obviates *or diminishes* the justifying force of the first-order evidence for you. The second disjunct is important. Quite generally, defeat can be partial: it doesn't always *completely extinguish* the support that some evidence supports for a conclusion, but sometimes merely diminishes it.[38] The latter diagnosis is particularly plausible when you're merely getting pro tanto evidence that your belief is irrational or the product of bias, rather than "learning" (for certain) that it is. (The latter, I would add, happens rather infrequently, even though the literature often proceeds as if it were the primary case.) By analogy, if rather than learning that a wall *is* being illuminated by a red light, you learn that it *might well be* being illuminated by a red light (perhaps you learn that a lot of walls around here are illuminated by red lights, but not all are), this diminishes the justifying force of the appearance of redness but doesn't totally extinguish it: provided that it's not certain that the wall is being illuminated by a red light, the appearances of redness is still *some* pro tanto evidence that the wall is actually red.

So too, it may be that the evidence that you're an above-average teacher still provides *some* justificatory support for this proposition even in the presence of the evidence you have irrationally

---

[35] Notice that what is an objective indicator of what is not (purely) *a priori*. It often depends on other *empirical* facts. Whether teaching evaluations are a good indication of teaching ability, for example, is not *a priori*. What is *a priori* is at most what your *total* evidence supports. Thus, Titelbaum's (2015) contention that evidential support relations are *a priori* (n. 1) doesn't help the proponent of Anti-Significance here.

[36] If it's not so strong, it merely means that your epistemic position with respect to the proposition that your first-order evidence objectively indicates that you're an above-average teacher is merely weakened somewhat. See the discussion of partial defeat below.

[37] I'm not begging the question here. As I said right at the outset, it's uncontroversial that higher-order evidence is epistemically significant at the level of higher-order propositions like *my evidence supports p*, *my evidence objectively indicates p*, *my belief in p is irrational*, etc. Evidence that your belief in *p* is irrational should, of course, reduce your confidence that your belief in *p* is rational! The question at issue is whether it should reduce your confidence in the first-order proposition *p*. I'm giving an argument for why, given its higher-order significance *and* the holistic nature of evidential support, it has first-order significance.

[38] See Pryor (2013: esp. 93-94).

overestimated your teaching abilities; the view is just that this support is diminished, or partially defeated. Partial defeat still calls for a reduction in confidence in the proposition you believe at the fine-grained, credal level. And depending on the details of the case, it *can* be enough to render your outright belief (on-off) unjustified. That will depend on how much justification you had to start with and how close-to-total the partial defeat is. If it leaves you with some, but insufficient, evidential support for your belief, this belief will not be (on-off) justified.

So far I've been focused on explaining how evidence of irrationality give rise to undercutting defeat in the case in which this evidence is misleading (i.e., the case in which your belief is *not*, in fact, irrational, at least prior to your getting the higher-order evidence in question). Let's now turn to the case in which the evidence is indicative of a truth (i.e., the case in which your belief *is*, in fact, irrational even prior to your getting the higher-order evidence). Recall that the charge (on behalf of the opponent of assimilating evidence of irrationality to undercutting defeat) is that in *this* case, since there is no justification to defeat, the undercutting defeat model cannot apply.

But our discussion of partial defeat brings out how this contention turns on an equivocation on 'justification'. As I've already been obliquely indicating, the word 'justification' itself admits of both a graded interpretation (on which a belief can be more or less justified) and an "on-off" interpretation (on which a belief is either justified or unjustified).[39] In the case in which your belief is in fact irrational, there is no *on-off* justification to be defeated (assuming an equivalence between [substantive] irrationality and unjustifiedness). But that's compatible with there being some pro tanto *graded* justification, or evidential support, to be defeated. And this can still be defeated (either totally or partially) by the evidence of irrationality that you get.

