Can Bayesians Give an Account of Higher Order Defeat?

Abstract: Sometimes we get evidence which suggests that our beliefs aren't rational. This might come from learning that people with the same evidence reach different conclusions, that we're under the influence of mind-distorting drugs, subject to implicit biases, or that our beliefs have been impacted in potentially problematic ways by social or evolutionary forces. Defeatists think that this type of evidence should sometimes induce a radical shift in our credences. In this paper I'll argue that thinking of such revisions as Bayesian, and motivated by principles constraining the relationship between our credences in "higher order" and "lower propositions," while possible, raises a puzzle, and brings with it a set of substantial and controversial commitments.

1. Introduction

To what extent can higher order evidence – evidence that bears on the rationality of our beliefs – defeat our first order opinions? This question, understandably, has been closely bound up with discussion of what are called "bridge" or "enkratic" principles. Such principles impose constraints on how a rational agent's higher order credences (credences in propositions about rationality) and first order credences (credences in propositions with purely descriptive contents) interact with one another. In particular, these principles are meant to rule out the rationality of certain forms of akrasia: for example, believing P, while also believing that believing P is irrational. I will argue that a Bayesian account of higher order defeat with enkratic commitments has some rather surprising consequences, yielding an epistemological picture that is somewhat stranger than one might have expected. I won't be arguing *against* such an account. Instead, I'll just draw out the consequences of the view with the hope of giving the reader the resources they need to evaluate it using their own philosophical tastes.

Throughout, when I talk about a "Bayesian" account of defeat or a Bayesian defeatist, I'm referring to a combination of three views:

(1) *Defeatism* – The view that higher order evidence (evidence concerning the rationality of your beliefs) can have a powerful defeating effect on our first order views, as exemplified in the specific cases I'll present below.
(2) *Classical Bayesianism* – probabilism and classical conditionalization
(3) *Bridging* – A principle that imposes constraints on how our higher order and first order credences interact with one another.

I'll begin with some discussion of each of the three components, and then proceed to show what follows when they're combined.

2. The Defeatism Component

Let's begin by considering a version of a now classic toy case from the literature on higher order evidence:

> HYPOXIA[1]: You're a cargo pilot who was hired to fly from London to Philadelphia to deliver some goods. While in the air you get a message from your employer. They inform you that, due to some recent developments, it would be optimal for the goods to be delivered to Los Angeles rather than Philadelphia. Unfortunately, they can't afford a refueling stop, but they're wondering whether you have enough fuel to safely make it to Los Angeles. At this point your credence that you have enough fuel to make it to Los Angeles (call this proposition L) is 0.5.
>
> You know that you can easily gather some evidence that will support either a high credence in L or (an equally) high credence in ~L (depending of course on what type of evidence you receive), and that upon evaluating the evidence, you will judge either that L is likely to be true given this evidence or that ~L is. At 12pm you gather some evidence by looking at your gauges dials and maps (call the evidence you get E). You judge that L is likely given the evidence and your credence in L increases significantly.
>
> You get a call from ground control at 12:05pm. They alert that for the last five minutes you've been flying at an altitude which puts you at risk for hypoxia: a condition that impairs one's ability to reason properly, even though everything seems normal. Pilots reasoning at your altitude, you are informed, reach the conclusions that are supported by their evidence at the same rate as chance. Call the information you get from ground control "H."

[1] I believe hypoxia was first discussed in the higher order evidence literature in Elga (ms.). It has been discussed by many others since.

Defeatist verdict: Your credence in L after you learn you H, should go back to 0.5.

Why 0.5 in particular?[2] Why not, say, 0.8 ? Or 0.9? To get a feel for what's motivating the defeatist verdict here consider the following variant of the case:

> PITTSBURGH-HYPOXIA: You're a cargo pilot who was hired to fly from London to Philadelphia to deliver some goods. While in the air you get a message from your employer. They inform you that, due to some recent developments, it would be optimal for the goods to be delivered to Pittsburgh rather than Philadelphia. Unfortunately, they can't afford a refueling stop, but they're wondering whether you have enough fuel to safely make it to Pittsburgh. At this point your credence that you have enough fuel to make it to Pittsburgh (call this proposition P) is 0.95. (Pittsburgh is quite close to Philadelphia so, from the get-go, you think it's very unlikely that you'd be in a situation in which you have enough fuel to make it to Philadelphia but not enough to make it to Pittsburgh).
>
> At 12pm you gather some evidence E' by looking at your dials gauges and maps and upon considering it you end up with a credence of 0.99 that you do indeed have enough fuel to make to Pittsburgh. You get a message from ground control at 12:05pm alerting you that for the past five minutes you've been flying at an altitude which puts you at risk for hypoxia. Pilots reasoning about such matters at this altitude reach the conclusions that are supported by their evidence at the same rate as chance.

What is the defeatist verdict for this case? The defeatist of course won't think that your credence that you have enough fuel to make it to Pittsburgh should shoot all the way down to 0.5. That would be a highly implausible verdict. Rather, they'll think your credence that you have enough fuel to make it to Pittsburgh in such a case should be (at least) 0.95. Why? Because learning that a certain process does no better than chance shouldn't take you to some magic number like 0.5, but to whatever your *prior* happened to be. To put the point another way: the defeatist intuition is based on the thought that learning that you're reasoning under conditions in which you're likely

[2] I don't believe much rests on the defeatist verdict being *exactly* 0.5. I'm quite confident that the consequences I describe contain equalities that would be approached as the defeatist verdict approaches 0.5, though I don't provide a proof for this.

to be impaired means that you should distrust the reasoning done *under the conditions in which you're likely to be impaired*.[3] In the original HYPOXIA case your prior credence was 0.5 and so distrusting the reasoning you did while at risk for hypoxia results in a dramatic reduction of confidence. In PITTSBURGH-HYPOXIA your prior is 0.95 and so distrusting the reasoning you did while at risk for hypoxia results in a moderate reduction of confidence.

It's worth noting that Isaacs (2021) raises an objection to a view he calls "calibrationism" – a view intended to deliver defeatist verdicts. The worry raised by Isaacs assumes that the calibrationist is committed to the thought that in a case like PITTSBURGH-HYPOXIA your credence should go down to 0.5. And, as Isaacs rightly points out, to think that in such a case your credence should go down to 0.5 is to commit the base-rate-fallacy: to ignore the effect your priors should have on your final credence. (For discussion of importance of priors in cases of defeat see also White (2009) and Pittard (2019)). Whether the particular authors Isaacs targets are indeed committing such a fallacy need not concern us here. The point is rather that the best version of a Bayesian account of defeat should *not* commit the base-rate-fallacy. Here then is a premise I'll be appealing to:

> REVERSION TO PRIORS: Defeatists should think that learning that a cognitive process you engaged in performs exactly as well as chance, should result in a revision to your credences that is no more dramatic than a reversion to your priors.

REVERSION TO PRIORS contains an implicit assumption about how defeaters work. To see why, consider a version of Isaacs' example:

> SHOCHET: Moishe is a shochet (a person hired to ritually slaughter animals in accord with Jewish law) and the first thing Moishe does when he is brought an animal is determine whether the animal is kosher. It is extremely unusual to bring a shochet a non-kosher animal, and so when Moishe learns that he's about to be presented with an animal, he is already .95 confident that the animal will be kosher. When the animal arrives Moishe takes

[3] This way of thinking about things sits well with the sorts of things that people who defend defeatist verdicts say: they claim that if you judged that P, and then get a defeater, your credence should equal the probability that you judged correctly *setting aside/bracketing/independently of the reasoning in question*. (See e.g. Christensen (2010), Horowitz and Sliwa (2015), Schoenfield (2015a)). In a case like PITTSBURGH-HYPOXIA, the reasoning in question is the reasoning done under hypoxia-inducing conditions. Setting aside the reasoning done under these conditions, you'll be 0.95 confident you have enough fuel to make it to Pittsburgh.

> a close look and confidently judges that the animal is indeed kosher. At this point Moishe's credence that the animal is kosher goes up to 0.99. Moishe is then told that he is quite drunk (so drunk, in fact, that he doesn't know what day it is or even that he's been drinking). He is also told that, when he is inebriated to this degree, he will often mistake a cow for a pig or a pig for a cow, and that his judgments about whether a given animal is kosher are no better than chance.

Isaacs points out (concerning a similar version of this case) that "Quite apart from Moishe's judgment, it's substantially more likely than not that the animal brought to him is kosher." (p.3). And so, Isaacs claims, it would be extremely implausible to suppose that Moishe's credence should shoot all the way down to 0.5 upon learning that his judgments are unreliable.

This seems right. But let's now add a wrinkle to the story. Suppose Moishe knows that he only gets drunk on Purim – a Jewish festival in which getting drunk is customary. Drunkenness, however, isn't the only custom on Purim. It's also customary to play practical jokes, and what better practical joke to play on a shochet than bringing him a pig? If that's right, it may well turn out that Moishe *should*, upon learning that he's drunk, reduce his confidence substantially (even all the way to 0.5!) in the proposition that the animal is kosher. But this reduction wouldn't come primarily from anything to do with Moishe's reliability, but from the fact that his drunkenness is *itself* evidence that he'll be presented with a non-kosher animal (since it is evidence for its being Purim). In such a case, Moishe shouldn't necessarily revert to his priors.

REVERSION TO PRIORS then is only plausible when nothing funny like that is going on – when the *only* way in which the defeater defeats, is by making it the case that the proposition that you formed a judgment that P provides no confirmation for P. This assumption can be formalized through a stipulation on priors (in Moishe's case by "priors" I mean his probability function prior to seeing the animal *and* prior learning that he's drunk). If Moishe's priors are such that Pr(I am brought a kosher animal|I am drunk) = Pr(I am brought a kosher animal|I'm not drunk), it will follow from Bayes' theorem, and the fact that Moishe is as reliable as chance when drunk, that Pr(I am brought a kosher animal|I am drunk and judge that the animal is kosher) = Pr(animal is kosher). (See Isaacs (2021) for more detail). If however, Moishe's priors are such that Pr(I am brought a kosher animal|I am drunk) < Pr(I am brought a kosher animal|I'm not drunk), as is the case when drunkenness is *itself* evidence for practical jokes and non-kosher-animal-bringings, REVERSION TO PRIORS won't be derivable from Bayesianism, nor would it be independently plausible.

