A Challenge for Approximate Coherentism

Abstract

In a series of articles, Glauber De Bona and Julia Staffel argue that imperfect Bayesian agents ought to hold credences which are approximately coherent, where the degree of incoherence is measured by divergence from a closest coherent credence function. This norm is intended to be a generalization of Joyce's coherence norm to non-ideal Bayesian agents. I argue, however, that as currently developed it cannot play this role. Whereas Joyce's coherence norm is stable under updating, in the sense that our verdict about the epistemic quality of a credence function over time does not depend on facts outside of the agent's control, De Bona and Staffel's approximate coherence norm is not.

Introduction

Glauber De Bona and Julia Staffel argue that the credences of a non-ideal Bayesian agent ought to be approximately coherent, where coherence is evaluated by an appropriate measure of divergence, such as normed distance or Kullback-Leibler divergence, from the closest coherent probability distribution (Staffel, 2015; De Bona and Staffel, 2017, 2018). I'll call this norm *Approximate* (italicized terms will be defined in the next section). It says that from a normative perspective, if credence function B is closer to coherence than credence function A, then credence function B is to that extent to be preferred to/ judged better than A. De Bona and Staffel show that for every incoherent credence function there exists an accuracy dominating less incoherent one. As a result, they propose Approximate as a generalization to the non-ideal context of its categorical relative, Joyce's *Coherence* norm, which says that for every incoherent credence function there exists an accuracy dominating coherent one (Joyce, 1998, 2009).

Both Coherence and Approximate should be distinguished from externalist criteria, such as truth or accuracy norms, which say that as between two beliefs or credence functions, the one that is in fact true or more accurate is better. What makes Coherence and Approximate different from truth or accuracy norms is that as regulative ideals, Coherence and Approximate are not susceptible to misfortune. I'll call this requirement *Weak Credal Internalism*, for reasons that will become clear below. I will also make it more precise. For now, the best way to illustrate the concern I have in mind is by example.

Consider, in ordinary decision-making, a norm that requires an agent to maximize actual utility. This is an externalist norm, in the sense that which option maximizes actual utility depends on external facts about the world, of which decision-makers are typically uncertain. In ordinary decision-theory, we would in general not use such a norm to evaluate the quality of an agent's choices. Rather, we subject them to a weaker, internalist norm,

namely, that of maximizing expected utility. We ask: did the agent do the best they could with the information they had available to them at the time of the decision? We do not want to punish them, so to speak, for factors that were outside of their control. Perhaps more accurately: we do not think that such factors should adversely affect our normative assessment of the quality of their choices.

I will suggest that Approximate, as a norm for evaluating the quality of an agent's beliefs, is more akin to the requirement to maximize actual utility in decision-making than it is like the requirement to maximize expected utility. In certain situations, it can punish agents for factors outside their control. In particular, it is possible to have the following situation: two agents, A and B, are such that B is currently less coherent. They perform a simple experiment (e.g. toss a coin) and update their credences by conditioning on the outcome. If the coin lands on tails, A becomes less coherent whereas if it lands on heads, B remains less coherent. Therefore, it is possible for either agent to observe an unfortunate sequence of data, from their perspective, which adversely affects our assessment of the quality of their doxastic state. A can go from doing better, to doing worse, solely due to the outcome of the coin toss. This is not possible for ordinary Coherence. If an agent starts with a coherent credence function, and updates their credences by Bayes' Rule, there is no possible world in which they end up with an incoherent credence function.

The paper proceeds as follows. In Section 1, I explore the normative status of accuracy and coherence/ approximate coherence. In Section 2, I define both norms carefully and describe their relationship to accuracy considerations. In Section 3, I motivate and explain the meta-norm I call credal internalism. In Section 4, I show that Approximate violates a minimally strong version of credal internalism. In Section 5, I offer some brief remarks on whether non-ideal theories of Bayesian rationality should be graded or not and consider some potential responses to the concern identified here. Section 6 concludes.