To make this more concrete, let's once again return to our central case. In the version of the case currently at issue, it's stipulated that you *have* irrationally overestimated your teaching ability. Let's suppose, just to fix some details (and with psychological plausibility) that this isn't because you've hallucinated good teaching evaluations (or whatever), but rather that you've *overweighted* the evidential weight of this evidence and *underweighted* the evidential weight of other countervailing evidence.[40] Still, it's not like there's *no* evidence for the proposition that you're an above-average teacher in this case: the good teaching evaluations are still *some* pro tanto evidence for this proposition (and hence, a source of graded justification for believing it). They just don't suffice to *on-off* justify belief in this proposition (at least, given the countervailing evidence you have). Provided they are a source of pro tanto graded justification or evidential support, this graded justification is a candidate for defeat by the evidence of irrationality. In short, your belief is by stipulation already unjustified when you irrationally overestimated your teaching abilities; but it is *even more* unjustified when you get evidence that makes this irrationality evident, or indicates it. Thus, I think, we can make sense of evidence of irrationality as giving rise to undercutting defeat both in the case where this evidence is misleading and in the case where it's indicative of a truth.

---

[39] Plausibly, the latter is a function of the former. If you don't like the idea that 'justification' admits of a graded interpretation, you can just substitute in the notion of evidential support, which clearly admits us such a graded interpretation.

[40] Such overweighting and underweighting is one of the primary mechanisms by which "motivated reasoning" occurs. Cf Kunda (1990).

5. Drugs and logic

With all this said, there's a special group of cases of evidence of irrationality that might seem especially challenging to assimilate to undercutting defeat. These are cases where one's evidence *deductively entails* some conclusion.[41] Such cases are familiar from the literature; for example, the following is a standard case:[42]

> **Logic-Impeding Drug.** You are a generally competent and reliable logician. You know two propositions, P and Q, and you are trying to figure out whether these propositions jointly entail R (such that R must also be true). You work on the problem, and after a little while you have what seems to you to be an impeccable proof of R from the premises P and Q. However, you are then informed that before being given the logic puzzle, a drug was slipped into your coffee (call this fact D). This drug makes people who are otherwise competent and reliable logicians make elementary logical mistakes about 80% of the time. Moreover, the effects of the drug are subjectively unnoticeable, causing the subjects who have made these mistakes to have the experience as of having reasoned impeccably.

We're then asked to suppose, for the sake of argument, that your proof of R, as a matter of fact, *was* impeccable; you are among the 20% of those who make no logical mistake even in the presence of the drug. Still, the thought is, you don't know this – after all, it's stipulated that if you *had* been affected by the drug, its effects would have been unnoticeable to you. Consequently, D is strong (albeit, in this case, misleading) evidence that your proof of R was mistaken. But can D undercut your evidence for R?

Some say: no! Your evidence for R consists in premises – P and Q – that *deductively entail* R. And deductive entailment (from known premises) is often thought to be the strongest kind of evidential support there is; indeed, it might even be thought to be *indefeasible*. Now, D doesn't undercut your justification for P and Q themselves (it affects only your logical reasoning ability, and we can suppose that you came to know P and Q by some other means, or before the drug was administered). But nor, obviously, does it make it the case that P and Q *stop deductively entailing* R. Thus, since deductive entailment constitutes an (extremely strong) kind of evidential support, it also doesn't make it the case that P and Q *stop supporting* R. Thus, the thought is, D can't be an undercutting defeater of your justification for (believing) R.[43]

Two points here. First: I don't this the choice of cases involving deductive entailments (as a focus of the literature on higher-order) is innocent. If deductive entailments really do provide indefeasible support, this makes them a very special case. And they are comparatively rare ones: we don't typically get evidence that we're under the influence of drugs making us incapable of performing logical inferences, but we do often get evidence that we're irrational or biased in assessing (defeasible) empirical evidence. Thus, even if evidence of irrationality doesn't constitute an undercutting defeater

[41] I'll focus on a case involving logic, but various other mathematical cases could also be used.
[42] Cf. Christensen (2010: 187).
[43] Cf. Christensen (2010: 194-5).

in the latter cases, this falls far short of showing that it doesn't constitute an undercutting defeater in cases of evidence of irrationality more generally. It would be a big mistake to take this verdict about cases like Logic-Impeding Drug that to be a guide to what we should say about other cases of evidence of irrationality, let alone higher-order evidence more generally.