One last note about defeat before moving on: Many of the real-life exciting cases of higher order evidence are more complicated and controversial than HYPOXIA and raise issues that don't arise in a simple toy case like this one. There are also many cases of defeat which don't involve some of the stipulations made in HYPOXIA: for example, cases where your uncertainty about what's rational involves more than two hypotheses (while in HYPOXIA you're certain that E supports one of L or ~L and whichever it supports it supports to the same degree), or cases in which you leave open the possibility that you'll look at the evidence, be baffled, and not arrive at any judgment at all about what the evidence supports (while in HYPOXIA you're certain that upon looking at the evidence you'll judge either that it supports L or that it supports ~L). Nonetheless, my thought is that if a Bayesian account of defeat is going to successful, it should apply to a simple and clean case like HYPOXIA. And if it doesn't, because of certain simplifications or idealizations about the case, that will point the way to questions about what kind of defeat cases a Bayesian account *could* potentially accommodate.

3. The Bayesian Component

The Bayesian picture I'm working with is one on which a rational agent's degrees beliefs (or credences) can be represented by a probability function, and the agent's credences get revised through classical conditionalization.

I am by no means the first to question whether Bayesianism and defeatism can live in harmony. An initial worry you might have about their compatibility is that, arguably, logical and mathematical beliefs can be defeated (Christensen (2010)), and uncertainty about logic and mathematics have never sat particularly comfortably with Bayesianism. I don't however take this to be a good motivation for giving up on a Bayesian account of defeat from the get-go for two reasons. First, logical omniscience is a general challenge to Bayesianism, and there is no reason the defeatists can't avail themselves to any of the variety of approaches to logical uncertainty Bayesians have defended.[4] Second, there are plenty of cases of defeat in the literature that don't involve any logical or mathematical reasoning.[5] And if Bayesianism can't account for defeat in non-logical cases, then the issue raised by defeat is distinct from the problem of logical omniscience.

---

[4] For a survey see Halpern and Pucella (2011). For a smattering of approaches see Hacking (1967), Garber (1983), Stalnaker (1991), Gaifman (2004), Seidenfeld et al. (2012), Dogramaci (2018), Skipper and Bjerring (forthcoming), Pettigrew (forthcoming) and Elga and Rayo (ms.)

[5] They include sleepy detectives reasoning about who committed a crime (Horowitz (2014)), people with implicit bias considering job applications (Christensen (2019)), wishful thinking about the behavior of a college president (Avnur

Setting such worries aside, there are a number of other problems that have been raised for a Bayesian account of defeat. I believe, however, that the arguments that have been offered so far that point to a tension between Bayesianism and higher order defeat are problematic. I survey these arguments in Appendix 1. Given that, as I argue in the appendix, existing arguments don't doom a Bayesian account, I believe a fresh investigation into the question is in order.

4. The Bridging Component

In this section I will discuss the final of the three positions which constitute what I'm calling "Bayesian Defeatism": a bridge principle that constrains the relationship between first order and higher order credences.

In bundling such a bridge principle together with Bayesian and defeatism, I am assuming that higher-order defeat is, indeed, higher order. What I mean by this is that the reduction of confidence in a first-order proposition (like L) that results from an encounter with a higher-order defeator (like H) is explained by a reduction of confidence in a proposition about the *rationality* of believing L. This is commonly assumed (either implicitly or explicitly) in the literature on higher order defeat. Christensen (2010) for example says that higher order evidence "rationalizes a change of belief precisely because it indicates that my former beliefs were rationally sub-par" (p.185).

I'll call the proposition that believing L is rational in response to evidence E "RatEL." Because we are assuming classical conditionalization, if it's rational to be confident in L when your evidence is E, it follows that the rational prior probability function, which I'll "Pr" is such that $Pr(L|E) = high$. Thus, RatEL will the proposition that $Pr(L|E) = high$.

It's worth noting here that RatEL is not a proposition about entailment relations. While it may be that in some spellings out of HYPOXIA some of the reasoning you do involves working out entailment relations, the reasoning in HYPOXIA needn't all be mathematical (you might be considering what flight path you'd most likely be routed on if you were to fly to L.A during this time of day for example). More importantly, as I mentioned above, many cases of defeat involve no entailment relations at all. So to allow the account to generalize to a broad range of defeat cases it is important to think of RatEL as a proposition about what makes L likely, not what makes L certain.

---

and Scott-Kakures (2015)), and subliminal messaging influencing our views about the efficacy of pet therapy (Vavova (2018)) to name just a few examples.

Since we're assuming that you know throughout that E supports one of L or ~L and that whichever it supports, it supports to the same degree, it will be useful to have a variable representing the degree to which you think either L or ~L is supported by E (as opposed to talking about the probability equaling "high").So let *r* be a number greater than 0.5 such that in HYPOXIA you're certain that either: Pr(L|E) = *r* or Pr(~L|E) = *r*.[6] Since you're certain that one or the other is true, ~RatEL will be the proposition that it is rational to be confident that ~L in response to E. Thus, we have that for $r > 0.5$:

RatEL: Pr(L|E) = *r*
~RatEL: Pr(~L|E) = *r* or equivalently: Pr(L|E) = 1-*r*

Shortly I'll state the bridge principle that will yield the result that reduction of confidence in RatEL brings about a reduction of confidence in L.[7] But first I'd like to address a potential big-picture concern that might arise at this point. You might be wondering though whether we really *should* think of defeatism as closely connected to some sort bridge principle. Isn't the interesting question simply whether *Bayesianism* is compatible with defeatist verdicts? Why care about these bridge principles? If things get complicated when bridge principles and funny higher-order propositions like RatEL get involved, you might say, so much the worse for such principles and propositions!

There are really two separate questions here. The first is, why we should be interested in a Bayesian view of defeat which takes any stand at all on the connection between higher order and first order credences. The second is: why assume that such connections can be neatly described by some systematic principle? I'll address each in turn.

The reason we should be interested in a Bayesian view of defeat which takes a stand on the connection between a rational agent's higher and first order credences is that such connections play an important role in motivating defeatist verdicts. If evidence suggesting that your belief

[6] You might worry that the fact that you're certain that E supports one of L or ~L and that whichever it supports, it supports to the same degree doesn't entail that, for some *r*, you are certain that E supports L to degree *r* or E supports ~L to degree *r*. Although I suspect the weaker claim is all that's needed for the derivations that follow, I will assume the stronger one throughout since it will make the derivations much simpler. It won't matter for the overall dialectic because presumably defeatist verdicts are meant to also apply to cases in which you do happen to know, for some *r*, that E supports L to degree *r* or E supports ~L to degree *r*. If I can raise trouble for a case with this feature that will suffice to generate the puzzle I'm interested in.

[7] For defenses and discussion of a variety of different bridge principles see e.g. Feldman (2005), Christensen (2012), Elga (2013), Greco (2014), Horowitz (2014), Ramussen et al. (2018), Dorst (2019) and Dorst et al. (ms.).

about the fuel is *irrational*, is to induce a change in your opinion *about the fuel*, you must be taking facts about the rationality of believing L as relevant to the question of whether L. And indeed, the standard line amongst those who *reject* defeatists verdicts is to say: "sure, H might give you evidence that your belief in L is irrational, but so long as your belief in L in fact is well-supported by the evidence, then this (misleading) evidence against the rationality of your belief, isn't a reason to doubt L itself. [8]

Here's another way to get a sense of why bridges between first order and higher order credences play an important role in the context of defeat: imagine a variant of the case in which ground control's message went as follows: "We need to alert you that you're at significant risk for hypoxia. Pilots at your altitude reason about such matters in a manner that is no better than chance. However, we have just examined all of the evidence E that you collected and we assure you that it does indeed support the claim that you have enough fuel to make it to Los Angeles." In this case, intuitively, there's no defeat (or the defeat is defeated). Why? Presumably because the defeat of L in the original case took hold by challenging your belief that L is well-supported by E. Once the case is set up so you have no reason to doubt that L is supported by E, H is no longer a threat to your belief that L.

Let's move on to the second part of the question though: even if we accept that defeatism involves *some* kind of enkratic commitment – a commitment to there being *some* connection between the higher and lower orders – why think that such a commitment can be encoded by some neat systematic principle? In fact, although I will be putting forward a principle and appealing to it at two junctures, I don't believe defeatists need to accept that any such principle holds universally. What is necessary for the arguments that follow isn't the principle itself. but the two particular applications of the principle that I'll be appealing to. Later in the paper I will provide some independent motivation for these two applications, but I appeal to the bridge principle throughout because I believe the principle gives an illuminating and unifying story about what's going on in ordinary defeat cases.

With this in mind, let's move on to the principle. The principle I'll propose is version of the weakest such principle endorsed in the current literature in the context of defeat that I'm aware of

[8] For instance, Wedgwood (2011), Coates (2012), Lasonen Aarnio (2014), and Weatherson (ms.).

– Elga's (2013) "New Rational Reflection" (NRR) (which is equivalent to Dorst's (forthcoming) "Hifi"). In fact, the principle I'll propose is, in a sense I'll explain below, even weaker than NRR. [9]

> BRIDGE: Let Pr be the rational ur-prior[10] and let Pr* be Pr(· |Pr is rational). Then, if your total evidence is $\mathcal{E}$, the rational credence function to adopt given that evidence, $Pr_{\mathcal{E}}$, is such that
>
> $$Pr_{\mathcal{E}} = Exp_{Pr_{\mathcal{E}}}(Pr^*(\cdot \mid \mathcal{E}))$$

This principle says that a rational credence function with total evidence $\mathcal{E}$ will equal the expectation of the value of a certain other credence function. Which credence function is that? Well, it is a credence function that is just like the rational one given $\mathcal{E}$, except that in addition to knowing $\mathcal{E}$, it knows what the rational ur-prior is.[11] I say "ur-prior" so as to the state the principle as generally as possible. But for our applications to the HYPOXIA case we can just as well think of "the ur-prior" as the rational credence function to have prior to getting E.