1 Accuracy and Coherence

DeBona and Staffel give two types of arguments for approximating coherence. One is a pragmatic defense – namely, that reducing one's degree of incoherence decreases the extent of their vulnerability to dutch books. I will not address that argument here. Indeed, I find it compelling – if one is exposing themselves to sure loss, it should be better to expose themselves to less of it. However, there remains a question about whether approximating coherence can be defended on non-pragmatic terms. This has been a common topic of discussion with respect to other norms in formal epistemology. For example, Joyce (1998, 2009) gives a non-pragmatic defense of probabilism and Greaves and Wallace (2006) give a non-pragmatic defense of updating by Bayesian conditioning. Fittingly, DeBona and Staffel also offer an epistemic defense for approximating coherence – namely, that approximating coherence improves one's accuracy outcomes. This is the issue I take up in this paper. In this section, I draw a distinction between two types of normative properties in epistemology – those which are of primary value, like truth and accuracy, and regulative ideals – i.e., those which, if followed, are in some way conducive to the things we value in and of themselves, like justification and coherence.

Joyce (1998) introduces the guiding ideal for Bayesian epistemology, namely, the norm of graded accuracy. The norm of graded accuracy implies that as between two credence functions, the one that is closer to the truth, where closeness is evaluated by a suitable measure of accuracy, is to be preferred. Accuracy is ordinarily evaluated using a loss function or scoring rule, which maps a pair of values – a probability assignment and the true state of the world – to a positive real number. For example, if someone forecasts the probability of rain tomorrow to be 0.8, a simple way to measure their accuracy would be the squared distance from the true outcome, $(I_v - 0.8)^2$, where I_v is an indicator variable that takes the value 1 if it rains and 0 otherwise.[1] This score, known as the Brier score, maps probability assignments to the closed unit interval where 0 denotes perfect accuracy and 1 denotes maximum inaccuracy (Brier, 1950).

Graded accuracy is similar to Goldman (2002)'s notion of veritism in ordinary epistemology (Pettigrew, 2017). For most subjective Bayesians, it is a feature of primary epistemic value regarding an agent's credal state. It is better to be closer to the truth than to be further away from it. However, graded accuracy is not a regulative ideal. By this I mean that an agent is not deemed epistemically irrational because she is inaccurate. We recognize that it is possible for an agent to formulate her credences diligently but end up inaccurate nonetheless. For example, suppose Alice is making a forecast about the weather tomorrow in a rainy city during a rainy season. She estimates the probability of rain to be 0.8. This is consistent with her evidence, including professional forecasts. We can further suppose that she updated her estimates by Bayes' Rule, that she is perfectly coherent, etc. In other words, she has done everything as well as we can expect. But in those improbable worlds where it does not rain tomorrow, Alice will be quite inaccurate, despite the fact she does not seem to have done anything wrong, so to speak, from an epistemic perspective. We can end up inaccurate by getting unlucky. Ordinarily, we do not hold such misfortune against an agent's epistemic rationality. Indeed, it is one of the lessons of Newcomb style problems that it does not always pay to be rational. This remains true when the 'payment' is given in terms of accuracy.

Rather, we use graded accuracy, together with an appropriate decision rule, in order to identify certain regulative ideals or norms for the assessment of an agent's credal state. For instance, Joyce (1998, 2009) defends coherence (the regulative norm) by showing that every incoherent credence function is dominated[2] (the decision rule) by some coherent credence

[1] There exist some important constraints on the shape of permissible scoring rules, some of which have drawn criticism, but since De Bona and Staffel generally agree with the formal setup laid out in Joyce (2009), I will not discuss them at length here. For the interested reader, the scoring rule on Joyce's framework should be continuous, monotonic, and strictly proper. Squared distance satisfies these conditions. For a recent discussion of these properties see Babic (2018). For a critical perspective on strict propriety, see Blackwell and Drucker (2018).