Second, though, I'm suspicious of the claim that entailing evidence is indefeasible. Suppose again that you know that P and that Q. Suppose also that you have a complex proof that P and Q entail R, and so you believe R on the basis of P and Q (and, perhaps, the proof). Now suppose a genius logician tells you (falsely) that the proof is faulty. I think it's plausible that the logician's testimony gives you at least *some* reason to reduce your confidence that P and Q entail R, and that in turn this at least somewhat defeats your justification for believing R on the basis of P and Q. If the mathematician's testimony can bring about defeat of your justification for believing R (even though you know things that *entail* R), then I don't see why discovering that you've taken the logic-impeding drug couldn't make the same difference: both defeat your justification for R (if they do) by (as it turns out, misleadingly) calling into question the soundness of your proof. Of course, neither the mathematician's testimony nor discovery of the logic-impeding drug make it the case that P and Q don't in fact entail R. But (just as in the case of your belief about your teaching ability) what this shows is only that P and Q don't cease to be objective indicators of R's truth; rather, it's that they cease to support R *for you* (or, at least, cease to support it as strongly) against the background of everything else you're aware of. Against a background of reasons to doubt that P and Q do not in fact entail R, P and Q no longer support R *for you* (to the same degree). As in the teaching ability case, we can explain that by appeal to the claim that you're no longer in a position to know that P and Q are objective indicators of R.[44]

6. Conclusion

Let's recap. In response to the common assumption that the significance of higher-order evidence (and in particular, evidence of irrationality) can't be assimilated to undercutting defeat, I've argued two things. First, I've argued that many of the cases of higher-order evidence that we might have *thought* involve evidence of irrationality *don't* in fact involve evidence of irrationality per se but rather of some other kind of cognitive imperfection, and that there's no challenge in assimilating these cases to undercutting defeat. Second, I've argued that even in cases of evidence of irrationality, there's a plausible argument to be made that they can be assimilated to undercutting defeat. The idea that evidence of irrationality cannot be assimilated to undercutting defeat turns out, on reflection, to turn on some combination of a variety of mistakes: an insufficiently holistic notion of evidential support; a conflation between something's being an objective indicator of a proposition's truth and its support

[44] A difference with the teaching ability case, though, is that whereas questions about what is an objective indicator of teaching ability are not *a priori* (n. 35), matters of entailment *are* a priori. Thus, someone who thinks that we have indefeasible justification for *a priori* truths might hold that knowledge that P and Q entail R, and hence knowledge that p and Q are objective indicators of R, is itself indefeasible. If this is so, then this can be used to defend the claim that your justification for believing R on the basis of P and Q might similarly be indefeasible. I concede that the availability of this strategy makes the logic cases at least somewhat better for the opponent of the significance of higher-order evidence (or of assimilating its significance to that of undercutting defeat) than the teaching ability case. Still, I myself find it very dubious that *a priori* knowledge is indefeasible.

that proposition *for a particular* agent; inattention to the possibility of partial defeat; and inattention to the property of graded justification. At most, there's one special category of cases of evidence of irrationality—that where one's first-order evidence *entails* one's belief—that resists assimilation to undercutting defeat. But these cases are outliers, and even with them there is a case to be made that they can be explained by appeal to undercutting defeat after all.

The upshot is twofold. First, opponents of the epistemic significance of higher-order evidence are in a worse position that they might have thought, unless they want to deny the phenomenon of undercutting defeat more generally, in classic cases like Red Light. Second, friends of the epistemic significance of higher-order evidence don't need fancy, revisionary machinery like bracketing, dispossessing defeat, or non-evidential reasons to suspect. They can simply use the well-accepted framework of undercutting defeat to explain the significance of higher-order evidence (evidence of evidence aside; this is explained differently). Perhaps, then, the significance of evidence that I'm biased or irrational should, after all, be no more controversial than the significance of evidence that the wall is being illuminated by a red light. The framework of undercutting is more widely applicable than we first thought.

References

Avnur, Yuval & Scott-Kakures, Dion (2015). How Irrelevant Influences Bias Belief. *Philosophical Perspectives*, 29, 7-39.