The difference between BRIDGE on the one hand, and NRR on the other is that, while BRIDGE tells you that your credences (when your evidence is $\mathcal{E}$) should equal your expectation of the rational credence function that knows both $\mathcal{E}$ and which credence function is *the rational ur-prior*, NRR says that your credences should equal your expectation of the rational credence function that knows both $\mathcal{E}$ and which credence function is *rational given* $\mathcal{E}$.

---

[9] Christensen (forthcoming) is the only defeatist I'm aware of who rejects NRR (a principle that avoids the worries raised by Williamson (2011,2014) and Horowitz (2014) about clocks and dartboards that arise for other bridge principles). Since my principle is a version of NRR, Christensen (forthcoming) would presumably reject mine as well. However, Christensen's reasons for rejecting NRR involve a specific type of counterexample (based on a case from Barnett (forthcoming)), in which one gets testimonial evidence that rationality is not accuracy-conducive. In effect, an expert recommends that you be akratic in these examples. Christensen writes: "The present cases are fully consistent with the usual silliness of frankly akratic subjects, and the irrationality of most akrasia. In fact, they help *explain* it: in most ordinary cases, subjects are rational to expect that rationality and accuracy go hand-in-hand" (11). The cases I'll be focused on fall squarely into the category of cases in which Christensen would regard akrasia as exhibiting "the usual silliness" so I take Christensen (forthcoming) to be amongst the defeatists who will be sympathetic to the kind of enkratic judgments I'll be appealing to in this paper. (And, as I mentioned, I will provide some independent motivation for these judgments later in the paper so that accepting the general principle is not crucial for what follows).

[10] Ur-priors are the opinions of the so called "Bayesian superbaby" – an imaginary being with great cognitive sophistication, but no evidence. In this context, the ur-prior is being used as a way to represent a rational agent's most fundamental opinions. I am simplifying here by assuming that there is a single rational ur-prior. However, if you're a subjective Bayesian, you can just as well think of this function as *the particular subject's* ur-prior (a probability function that will encode the *particular subject's* most fundamental opinions, as opposed to the most fundamental opinions of *any rational subject*).

[11] If you don't like talk of ur-priors, you can just well think of Pr* as $Pr_{\mathcal{E}}$ conditional on the true proposition of the form: "Function *F* is the function from bodies of evidence to probability functions that describes which probability function is rational in respone to a given body of evidence."

The problem with appealing to NRR in a defeat context is that such a principle doesn't do a great job at motivating or explaining defeatist verdicts. Here's why: In HYPOXIA you're meant to become uncertain, upon learning H, about what it is rational to believe *given E*. But so long as you're not completely out of your senses, you may still know that, in response to a defeater like H, you should give up your belief that L. In other words, while you may not be in a position to be confident about what *E* supports, you may well be in a position (because you know that defeatism is true) to be confident about what *E&H* supports (namely, a 0.5 credence in L). If that's right, then once you have H in hand, you may have no uncertainty at all about what's rational given your *total* evidence. So insofar as rational uncertainty is going to play a role in explaining the defeatist verdicts, the uncertainty will concern how likely L is *given E,* not how likely L is *given E&H.* (Recall also Christensen (20210) saying that higher order evidence "indicates that my *former* beliefs were rationally sub-par" (p.185, my emphasis)). Thus, if H is meant to change your expectation of the value of some credence function, and this change in expectation is meant to change your credence in L, it had better be a credence function that knows what's rational given E, and not a credence function that knows (only) what's rational given E&H.

Although BRIDGE is neither *logically* weaker or stronger than NRR, there is a sense in which it is weaker than NRR: Principles like NRR, or the Principal Principle are in the family of "expert deference" principles. They tell you to think of some probability function as an expert and, in the sense made precise by the principle, "defer" to it. The Principal Principle tells you to defer to the objective chance function and NRR tells you to defer to the opinion of a rational agent who knows what is rational given your evidence. But the "expert" that BRIDGE is talking about is an even *greater* expert than the one NRR is talking about. This expert not only knows what's rational given your evidence – they are rationally omniscient – they know *all* there is to know about rationality. For (assuming classical Bayesianism) the rational ur-prior indicates what is rational given *any* body of evidence. So if you thought you should defer to an agent who has all your evidence, is rational, and knows a fact or two about rationality, you should be even more enthusiastic about deferring to an agent who has all your evidence, is rational, and knows *all* the facts about rationality.

I'll now show how BRIDGE, supplemented with two additional assumptions about HYPOXIA, can be used to derive the defeatist verdict. By "derive the defeatist verdict" I mean: show that *if you should become agnostic about RatEL when you learn H*, you should also become agnostic about L when you learn H. Here is the first of the two additional assumptions needed for the derivation:

Assumption 1: Throughout the cases under discussion, for real numbers r and r^* both greater than 0.5 you are certain that (a) $Pr(L|E) = r$ if and only if $Pr^*(L|E) = r^*$ and (b) $Pr(\sim L|E) = r$ if and only $Pr^*(\sim L|E) = r^*$. (As a reminder: $Pr^* = Pr(\cdot\ |\text{Pr is rational})$).

Recall that we stipulated in HYPOXIA that you are certain that E supports one of L or ~L, and whichever it supports, it supports to the same degree. Assumption 1 adds that this holds relative to Pr* as well. Assumption 1 also adds that Pr and Pr* *agree* about which of L or ~L the evidence supports. To deny this would involve thinking the following: if you learn E you should be confident that L, but if you then become certain about what the rational prior is (and so become certain that E *does* indeed support L), you should become confident that ~L. There may be special cases in which learning about the evidential support relations flips their valence, but I'll be assuming that the cases under discussion lack such exotic features.

Assumption 2: H has its defeating effect on L through inducing uncertainty about what's rational. Formally: $Pr(L|\text{Pr is rational} \,\&\, E\&H) = Pr(L|\text{Pr is rational} \,\&\, E)$

Assumption 2 is just a way of making sure that H doesn't have some indirect bearing on L– one that doesn't go via higher order uncertainty. For example, if you thought that the fact that you're at risk for hypoxia is itself evidence that you're using a lot of fuel (perhaps because hypoxia only becomes a risk at a certain altitude and you need to use a lot of fuel to reach that altitude), then you might think H is evidence against L even setting aside any uncertainty about rationality.

BRIDGE in combination with Assumptions 1 and 2 allows us to derive the following:

Consequence 1: If your total evidence is E&H, it's rational to assign a 0.5 credence to RatEL if and only if it's rational to assign a 0.5 credence to L.

(Here, and throughout the paper, the proofs can be skipped for readers who aren't interested in the technical details).

Proof of left to right: Assume that, it's rational to be 0.5 confident that RatEL once you've learned E&H in HYPOXIA. To figure out what constraints BRIDGE imposes on your credence in L, we need

to consider what Pr* conditional on E&H, will think about L. Since Pr* is certain about what is rational given E, it follows from Assumption 2, that H will have no effect on Pr*'s credence in L. That is Pr*(L|E) = Pr*(L|E&H). It follows from this fact in combination with BRIDGE that your credence in L given *E&H* should be the expectation of Pr*'s credence in L *given E*. What will your expectation of Pr*(L|E) be in this case? If you're 0.5 that RatEL, then you're 0.5 confident that Pr(L|E) = r >0.5 and 0.5 that Pr(~L|E) = r. By Assumption 1, Pr(L|E) = r if and only if, for r^*>0.5, Pr*(L|E) = r^*. Additionally: Pr(~L|E) = r if and only if Pr*(~L|E) = r^*. It follows (from the fact that you must assign equal probability to propositions you regard as equivalent) that you're 0.5 confident that Pr*(L|E) = r^* and also 0.5 confident that Pr*(~L|E) = r^*. By probabilism, Pr*(~L|E) = r^*, if and only if Pr*(L|E) = 1-r. Thus, your expectation of Pr*(L|E) = (.5)r^*+(0.5)(1-r^*) = 0.5. Since your credence in L should be your expectation of Pr*(L|E), your credence in L should be 0.5.

Proof of right to left: Start by assuming that you should end up 0.5 confident that L when your total evidence is E&H. That is: where Pr_{EH} is the rational probability function to adopt upon learning E and H, Pr_{EH}(L) = 0.5. Suppose for reductio that Pr_{EH}(RatEL) ≠ 0.5 and without loss of generality that it is greater than 0.5. Thus we're assuming that (for r > 0.5)

Pr_{EH}(RatEL) =Pr_{EH}((Pr(L|E) = r) >0.5

By Assumption 1, and the fact that you must assign equal probability to equivalent propositions it follows that Pr_{EH}(Pr*(L|E)=r^*) >0.5.

Since the propositions in: {Pr*(L|E)=r^*, Pr*(L|E)=1-r^*} form a partition (Assumption 1 and probabilism), if we let Pr_{EH}(Pr*(L|E)=r^*) = x it follows that Pr_{EH}(Pr*(L|E)=1-r^*) = 1-x

This means that the expectation of Pr*(L|E) relative to Pr_{EH} will be xr^*+(1-x)(1-r^*). Since x>0.5, and r^*>0.5, we have that:

Exp_{PrEH}(Pr*(L|E))>0.5.

From BRIDGE we have that

Pr_{EH}(L) = Exp_{PrEH}(Pr*(L|E&H)).

As argued above, it follows from Assumption 2 that H has no defeating on Pr* and so:

Exp_{PrEH}(Pr*(L|E&H)) = Exp_{PrEH}(Pr*(L|E)) .

We established that Exp_{PrEH}(Pr*(L|E))>0.5. And so it follows from the above two equalities that Pr_{EH}(L)>0.5 contrary to our assumption (/the defeatist verdict).

5. Reasoning-for-fun

Now that I've laid out the three components of the view under discussion – Bayesianism, defeatism, and BRIDGE – and shown how they fit together, I want to present a variant of HYPOXIA that will be play a role in the arguments that follow:

> REASONING-FOR-FUN*:* While flying your plane at 12pm you consider, just for the fun of it, how likely L is to be true, on the supposition that E. You are certain that E supports one of L or ~L and that whichever it supports it supports to the same degree. You are also certain that by 12:05pm you will have judged either that L is likely given E, or that ~L is likely given E. You don't have evidence E in hand – you are thinking about the matter purely suppositionally. You judge that L is likely given E. You then learn H.