[2] Dominance says that if one credence function is more accurate than another in at least one world and at least as accurate in every other world, then the other credence is epistemically defective, because it exposes the agent who adopts it to sure loss. This is intended as a floor on epistemic rationality – i.e., while we may disagree with respect to many decision rules, everyone ought to agree, at a minimum, that a dominated credence function is defective, provided they endorse the measure with respect to which the credence function is dominated. Note that dominance can be taught of as an especially strong case of expected value maximization (EVM). Anyone who accepts EVM will accept dominance (but not vice versa).

function with respect to the underlying measure of accuracy (the feature of primary epistemic value). If an agent is incoherent then we say she is to that extent epistemically irrational. Likewise, Greaves and Wallace (2006) defend Bayesian updating (the regulative norm) by showing that under certain relatively general circumstances, for every learning experience, updating by Bayes' Rule maximizes the posterior expected value (the decision rule) of an underlying measure of accuracy (the feature of primary epistemic value).

2 Coherence and Approximate Coherence

In this section, I make the relevant notions of coherence and approximate coherence more precise and suggest that approximate coherence, as De Bona and Staffel characterize it, is intended to be a generalization to the non-ideal context of its categorical relative – i.e., a regulative norm justified on the basis of accuracy considerations. I consider cases where we are formulating a credence about a real-valued unknown one-dimensional parameter. For example, the unknown bias of a certain coin, the stylized objective chance of rain tomorrow, or the unknown mean of a normally distributed random variable. I define the relevant concepts accordingly (similar definitions could be given for discrete valued parameters, or for vector values ones).

Coherence Norm.

Let Ω denote our parameter space containing all possible values of the true unknown quantity $\theta \in \mathbb{R}$ for a one-dimensional Bayesian inference problem. A function of θ, $p(\theta)$, is coherent if (1) $p(\theta) \geq 0$ and (2) $\int_{\Omega} dp(\theta) = 1$.[3]

A coherent function is to be preferred to an incoherent one.

Coherence, as noted, is valuable because it guarantees improvements in accuracy. If we measure accuracy with a truth-directed, continuous and strictly proper score, then for every incoherent credence function there exists an accuracy dominating coherent one. Coherence plays a role in Bayesian epistemology that is similar in some respects to justification in traditional epistemology. Holding justified beliefs is valuable because this promotes the ultimate goal of believing true propositions. Holding coherent credences is valuable because this promotes the ultimate goal of holding accurate credences. De Bona and Staffel propose the norm of approximate coherence as a suitable alternative to coherence for ordinary non-ideal agents. Since it is unrealistic to expect ordinary agents to be coherent, they suggest, we can evaluate them in terms of how closely they approximate the ideal. On this approach, approximating coherence similarly promotes the ultimate goal of holding accurate credences. We define this norm using the same framework as above.

Approximate Norm.

Let Ω denote our parameter space containing all possible values of the true unknown quantity $\theta \in \mathbb{R}$. Let $f(\theta)$ be a function that violates (1) or (2), from

[3]The two conditions above commit us to countable additivity as well since the density of a countable union of disjoint regions is the sum of the densities of the individual regions due to the linearity of integration.

above, and $p(\theta)$ a probability function. Let $D : f \times g \to \mathbb{R}^+$ be a measure of divergence between f and g. The incoherence of f according to D is measured as $I_D(f) = \arg\min_p D(f, p)$.

As between two credence functions f and g, if $I_D(f) < I_D(g)$ then f is to that extent to be preferred to g.

De Bona and Staffel defend Approximate by relying on the property of final epistemic value, accuracy, in the same way that Joyce defends Coherence. In particular, De Bona and Staffel (2018) show that as between two incoherent credence functions, the one that is closer to coherence guarantees improvements in accuracy, as measured by a suitable divergence function (Proposition 2). In general, a divergence function is suitable if its associated scoring rule satisfies the conditions required by Joyce (2009) (continuity, truth-directedness, and strict propriety) and it is additive. A scoring rule is additive if we evaluate the inaccuracy of a forecast by adding up the inaccuracies of the probabilities assigned to every possible outcome in the relevant partition or algebra of events, rather than just the outcome that in fact occurred.