Bertrand, Marianne & Mullainathan, Sendhil (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94(4), 991-1013.

Chen, Yan & Worsnip, Alex (2025). Disagreement and Higher-Order Evidence. In M. Baghramian, J. A. Carter & R. Cosker-Rowland (eds.), *The Routledge Handbook of the Philosophy of Disagreement*, 223-236. Routledge.

Christensen, David (2010). Higher Order Evidence. *Philosophy and Phenomenological Research* 81(1), 185-215.

Eder, Anna-Maria & Brössel, Peter. (2019). Evidence of Evidence as Higher Order Evidence. In M. Skipper & A. Steglich-Petersen (eds.), *Higher-Order Evidence: New Essays*, 62-83. Oxford University Press.

Feldman, Richard (2005). Respecting the Evidence. *Philosophical Perspectives* 19, 95-119.

Fitelson, Branden (2012). Evidence of Evidence is Not (Necessarily) Evidence. *Analysis* 72(1), 85-88.

Friedman, Jane (2013). Suspended Judgment. *Philosophical Studies* 162(2), 165-181.

Gilovich, Thomas, Vallone, Robert & Tversky, Amos (1985). The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive Psychology* 17(3), 295-314.

Goldman, Alvin (1986). *Epistemology and Cognition*. Harvard University Press.

González de Prado, Javier (2020). Dispossessing Defeat. *Philosophy & Phenomenological Research* 101(2), 323-340.

Green, Brett & Zwiebel, Jeffrey (2018). The Hot-Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments? Evidence from Major League Baseball. *Management Science* 64(11), 5315-5348.

Kelly, Thomas (2014). Evidence. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/evidence/

Kunda, Ziva (1990). The Case for Motivated Reasoning. *Psychological Bulletin* 108(3), 480-498.

Lasonen-Aarnio, Maria (2010). Unreasonable Knowledge. *Philosophical Perspectives* 24, 1-21.

-------- (2014). Higher-Order Evidence and the Limits of Defeat. *Philosophy and Phenomenological Research* 88(2), 314-345.

Lord, Errol & Sylvan, Kurt (2021). Suspension, Higher-Order Evidence, and Defeat. In J. Brown & M. Simion (eds.), *Reasons, Justification, and Defeat*, 116-145. Oxford University Press.

Matthews, Genae (ms.). Social Location as Higher-Order Evidence.

Meehl, Paul E. (1956). Wanted – A Good Cook-Book. *American Psychologist* 11(6), 263-272.

Pollock, John (1970). The Structure of Epistemic Justification. In N. Rescher (ed.) *Studies in the Theory of Knowledge* (*American Philosophical Quarterly* Monograph Series No. 4), 62-78. Blackwell.

Pryor, James (2013). Problems for Credulism. In C. Tucker (ed.), *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservativism*, 89-132. Oxford University Press.

Pust, Joel (2000). *Intuitions as Evidence*. Routledge.

Tal, Eyal (2020). Is Higher-Order Evidence Evidence? *Philosophical Studies* 178(10), 3157-3175.

Titelbaum, Michael G. (2015. Rationality's Fixed Point (or: In Defense of Right Reason). *Oxford Studies in Epistemology* 5, 253-292.

Vavova, Katia (2018). Irrelevant Influences. *Philosophy and Phenomenological Research* 96(1), 134-152.

Weatherson, Brian (2019). *Normative Externalism*. Oxford University Press.

White, Roger (2005). Epistemic Permissiveness. *Philosophical Perspectives* 19, 445–459.

-------- (2010). You Just Believe that Because… *Philosophical Perspectives* 24, 573-615.

Worsnip, Alex (2018). The Conflict of Evidence and Coherence. *Philosophy and Phenomenological Research* 96(1), 3-44.

-------- (2021). *Fitting Things Together: Coherence and the Demands of Structural Rationality*. Oxford University Press.

-------- (2023). Suspiciously Convenient Beliefs and the Pathologies of (Epistemological) Ideal Theory, *Midwest Studies in Philosophy* 47, 237-268.