I will make the following two additional assumptions:

Assumption 3: In REASONING-FOR-FUN, if after learning H, the agent proceeds to learn E, they will end up with the same total evidence as the agent in in the original HYPOXIA case.

In other words, learning E in REASONING-FOR-FUN would amount to "catching up" to the agent in HYPOXIA from an evidential perspective. Or to put the point yet another way: the only relevant evidential difference between the two cases is that in REASONING-FOR-FUN the agent lacks E and in HYPOXIA the agent has E.

Assumption 4: At every point in time, in the cases under discussion, E and RatEL are independent: E doesn't confirm or disconfirm which credence in L is rational given E.

The assumption that a body evidence doesn't confirm what is rational given that evidence plausibly holds in a wide range of cases (if it didn't, conditionalization wouldn't look like a very good update procedure). When might it not? Perhaps a case in which the oracle tells you that they'll give you some evidence $\mathcal{E}$, if and only if a proposition P is likely given $\mathcal{E}$? Perhaps. But in such a case your total evidence of course includes more than $\mathcal{E}$: it also includes the oracle's

testimony. Perhaps some clever case could be devised which violates Assumption 4, but I'll assume that in the cases under discussion, Assumption 4 holds.

I want to pause for a moment to make a methodological remark about these claims I'm calling "assumptions." These assumptions are in fact borderline-stipulations. What I mean by this is that they are aimed at fleshing out the details of HYPOXIA and its variants with more specificity than in the rough sketches presented in the indented case descriptions. The reason these assumptions aren't simply stipulated as part of the case is that, although I think they are plausible ways of filling out such cases, I want to flag the places the Bayesian defeatist may have some room to maneuver. I make explicit those opportunities as we go by calling these aspects of the cases "assumptions." But the methodology of these assumptions works like this: For any such assumption *A*, you can think of me appealing to the following premise: *If defeatism is plausible, then the defeatist verdict should apply to a case in which A holds*. In other words, the defeatist who wants to reject my conclusions by resisting the assumptions will be one who is happy to restrict the scope of their view to cases in which these assumptions don't hold.

Back to the main thread: I will now make use of Assumptions 3 and 4, as well as BRIDGE to argue for:

> ***Consequence* 2**: In REASONING-FOR-FUN, after learning H, your credence in RatEL should be 0.5.

Proof of Consequence 2: Let Pr' be the agent's credence function in REASONING-FOR-FUN after learning H. By Assumption 3, if the agent were to learn E at this point, they'd end up with the same total evidence as the agent in HYPOXIA. Since the order in which you learn evidence won't matter for a classical Bayesian, it will follow that if the agent came to learn E, they should end up with the same credences as the agent in HYPOXIA. Since, by defeatism and Consequence 1, the agent in HYPOXIA will assign a credence of 0.5 to RatEL, it will follow that if the reasoning-for-fun agent were to learn E, they should also assign a credence of 0.5 to RatEL. Since we're assuming the agent would be updating by classical conditionalization it follows that $Pr'(RatEL|E) = 0.5$. By Assumption 4, E isn't evidence for or against RatEL. So from $Pr'(RatEL|E) = 0.5$ it follows that $Pr'(RatEL) = 0.5$

6. Some Consequences of Bayesian Defeatism

In this section I'll prove the following two consequences of Bayesian Defeatism:

Consequence 3 – *Blank Slate About Rationality:* Your prior credence in RatEL should be 0.5. That is, Pr(RatEL) = 0.5.

Consequence 4 – *Blank Slate Conditional Probabilities*: Your prior credence in L given E should be 0.5. That is, Pr(L|E) = 0.5.

Proof of Consequence 3: Pr(RatEL) = 0.5

Given that the defeatist thinks that learning H should bring the reasoning-for-fun agent's credence in RatEL down to 0.5 (Consequence 2), but also thinks that defeat should never induce a more dramatic change than reversion to priors (REVERSION TO PRIORS), it must be that the agent's *prior* credence in RatEL is 0.5. If the agent's prior credence in RatEL were anything *above* 0.5 – say it were 0.95 – then we'd be in a situation like PITTSBURGH-HYPOXIA: one in which, even in the presence of a defeater, your credence shouldn't zoom all the down to 0.5 (where would the 0.5 number even come from?) but to no less than whatever your prior happened to be. To think otherwise would amount to (as Isaacs (2021), Pittard (2019) and White (2009) point out) committing the base rate fallacy.

Proof of Consequence 4: *Pr(L|E) = 0.5*

From the fact that Pr(RatEL) = 0.5 (Consequence 3) we get that that for r>0.5

$$0.5 = Pr(Pr(L|E) = r)$$

From the fact that E isn't evidence for the rationality facts (Assumption 4), we get that:

$$Pr(Pr(L|E) = r) = Pr(Pr(L|E)=r|\ E)$$

From the equivalence of the propositions: {Pr(L|E) = r, Pr*(L|E) = r*} (Assumption 1) we get

$$Pr(Pr(L|E) = r\ |\ E) = Pr(Pr^*(L|E) = r^*\ |\ E)$$

Applying transitivity to the previous three equations we get that

$0.5 = Pr(Pr^*(L|E) = r^* \mid E)$

Because the propositions $\{Pr^*(L|E)=r^*, Pr^*(L|E)=1-r^*\}$ from a partition, it follows from the above equality that

$0.5 = Pr(Pr^*(L|E) =1- r^* \mid E)$

Where Pr_E is the credence function you ought to have upon learning E, BRIDGE tells us that $Pr_E(L) = Exp_{PrE}(Pr^*(L|E))$. And by conditionalization, $Pr_E(L) = Pr(L|E)$.
So we have that

$Pr(L|E) = Exp_{PrE}(Pr^*(L|E))$.

(In other words, your conditional credence in L given E should equal your *conditional-on-E* expectation of $Pr^*(L|E)$)

Since we've shown that: $Pr(Pr^*(L|E) =r^* \mid E) = Pr(Pr^*(L|E) = 1-r^* \mid E) = 0.5$

We have that:

$Exp_{PrE}(Pr^*(L|E)) = 0.5(r^*)+(0.5)(1-r^*) = 0.5$, and so by BRIDGE $Pr(L|E) = 0.5$

6. The Puzzle

Here's where we are: By appealing to BRIDGE, REVERSION TO PRIORS, and the four assumptions about the cases listed above I showed that the Bayesian defeatist's priors must be such that:

(a) $Pr(RatEL) = 0.5$, and
(b) $Pr(L|E) = 0.5$.

But now we have a problem because the story as told so far is inconsistent. Here's why:

(1) *Inconsistency* #1: We started with the stipulation that you know that E supports one of L or ~L, but we've just derived that it doesn't support either of L or ~L (For Pr(L|E) = 0.5).

(2) *Inconsistency #2*: We've also assumed classical conditionalization, and that it was rational for you to be confident that L upon learning E (and prior to learning H). But if Pr(L|E)=0.5 and you learn E, your credence should only be 0.5 in L.

(3) *An Additional Tension*: If we think that even once you've read this paper you should be sensitive to defeat, there is a further problem. The considerations here allow you to *deduce* that E supports neither L nor ~L. How then can uncertainty about *which* of L or ~L is supported by E be responsible for your reduction of confidence in L?

Arrival at an inconsistency is always a good time to make sure we like the premises that got us here. I won't say much about the four assumptions since, as I mentioned earlier, these are really more like stipulations. The premise *associated* with each assumption is, recall: "If defeatism is plausible, the defeatist verdict should apply to a case in which _____ holds." I also won't say much more here about REVERSION TO PRIORS since I think all *Bayesian* defeatists should agree that some such constraint is needed to avoid the base rate fallacy. But see Appendix 3 for a possible objection and response.

The remaining premise to resist is BRIDGE. As I mentioned at the outset, resisting the results by rejecting BRIDGE, will really involve rejecting the *particular applications* of it that are appealed to. So it might be helpful to see which entailments of BRIDGE actually played a role in the argument and see whether the defeatist might want to reject either of them.

BRIDGE was appealed to twice in the argumentation above. The first was simply to show how the principle motivated the defeatist verdict – that is, I used BRIDGE to derive Consequence 1 which says that "if your total evidence is E&H, it's rational to assign a 0.5 credence to RatEL if and only if it's rational to assign a 0.5 credence to L."

The puzzle itself though doesn't rely on Consequence 1 (which is a biconditional), but only on the claim that your credence in RatEL in HYPOXIA upon learning H should be 0.5. So one way the defeatist could reject where we ended up is by appealing to the following strategy:

Strategy #1: Claim that your credence in RatEL in HYPOXIA upon learning H should *not* be 0.5 – in fact it should be greater than 0.5.

While this is certainly a way one *could* go, it does make the defeatist verdict in the original case rather puzzling. If E in fact *does* support L, and, even once you learn H, you should still think it's more likely than not that it supports L, why would you be completely neutral about L?[12]

The second time BRIDGE was appealed to was in the derivation of Consequence 4 which says that Pr(L|E) = 0.5.

More specifically, the role that BRIDGE played in deriving Consequence 4 was in arguing for the following conditional claim:

If Pr(RatEL) = 0.5, then Pr(L|E) = 0.5. So the second BRIDGE-rejecting strategy would be:

Strategy 2: Reject that if Pr(RatEL) = 0.5, then Pr(L|E) = 0.5.

In fact, however, Strategy 2 won't work. For even without relying on BRIDGE we can get the result that if Pr(RatEL) = 0.5, then Pr(L|E) = 0.5. Here's how: On the Strategy 2 story you're meant to start out 0.5 confident that RatEL, but your credence in L given E should be high. Now imagine that you go on to learn E. Your credence in RatEL will still be 0.5, since, by Assumption 4, E isn't evidence for or against RatEL. So at this point you're still 50% confident that RatEL but more than 50% confident in L (because you conditionalized). But then what happens if you learn H? You're already only 0.5 confident that RatEL, so learning H can't induce any more uncertainty about what's rational than you already have. But, by Assumption 2, H only induces a reduction of confidence in L *by* inducing uncertainty about RatEL. This means that if H no longer induces any uncertainty about the rationality of believing L given E, then H will have no impact on your credence in L. And this of course is incompatible with defeatism. So, given the other assumptions

[12] Imagine a version of the case where the folks at ground control say: "you're flying at an altitude that puts you at risk for hypoxia. We have set our team of rationality experts to work and I *think* what they reported is that your evidence supports L. Things were a bit noisy though, so I might have misheard – I'll verify and get back to you shortly." At this point, suppose, your credence in RatEL is reasonably greater than 0.5. Will the defeatist really want to say that your credence in L should be 0.5? On what basis? Where would the number 0.5 come from when both your initial evidence, and your higher order evidence seem to favor L over ~L?

in place, Strategy 2 won't work. That is, even without accepting BRIDGE, the defeatist should think that if Pr(RatEL) = 0.5 then Pr(L|E) = 0.5.

7. Solving the Puzzle

7.1 The First Modification: Two Normative Notions

First, we'll need to introduce an additional normative notion. For if we think of "Pr" as the rational prior, and Pr behaves in the way the Bayesian defeatist wants it to, then the "Rat" in "RatEL" simply can't be talking about rationality in the sense of the "rational" prior Pr. For, at least once you're aware of the reasoning in this paper, and so you know that relative to the rational prior, Pr, E neither confirms nor disconfirms L, it can't be that uncertainty about whether Pr(L|E) is high or low is responsible for your reduction of confidence in HYPOXIA. To put the point another way, whatever notion of rationality involves *sensitivity* to defeat, cannot be notion of rationality, uncertainty about which *is responsible* for defeat. There are simply too many jobs Pr is being asked to do and it cannot do them all.

Once we add a second notion of rationality to the story many questions will arise about what distinguishes these two notions, how they work, and whether and to what extent they interact. But for now it will be helpful to just have names for them. I'll continue to call the defeat-sensitive notion of rationality – the notion that "Pr" describes – "rationality." I'll call the sense of rationality that "RatEL" is about – the notion, uncertainty about which, is *responsible* for defeat – "*Rat*inality" (to remind us of RatEL).

There is one question though about the relationship between Ratinality and rationality that needs to be settled before we go any further. And that is the question of which of these notions get plugged into BRIDGE. We can figure this out by recalling the work BRIDGE was doing for us. First, recall that RatEL is the proposition that H is meant to make you uncertain about. This means uncertainty about *Rat*inality is what's meant to drag down your credence in L. Second, note that the mechanism by which BRIDGE was supposed to convert higher order uncertainty into uncertainty about L, was by having the higher order uncertainty change your expectation of a certain credence function – the one we were calling Pr*. But now we can see that BRIDGE as stated simply can't work that way. For uncertainty about what's *Rat*inal won't result in any change in expectation to Pr*. This is because Pr*'s expertise does not concern Ratinality at all but rationality. The great insight that Pr* has to offer you is just that it's rational to regard E as neutral with respect to L. It won't tell you anything then about which of L or ~L is *Ratinal* to believe given E. This

means that if Ratinal uncertainty is going to play a role in an account of defeat we need to modify BRIDGE as follows:

Let P be the Ratinal ur-prior, and P* = P(· | P is Ratinal). Unlike Pr, which regards E as neutral with respect to L, P will be such that P(L|E) = high. Our new version of BRIDGE will then be:

> RATINAL BRIDGE: Where $Pr_{\mathcal{E}}$ is the rational credence function to adopt when your evidence is $\mathcal{E}$, $Pr_{\mathcal{E}}$ is such that: $Pr_{\mathcal{E}} = Exp_{Pr_{\mathcal{E}}}(P^*(\cdot \mid \mathcal{E}))$

RATINAL BRIDGE says that the *ratio*nal credence to adopt given $\mathcal{E}$ will be the same the rational credence function's expectation of the *Rat*inal credence function to adopt given $\mathcal{E}$, conditional on the Ratinality facts.

So the first modification to the picture involves rejecting BRIDGE and replacing it with RATINAL BRIDGE. Additionally, because we don't want to stipulate that you know that it's *ratio*nal to be confident in one of L or ~L given E, but only that it is *Rat*inal to, we'll want to replace Assumption 1 with Assumption 1' which differs from Assumption 1 only in that it replaces the "Pr"s with "P"s:

Assumption 1': Throughout the cases under discussion, for some real numbers *s* and *s** both greater than 0.5, you are certain that (a) P(L|E) = *s* if and only if P*(L|E) = *s** and (b) P(~L|E) = *s* if and only P*(~L|E) = *s**.

And the same will go for Assumption 2: rather than assuming that H induces uncertainty about rationality, we'll want to assume that H induces uncertainty about *Rat*inality. So the upshot of Assumption 2 will be what I'll call

Assumption 2': Pr(L|P is Ratinal & E&H) = Pr(L|P is Ratinal & E)

But there is another apparent problem that has yet to be addressed: Pr, recall, is meant to describe *ratio*nality. We showed that Pr(L|E) = 0.5, which means that if you get up on the plane and learn E, the rational credence in L to adopt will be 0.5. But if your credence in L begins at 0.5 (as is stipulated) and remains at 0.5 even once you learn E, then H won't have any defeating work to do. This means, then, that even once Ratinality is introduced into the picture to accommodate the fact

that there *is some sense* in which E supports one of L or ~L (and so learning H can induce uncertainty about *which* of L or ~L is *Ratinal*), since Ratinality doesn't concern the prior that the defeat-sensitive agent will be conditioning on throughout their epistemic life, we're still left wondering how the Bayesian defeat-sensitive agent who starts out with Pr, will get their credence in L up above 0.5 upon learning E.

7.2 The Second Modification: Judgments

In fact, though, I think this aspect of the puzzle is, in a certain sense, illusory. What I mean by this is that, even setting aside the results in this paper, the defeatist shouldn't be thinking that the *only* thing you learn on the plane is E. In particular, given that we're assuming that, in the cases at hand, the defeater defeats by way of telling you that your judgments are unreliable, and that there are no other ways that H and RatEL are evidentially linked (no crystal balls telling you that you'd find yourself hypoxic if and only if RatEL is false), defeatism will be most plausible when, in addition to coming to know E on the plane, you come to know what you judge.

To make this thought vivid, let's go back to REASONING-FOR-FUN. Call the credence function at the time during which your credence in RatEL is high "Pr_1" and your initial credence function, as before, will be Pr. Note that for H have its defeating effect we need that $Pr_1(RatEL|H) < Pr_1(RatEL)$. This will entail $Pr_1(H|RatEL) < Pr_1(H|{\sim}RatEL)$. In other words, once you have Pr_1 you should think that the probability that you're at risk for hypoxia is higher if L is likely to be true given E, than if ~L is likely to be true given E.

But why would a fact about Ratinality, *all on its own,* have any evidential relevance to a proposition about your cognitive capabilities? The most natural story to tell on which a Ratinality fact is evidence about your cognitive capacities is a story on which *you know which Ratinality-related proposition your cognitive capacities produced*. If you know that you judged RatEL, for example, then it's very sensible to think that, conditional on ~RatEL, there's a good chance your cognition is not working properly. But if you don't know what you judged, the bearing of RatEL on H is quite mysterious. As an analogy consider the following: suppose somebody told you that they take the fact that it's raining to be evidence that the meteorologist is unreliable. This would make the most sense if part of the background information includes the fact that the meteorologist claimed that it *wouldn't* rain. But suppose we added to the story that this person has no idea know what the meteorologist predicted. Then it would be very puzzling why they think rain is evidence of unreliability. (Maybe on rainy days people are in worse moods and think less clearly?)

All this is to say that defeatism makes the most sense from a Bayesian perspective when we assume that agents are aware of what they judged. So I'm going to assume the following:

Assumption 5: Upon forming a judgment concerning which of L or ~L is supported by E, you come to have as part of your evidence the proposition that you formed this judgment. [13]

With this in mind, we can put forward the following proposal, which will address the fact that your credences in L and RatEL end up high despite the fact that Pr(L|E) = 0.5 and Pr(RatEL) = 0.5

Proposal:
Where JEL is the proposition "I judge L to be likely given E"

Pr(L|E&JEL) = high
Pr(RatEL|JEL) = high

The upshot of this picture will be that it is evidence about your *judgments* that explains how your high confidence that RatEL, and your high confidence that L given E, arise from your blank-slate-ish priors. (Whether this is a desirable upshot or not, I'll leave for you to decide).

But at this point it's worth asking: what exactly is a judgment? The simplest account of a judgment might be something like a belief or a high credence. And if we didn't adopt Proposal this perhaps would be a fine way to think about judgments. However, if the Bayesian defeatist adopts Proposal, thinking of a judgment in this way is a bit awkward. JEL, recall, is meant to be the proposition that explains *how* your credence in RatEL and in L get above 0.5, given that your prior credence in RatEL, and in L given E is 0.5. There isn't anything full-out incoherent about the

[13] If you're somebody that gets squeamish about claims that suggest that our mental states are "luminous" (e.g. Williamson (2000)) you might not like Assumption 5 very much. So let me point out that the considerations above, concerning how H does and doesn't evidentially interact with RatEL, don't require agents to become *certain* of what they judged in order for H to have its defeating effect. The considerations do however require that they obtain *some* evidence which at least impacts the probability of what they judged upon forming a judgment – for without *some* change in opinion about what judgments were formed, a proposition about the reliability of *judgments* shouldn't have any bite (absent the deviant evidential routes we are intending to rule out). Nonetheless, it will be simplest to just assume that upon forming a judgment agents learn that they form these judgments, and the proposal I'll offer on behalf of the defeatist will proceed in that fashion. The proposal could, however, be modified by replacing JEL with some proposition which instead merely raises the probability of JEL. But if you think that in the case under discussion RatEL makes it likely that you're hypoxic, absent any information that bears on your judgments, you will be rejecting the stipulation I discussed in REVERSION TO PRIORS about how H and RatEL are evidentially linked.

claim that the proposition *that* you formed a high credence in RatEL is also the proposition that explains *why* you formed a high credence in RatEL. But it is awkward (A: Hey there B – why are so you confident that RatEL? B: Oh – simple– I conditionalized on the proposition that I'm highly confident that RatEL. A: Yeah, but why did you do that? B: Why did I conditionalize on my evidence? A: Well – I guess what I'm really asking is: how did it come to be your evidence that you formed a high credence that RatEL? B: That's simple too. I in fact *did* a form a high credence that RatEL and I was lucky enough to find out that I did. A: Yeah but *why* did you form it? B: I already told you– I formed it because I conditionalized on the proposition that I formed it. A: Huh?)

So given Proposal we might want to think of a judgment as something less committal than a belief: perhaps it's a *seeming as if* RatEL is true. But we needn't decide the matter here: there are no doubt a variety of ways you might think about what a judgment is in this context. What's crucial though for Proposal to work is that a judgment be something such that learning that you have it explains the increase in credence to RatEL.

I'll now state a consequence that follows from the adoption of Proposal. (If Proposal is rejected, the Bayesian needs to find some other proposition to do the relevant work).

> ***Consequence 6*** **–** *Believe What Seems Right*: If, upon considering E, you judge that E supports L, you should be confident that L, but if, upon considering E, you judge that E supports ~L you should be confident that ~L.

The proof of Consequence 6 is a bit more cumbersome than the previous proofs so I've chosen to relegate it (along with the two supplementary assumptions needed to derive it) to Appendix 2.

At this point you might have the following worry: Didn't we go along proving all sorts of results that assumed that the agent's total evidence was E or E&H (as opposed to E&JEL, or E&H&JEL)? And don't our assumptions concern rationality rather than Ratinality? Given the various modifications introduced here do the earlier proofs still work? I tie up these loose threads in Appendix 3. Let's now move on to consider the philosophical upshots of this picture.

8. Conclusion: What does the Resulting View Amount to Philosophically?

In this paper I presented some consequences of combining Bayesianism with defeatism. (And of course the results above won't be true of just L and E, but of any defeasible proposition-

evidence pairs with the relevant structural features). This combination of views has four features worth thinking through carefully.

(1) Bayesian defeatism requires making use of two notions: rationality and Ratinality.[14] On the resulting picture there is indeed a sense in which E supports L rather than ~L that will be encoded in the fact that the Ratinal prior which we called "P" is such that P(L|E) = high. But a Bayesian agent who starts out with P and updates by conditionalization won't be susceptible to defeat. So insofar as responding to defeat in the way the defeatist recommends is rational, there is some *other* probability function with something normatively good going for it, Pr, according to which E is completely neutral when it comes to L. But can we say anything about what notions P and Pr represent?[15] If one is an expressivist about normative notions like rationality, multiple versions of epistemic rationality are going to require distinct accounts of what sorts of attitudes are expressed by claims involving these notions, what distinct functional role they play and so forth. We might also wonder, insofar as these notions conflict, which we really should be guided by, or which to pursue for the purposes of getting at the truth.

(2) The second aspect of the view we might want to consider is the fact that a rational agent will have to start out with blank-slate-ish conditional credences. In HYPOXIA, for example, before learning E, your credence in L given E had to be 0.5. This might require a modification to how we think about the psychology of credence and conditional credence. For you might ask: what does it take for an agent to have a high conditional credence in L given E? If you think that we can have non-occurrent conditional credences, then these will be had, presumably, in virtue of certain dispositional facts about us. But given the Bayesian defeatist's picture this will get a bit complicated. Because if we look at the dispositional facts, it may well seem that the (rational) agents in these cases *do* have a high conditional credence in L given E. They might be disposed, for example, to take the

[14] This won't be the first time that appealing to multiple normative notions has been suggested to deal with issues surrounding defeat (though the reasons here are different). See for example van Wietmarschen (2013), Smithies (2015), Schoenfield (2015b, 2018), Steel (2018) and Worsnip (2018). Christensen (2010) while not going quite so far as saying that there two varieties of epistemic rationality says that responding in the defeatist way, while being the "epistemically best response" also involves falling short of some rational ideal. Horowitz (2019) and Lasonen Aarnio (2020) raise challenges to the "multiple normative notions" approaches.

[15] Note that we can't just think of P as the rationally omniscient version of Pr: in other words, P is *not* Pr(· | Pr is rational). For knowing that Pr is rational will just amount to knowing that E has no bearing on L, whereas P says E *does* have a bearing on L. The two notions are thus not in any straightforward way reducible to one another.

relevant kinds of conditional bets, or have whatever other dispositions you think go along with having a high credence in L given E.[16] So Bayesian defeatism may also incur some commitments about the psychology of non-occurrent credences (assuming you think such things exist).

(3) If the Bayesian defeatist accepts Proposal you might worry that "Believe What Seems Right" (Consequence 6) makes being rational too easy.[17] If the Baysesian defeatist rejects Proposal they will need to come up with some other proposition that is responsible for the credal bump in the cases under discussion.

(4) We started out with the thought that there was something called *higher order* defeat – we asked what the *rational* thing to believe is when we're uncertain about *what's rational*. It turned out though, that to say all the things we wanted to say, we, in a certain sense, got rid of higher order defeat. We no longer have a story about how to rationally be uncertain about what's rational – we have a story about how to rationally be uncertain about something else – what's *Rat*inal. But where does this leave uncertainty about *ratio*nality? Is it ever rational to be uncertain about what's rational rather than what's Ratinal, and if so, what does the theory of *rational* uncertainty look like? Is there some version of defeatism that applies to uncertainty about rationality? Will we need to retreat to some third normative notion to deal with this kind of uncertainty?

I don't want to claim that any of these issues pose an insurmountable challenge. What I've raised are simply a number of questions that a Bayesian defeatist will want to address in fleshing out the account. But these considerations may also suggest that something funny is going on in cases of defeat – something that is not a purely Bayesian phenomenon – and that there may be

[16] You might think that insofar as the reasoning process takes (a nontrivial amount of) time to perform the defeatist view is making the right predictions. After all, you might say, the agent won't be disposed to *immediately* take the relevant conditional bet, or to *immediately* to act as if L upon learning E. They'd need some time to think through the matter first. And perhaps we should only attribute to an agent a high conditional credence in L given E if they are disposed, to *immediately* [fill-in-the-blank]. While this may be a promising approach in the case at hand, I don't think it will work in general. For defeatists verdicts are also meant to apply to cases in which opinions are formed "instantaneously." For example, suppose we are watching a documentary about two children A and B. After the documentary I say: "one of these children commits a heinous crime in adulthood. Which do you think it is?" It may seem immediately obvious to you that it is A. But now imagine I tell you that the film has included subliminal messaging impacting your response, or that you're susceptible to certain relevant biases. If defeatists think their verdicts apply to such cases (and they do seem to think so), they won't be able to appeal to (non-trivial) stretches of time spent reasoning to explain the agnostic conditional credences in such cases.

[17] The concern that defeatist views will makes rationality "too easy" also appears in Kelly (2011) and Schoenfield (2015a). Although Horowitz and Sliwa (2015) present a version of defeatism which is meant to get around the concerns from Kelly and Schoenfield, the results here show that their suggestion won't be available to a *Bayesian* defeatist.

rich terrain to explore by considering the question of whether we might be *departing* from Bayesianism when faced with a defeater, and whether such departures can be explained and motivated. [18]

Appendix 1: Existing Arguments for a Tension Between Bayesianism and Defeat

In the literature on higher order evidence a number of problems have been raised for a Bayesian account of defeat. I do not think these arguments are decisive and in this appendix I'll explain why.

The first problem I'll discuss is what I'll call *the problem of independence*.[19] Here's the idea: Suppose that the whole flying episode in HYPOXIA happens on Monday. Before you ever enter this plane, say on Sunday, you had some conditional credence in L given E. (Maybe you never *consciously* considered the question of how likely it was that you'd have enough fuel to make it to L.A given that the dials and gauges are such and such, but deep down in your credal core – or so the Bayesian thought goes – you had a view of how likely L was given E). Now suppose we examine your credences on Sunday, pre-flight and consider how likely you think it is that you'd have enough fuel to make it to L.A given that the dials and gauges are such and such, *and* that you suffer some mental impairment on *Monday*. On a plausible spelling out of the case, adding the mental impairment fact does nothing: *On Sunday* you don't think that conditional on facts about your mental states *on Monday*, L is more or less likely given E. This then is the problem of independence: On Sunday you regard propositions concerning your Monday mental impairments as independent of L on the supposition that E. The Bayesian mechanism then, will require that once you get E on Monday, you continue to regard any Monday mental impairments as irrelevant to the question of whether you have enough fuel to make it to Los Angeles.

---

[18] Here are some views that involve a departure from classical conditionalization that one might wish to consider at this point. The most moderate modification would involve moving to a Jeffrey-based account. One thing that makes a Jeffrey-type account of defeat potentially both very easy and very unsatisfying is that using Jeffrey-conditionalization, you can move from any probability function Pr_1 to any other function Pr_2 (over the same algebra), simply by thinking of the new probability function as the input partition. Given this fact about Jeffrey-conditionalization, the question of whether it can accommodate defeat is in some sense quite uninteresting without imposing some constraints on what the input partition should be (see Weisberg (2015) and Miller (2016) for discussion of related points). Another view that involves a departure from classical conditionalization suggests that we think of what happens in cases of higher order defeat as information loss (Levinstein (ms.)). And yet another (Schoenfield (forthcoming) involves thinking of defeat as a departure from Bayesnism motivated by taking up what she calls "the perspective of doubt."

[19] For discussion of various versions of this problem see Christensen (1994, 2010), Weisberg (2015), Miller (2016), Schoenfield (2018), Levinstein (ms.), and for a response to Schoenfield (2018) see Bradley (2020).

There has been some discussion about whether the problem of independence could be solved by thinking of the higher order evidence as essentially indexical where the relevant proposition isn't "On Monday so and so is impaired" but rather "*I* am impaired *now*." (This suggestion first appears in Christensen (2010)). Even on Sunday, says this proposal, the proposition "*I* am impaired *now*" should be relevant to how likely you think L is given E. Whether this solution works depends on your views about how to update on essentially indexical evidence. Schoenfield (2018) argues that the indexical move doesn't help. Bradley (2020) argues that Schoenfield is wrong. The differences between them seem to turn on questions about how to deal with certain variants of the sleeping-beauty problem, and while these issues are certainly intriguing, I won't be delving into a discussion of them here. Suffice it to say, that given that there is a theory of self-location (Bradley (2020)) that allows the Bayesian to skirt the problem of independence, I think the arguments appealing to the problem of independence are not decisive.[20]

Second, you might think that Bayesianism is a theory of *ideal* rationality, whereas defeatism is really a view aimed at agents that are not ideal, and so any view which is going to be sensitive to defeat won't be a Bayesian one (see Smithies (2015) for remarks in this general spirit). However, most defeatists think of defeat as something that can afflict the most ideal of agents. This is because a case of defeat doesn't involve the agent *being* irrational but simply getting *evidence* that they are irrational. And no matter how rational you are, you are susceptible to the possibility of getting misleading evidence suggesting that you are not. (This point is made forcefully by Christensen (2008)).

Third, Pryor (2013) and White (2006) discuss a Bayesian challenge to defeat *for propositions for which we have immediate justification*. Pryor argues that any view which takes there to be some class of propositions that are defeasible but for which our justification is immediate (in a particular sense he discusses in his paper) will run into Bayesian challenges. (See Miller (2016) for a response to some of these). Pittard (2019) raises challenges for certain defeatists views concerning what describes as "fundamental" matters. Here I am focusing on a proposition L – that you have enough fuel to make it to Los Angeles – and that proposition is neither immediate nor fundamental the

[20] I'll also note here that although the problem of independence as stated can be avoided by introducing self-locating beliefs, I have struggled to fill in all of the details of a self-locating Bayesian account of defeat. There are many choice points concerning exactly which propositions are best thought of as self-locating, what attitudes should be had towards the non-self-locating versions of these propositions, and how best to account for the role our prior opinions play in the context of defeat when the defeating propositions are self-locating. This isn't to suggest that these problems are insolvable, but just to flag that there is, I believe, more work to be done in filling out a Bayesian account of defeat that appeals to self-location.

relevant senses. However, see Appendix 3, for a connection between the ideas in this paper and Pittard's discussion of fundamental disagreements.

Appendix 2 - Assumptions and Derivation of Consequence 6

> ***Consequence 6*** – *Believe What Seems Right*: If, upon considering E, you judge that E supports L, you should be confident that L, but if, upon considering E, you judge that E supports ~L you should be confident that ~L.

I make use of the proposal on behalf of the defeatist that I described in the main text and two additional assumptions to derive Consequence 6. The first assumption is:

Assumption 6: Pr(JEL) = 0.5.

Why is this plausible? First, note that it's impossible for Pr(JEL)>0.5. For by the theorem of total probability:

Pr(RatEL) = Pr(RatEL|JEL)Pr(JEL)+Pr(RatEL|~JEL)Pr(~JEL)

We know that Pr(RatEL|JEL) >0.5. (Proposal)
Let's call Pr(RatEL|JEL) = c.
We also have that Pr(RatEL) = 0.5. (Consequence 2)
So if Pr(JEL) = d > 0.5 we'd have

0.5 = cd + Pr(RatEL|~JEL)(1-d)
(0.5-cd)/(1-d) = Pr(RatEL|~JEL)

If d > 0.5, this would make Pr(RatEL|~JEL) < 0.

So the only way Pr(JEL) could fail to be 0.5 is if Pr(JEL) were *less* than 0.5. Recall however that we've stipulated in the REASONING-FOR-FUN case that you're certain you'll judge either that L is likely given E or that ~L is likely given E. So if Pr(JEL) is less than 0.5, that means you think it's

more likely than not that you'll judge that ~L is likely given E. Perhaps there are some spellings out of the case in which it's rational to think that probably you'll judge that ~*L* is likely given E (despite the fact that it's *L* that is likely given E). But it certainly seems like in at least *some* natural ways of spelling things out, given that you're 0.5 about which of L or ~L is likely given E, you'll also assign a 0.5 credence to which of L or ~L you'll *judge* E to support once you consider the matter.

We need one more assumption for the proof. What this final assumption says is that, conditional on the Ratinality facts and E, propositions about your Ratinality-judgments have no relevance to L. More specifically, the thought is that insofar as JEL has any relevance at all to the proposition that you have enough fuel to make it Los Angeles in the presence of E, it has that relevance *through* bearing on RatEL. This means that a credence function that was already certain about RatEL and E, won't take facts about what you judge about Ratinality to either confirm or disconfirm L. Thus:

Assumption 7: Where P is the Ratinal prior:
P(L|E& P is Ratinal) = P(L|E& P is Ratinal & JEL)
P(L|E& P is Ratinal) = P(L|E& P is Ratinal & ~JEL)

Using these assumptions we can prove *Believe What Seems Right*:

Proof:
By the theorem of total probability:

(*) Pr(RatEL) = Pr(RatEL|JEL)Pr(JEL) + Pr(RatEL|~JEL)Pr(~JEL).

Let Pr(RatEL|JEL) = c > 0.5 (by Proposal)

Recall that Pr(RatEL) = 0.5 (Consequence 3) and Pr(JEL) = 0.5 (by Assumption 6). Plugging in these values to (*) we get:

0.5 =(0.5)c + (0.5)Pr(RatEL|~JEL)

It follows that Pr(RatEL|~JEL) = 1-c. And so

$Pr(\sim RatEL | \sim JEL) = c$

Since in REASONING-FOR-FUN it's stipulated that you know that the E supports L or E supports ~L,[21] and that you'll judge either that E supports L or that E supports ~L it follows from the above equality that:

$Pr(\text{E supports } \sim L | \text{I judge E supports } \sim L) = c$

By Assumption 5, if you judge that E supports ~L, you'll have ~JEL as part of your evidence. Thus, if you judge that E supports ~L, conditionalization will require (given the above equality), that your credence that E supports ~L is *c*, (which, recall, is the same as what your credence in RatEL should be if you judge that E supports L).

So suppose you go ahead and judge that ~L is likely given E. We've established that you should be *c* confident that E supports ~L. But what you should think about ~L if you go on to learn E? Here's what we'll need to figure this out:

(1) Your evidence at this point is E&~JEL. (Stipulation and Assumption 5)

(2) RATINAL BRIDGE $Pr_{\mathcal{E}} = Exp_{Pr_{\mathcal{E}}}(P^*(\cdot \,|\, \mathcal{E}))$

It follows from (1) and (2) that your credence in L, having judged that ~L is likely given E should be your expectation of $P^*(L|E\&\sim JEL)$. By Assumption 7, since P^* is certain of the Ratinality facts, $P^*(L|E\&\sim JEL) = P^*(L|E)$.

So to figure out your credence in L, we need to figure out your expectation of $P^*(L|E)$.

By Assumption 1', you're certain, for some $s>0.5$, that $P^*(L|E) = s^*$ or $P^*(L|E) = 1\text{-}s^*$.

[21] Recall that the usages, in this context, of "supports" and "likely" all concern *Rat*inality – for you may well know that L is neither likely nor unlikely given E understood in the sense of *ratio*nality.

And we've shown that for c>0.5 your credence in ~RatEL (that is, the proposition that P(L|E) = 1-s) is c. Also by Assumption 1' we have that P(L|E) = s if and only P*(L|E) = s* and P(L|E) = 1-s if and only if P*(L|E) = 1-s*. This means that your expectation of P*(L|E) = c(1-s*)+(1-c)(s*).

Since we showed that your credence in L if you judge that ~L is likely given E and your evidence is E, should be your expectation of P*(L|E), it follows that your credence in L in this case should be c(1-s*)+(1-c)(s*) = c+s*-2cs*. And so your credence in ~L should be 1-(c+s*-2cs*).

Recall that c was stipulated to be what your credence in RatEL should be if you judge that L is likely given E.
So, in the original case, where you judge (truly) that L is likely given E, your credence in L should be your expectation of P*(L|E) which will be: cs*+(1-c)(1-s*) = cs*+1-s*-c -2cs* = 1-(c+s*-2cs*).
This is the same as what your credence in ~L should be if you judge that ~L is likely given E.

Thus, whichever of L or ~L you judge is likely given E, you should be confident in that proposition to degree 1-(c+s*-2cs*).

Appendix 3: Some Loose Threads

The other loose thread that needs tying up concerns how the assumptions and proofs go through once we add the modifications described in section 7 – in particular, Assumption 5 which says that your evidence will include facts about what you judge, and the addition of Ratinality as a distinct normative notion.

There are two junctures at which the modifications become relevant. The most important is that Consequence 1 will need to be replaced with Consequence 1' which will say: "If your total evidence is E&H&*JEL*, then it's rational to assign a 0.5 credence to RatEL if and only if it's rational to assign a 0.5 credence to L." The proof goes through in the same way as the proof of Consequence 1 except that in addition to Assumption 2' (a modification of Assumption 2 to deal with the shift to Ratinality, see p.21), we'll add Assumption 2": Pr(L|P is Ratinal & E&H&*JEL*) = Pr(L|P is Ratinal & E). The justification for Assumption 2" is that just as we are assuming that H has no relevance to L conditional on facts about what the Ratinal ur-prior is (which is what Assumption 2' says), facts about what you *judge* to be Ratinal also have no relevance to L conditional on facts about what the Ratinal ur-prior is. From Assumptions 2' and 2" we'll get that

P*(L|E&H&JEL) =P*(L|E). And from this, in combination with RATINAL BRIDGE, we get that your credence in L if your evidence is E&H&JEL should be your expectation of P*(L|E). The rest of the proof proceeds in a similar fashion to the original proof. So Consequence 1' together with the defeatist verdict will give us that Pr(RatEL|E&H&JEL) = 0.5. Given that E isn't evidence for or against RatEL this will also give us that Pr(RatEL|H&JEL) = 0.5.

The only other modification that needs to be made concerns the motivation offered for Assumption 4. Assumption 4 says that E isn't evidence for or against RatEL. In motivating Assumption 4, I gave some reasons for thinking that usually, a body of evidence doesn't confirm a proposition about what's *rational* to believe in response to that evidence. But now the relevant question will be whether, a body of evidence confirms a proposition about what's *Ratinal* to believe given that evidence. Since we don't know much about Ratinality, any such general claim is hard to assess. However, we don't actually need to assess the claim in general. The first reason for this is that, as with all the assumptions, the question we need to ask is: will the defeatist verdict apply to *some* cases in which the assumptions hold (for then the various results I derive will still hold true of those cases). The other reason we don't need to assess the general claim is that we can think of the dialectic of the paper as a kind of reductio. Starting out with a single notion of rationality leads us into contradiction. So a modification to the standard picture needs to be made for the Bayesian to account for defeat.

Finally, I want to consider a way one might try to resist REVERSION TO PRIORS as applied to RatEL. John Pittard (2019) also argues that REVERSION TO PRIORS will be a commitment of Bayesian defeatism. However, he claims that when it comes to fundamental matters, defeat simply doesn't make sense because there are no " priors" to revert to. [22] So one might resist my argument

[22] See his paper for a detailed argument for the claim that defeat doesn't make sense in such cases but here's one way one might get at the point: recall that REVERSION TO PRIORS was motivated by the thought in the cases under discussion the *only* way in which the defeater defeats, is by making it the case that your attitude concerning P provides no evidence for P. (We are intending to rule out deviant routes like the one in which drunkenness makes it likely that someone has played a joke on you). In the original case of the shochet, where it's plausible that the shochet should revert to his priors (because jokes *aren't* probabilified by drunkenness) the argument for reverting to priors appealed to the fact that Moishe's priors are such that Pr(I'm brought a kosher animal|I'm drunk)= Pr(I'm brought a kosher animal|I'm not-drunk). The manner, then, by which "I'm drunk" leads Moishe to reduce confidence in the proposition that he was brought a kosher animal, it that Moishe is also such that Pr(I'm brought a kosher animal|I'm drunk and I *judge that the animal is kosher*) < Pr(I'm brought a kosher animal|I'm not drunk and *I judge that the animal is kosher*). But notice that if Pr already knew what Moishe judged, this difference would disappear. We couldn't claim *both* that Pr(I'm brought a kosher animal|I'm drunk)= Pr(I'm brought a kosher animal|I'm not-drunk) – in order to rule out the deviant evidential routes – *and also* that Pr(I'm brought a kosher animal|I'm drunk and I judge that the animal is kosher) < Pr(I'm brought a kosher animal|I'm not drunk and I judge that the animal is kosher). So if your priors know their own opinions, then, while it's certainly open for you to say that drunkenness is evidence against kosherness, what's not so easy to say is: drunkenness is evidence against kosherness *because it undermines the reliability of your judgments.* Another

by saying that RatEL can't be defeated because it is fundamental in Pittard's sense. But this won't solve the problem for the Bayesian defeatist. For if RatEL is fundamental in some way that makes it indefeasible, then, given the assumptions in this paper, it will turn out that L isn't defeasible either. In other words, if you think that RatEL is a fundamental and thereby indefeasible proposition, the upshot of this paper for you will be that, given the enkratic commitments I described, higher order defeat is impossible for non-fundamental propositions as well.

Bibliography

Avnur, Y. and Scott-Kakures, D. (2015). "How irrelevant influences bias belief" *Philosophical Perspectives* 29 (1):7-39

Bradley, D. (2020). " Self-Locating Belief and Updating on Learning. *Mind* 129 (514):579-584

Christensen, D. (1994). "Conservatism in Epistemology." *Noûs* 28(1): 69-89.

Christensen, D. (2008). "Does Murphy's Law Apply in Epistemology?" *Oxford Studies in Epistemology 2*: 3-31. Oxford University Press.

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

Christensen, D. (2012). "Rational Reflection." *Philosophical Perspectives: Epistemology* 24: 121–140.

Christensen, D. (2019). "Formulating Independence" In M. Skipper & A. Steglich-Petersen (eds.), *Higher-Order Evidence: New Essays.* Oxford University Press Oxford University Press

Christensen, D. (forthcoming). "Akratic Epistemic Modesty." *Philosophical Studies*.

Coates, A. (2012). "Rational Epistemic Akrasia". *American Philosophical Quarterly* 49 (2): 113–124.

Dogramaci, S. (2018). "Solving the Problem of Logical Omniscience." *Philosophical Issues* 28 (1):107-128

Dorst, K. (2019). "Higher Order Uncertainty." In M. Skipper & A. Steglich-Petersen (eds.), *Higher-Order Evidence: New Essays.* Oxford University Press, 35–61.

Dorst, K, Levinstein, B., Salow, B., Husic, B.E. and Fitelson, B. (ms.). "Deference Done Better."

---

way to put it is this: if defeaters are only meant to defeat by removing the evidential relevance of your attitudes, but the "earliest" probability function around is already certain of what those attitudes are, talking about the effect of "removing" their relevance becomes tricky on a Bayesian picture. (One might think of this as related to the problem of old evidence).

Elga, A. (2012). "The Puzzle of the Unmarked Clock and the New Rational Reflection Principle" *Philosophical Studies* 164 (1):127-139

Elga, A. (ms.). "Lucky to be Rational."

Elga, A. and Rayo, A. (ms.). "Fragmentation and Logical Omniscience."

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

Gaifman, H. (2004) "Reasoning with limited resources and assigning probabilities to arithmetical statements." *Synthese*, 140(1/2):97–119.

Garber, D. (1983) Old evidence and Logical Omniscience in Bayesian Confirmation Theory. In J. Earman, (ed.) *Minnesota studies in the philosophy of science*, Volume 10. University of Minnesota Press.

Greco, D. (2014). "A Puzzle about Epistemic Akrasia." *Philosophical Studies* 167(2):201-219.

Hacking, I. (1967). "Slightly more realistic personal probability." *Philosophy of Science*, 34(4):311–325.

Halpern, J.Y. and Pucella, R. (2011). Dealing with logical omniscience. *Artificial Intelligence*, 175(1):220–235.

Horowitz, S. (2014). "Epistemic Akrasia" *Noûs* 48(4): 718-744.

Horowitz, S. (2019). "Predicably Misleading Evidence." In M. Skipper & A. Steglich-Petersen (eds.), *Higher-Order Evidence: New Essays.* Oxford University Press.

Horowitz, S. and Sliwa, P. (2015). "Respecting all the Evidence." *Philosophical Studies* 17 (11):2835-2858

Isaacs, Y. (2021). "The Fallacy of Calibrationism." *Philosophy and Phenomenological Research* 102 (2):247-260.

Kelly, T. (2011). "Peer Disagreement and Higher Order Evidence." In A.Goldman and D. Whitcomb (eds.) *Social Epistemology: Essential Readings*. Oxford University Press.

Lasonen Aarnio, M. (2014). "Higher Order Evidence and Limits of Defeat." *Philosophy and Phenomenological Research* 88 (2):314-345.

Lasonen Aarnio, M. (2020). "Enkrasia or Evidentialism? Learning to Love Mismatch." *Philosophical Studies* 177 (3):597-632.

Levinstein, B. (ms.) "Higher Order Evidence as Information Loss."

Miller, B.(2016). "How to Be a Bayesian Dogmatist." *Australasian Journal of Philosophy* 94 (4):766-780

Pettigrew, R. (forthcoming). "Logical Ignorance and Logical Learning." *Synthese*

Pittard, J. (2019). "Fundamental Disagreements and the Limits of Instrumentalism." *Synthese* 196 (12):5009-5038

Pryor, J. (2013). "Problems for Credulism." In Chris Tucker (ed.), *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism*.

Rasmussen, M.S, Steglich-Petersen, A. and Bjerring, J.C. (2018). "A Higher Order Approach to Disagreement." *Episteme* 15 (1):80-100.

Schoenfield, M. (2015a). "A Dilemma for Calibrationism." *Philosophy and Phenomenological Research* 91(2): 425-55

Schoenfield, M. (2015b). "Bridging Rationality and Accuracy." *Journal of Philosophy* 112 (12):633-657.

Schoenfield, M. (2018). "An Accuracy Based Approach to Higher Order Evidence." *Philosophy and Phenomenological Research* 96(3): 690-715

Schoenfield, M. (forthcoming). "Meditations on Beliefs Formed Arbitrarily."

Seidenfeld, T. Schervish, M.J. and Kadane, J.B. (2012). "What kind of uncertainty is that? Using personal probability for expressing one's thinking about logical and mathematical propositions". *Journal of Philosophy*, 109(8):516–533.

Skipper, M. and Bjerring J.C. (forthcoming). "Bayesianism for nonideal Agents." *Erkenntnis*

Smithies, D. (2015) "Ideal Rationality and Logical Omniscience." *Synthese* 192.9: 2769-2793.

Steel, R. (2018). "Anticipating Failure and Avoiding It." *Philosophers' Imprint.*

Stalnaker, R. (1991). "The problem of logical omniscience." *Synthese*, 89(3):425– 440.

van Wietmarschen, H. (2013). " Peer Disagreement, Evidence, and Well-Groundedness." *Philosophical Review* 122 (3):395-425

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

Weatherson, B. (ms.). "Do Judgments Screen Evidence?"

Weisberg, J. (2015). "Updating, Undermining and Independence." *British Journal of Philosophy of Science* 66(1): 121-159.

White, R. (2006). "Problems for Dogmatism." *Philosophical Studies* 131 (3):525-557.

White, R. (2009). "On Treating Oneself and Others as Thermometer" *Episteme* 6(3):233-250

White, R. (2010). "You just believe that because…" *Philosophical Perspectives* 24(1):573-615.

Williamson, T. (2011). "Improbable Knowing." In T. Dougherty (ed.), *Evidentialism and its Discontents*. Oxford University Press

Williamson, T. (2014). "Very Improbable Knowing." *Erkenntnis* 79 (5):971-999

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