3 Credal internalism

In this section, I explain what I mean by credal internalism, both weak and strong. To start, consider the following from Wedgwood (2002),

> [T]he rationality of a belief supervenes purely on 'internal facts' about the thinker's mental states ... not on facts about the external world ...Moreover, this seems to be a completely general feature of rationality ... When we assess a choice or decision as rational or irrational, we are assessing it on the basis of its relation to the agent's beliefs, desires, and other such mental states – not on the basis of its relation to facts about the external world that could vary while those mental states remained unchanged (pgs. 349-350, emphasis added).

The insight I want to highlight, and endorse, from this passage is that the rationality of a credence should generally depend on the agent's mental state, rather than external facts outside her control. The clearest illustration of this guideline is in ordinary decision theory, where we focus on whether an agent can be represented as maximizing expected utility, not on whether they in fact maximized actual utility. We generally stay away from actual utility because we recognize it is possible for an agent to be maximally diligent and yet have nature deal her an unlucky hand such that her diligent choice produces low utility. What I am particularly interested in is how a credence function evolves over time. This will require a little bit more care in how we define the internalist guideline that Wedgwood highlights. In particular, we will see that it is difficult to sever all connections to facts about the external world. Instead, we want to make this relation to external facts as weak as possible.

In general, I want to say that when an agent updates their beliefs under conditions of uncertainty, our assessment of their credence function should not depend on contingent

facts about the external world, just as our assessment of a decision does not depend on contingent facts about the external world. However, there are two ways we can characterize this dependence relation. First, we can say that a regulative norm for the assessment of an agent's credence function should not produce verdicts regarding the quality of the agent's credences that depend in any way on contingent external facts. In other words, such facts are strictly irrelevant to the assessment and could never affect our verdict about the quality of a credence function. I'll call this strong credal internalism (SCI).

> **Strong Credal Internalism**.
>
> Let X be a random variable representing an uncertain event, whose possible realizations are denoted by an integer from the finite set $\{1, 2, 3, ...N\}$ (X does not have to be discrete or finite valued, this is to keep things simple for now). Suppose further we have an agent who has a credence function defined on X, $c(X)$.
>
> An assessment regarding the quality of $c(X)$ should not produce a verdict that is sensitive to which value from $\{1, 2, 3, ...N\}$ X in fact realizes.

This is similar in some respects to the likelihoodist dictum in statistical theory that test statistics should not depend on irrelevant information. In particular, it resembles Birnbaum (1962)'s conditionality principle, which implies that our inferences should not depend on arbitrary features of an experiment. To illustrate, suppose that before performing an experiment, we have to choose which lab to conduct the experiment in, A or B. The labs have similar instruments, but are slightly differently calibrated, with slightly different measurement error rates, and so forth. To decide which lab to use, we flip a coin. The outcome suggests lab B. We now perform the experiment in lab B and obtain data $\mathbf{x}_B$. We make an inference about an unknown quantity of interest, θ. The fact that we did not perform the experiment in lab A, where we would have observed data $\mathbf{x}_A$, should be irrelevant to our inferences about θ. This example resembles the one given in Cox (1958) and is an instance of the so-called weak conditionality principle (WCP).

WCI is similar to WCP in the sense that if we say an agent's credence function is rational/irrational after updating on $\mathbf{x}_B$, it should not be possible for this verdict to change had we instead updated on $\mathbf{x}_A$. Notice that the issue here is not about misleading evidence. The idea is that if a credence function about the bias of a coin is said to be rational/irrational after updating on $\langle Heads, Heads, Tails\rangle$ it should be equally rational/irrational had the agent updated instead on $\langle Tails, Heads, Tails\rangle$, or any other permutation of the three tosses. Of course, an agent's actual accuracy can be affected by misleading evidence. For instance, if they observe a sequence of ten heads with a coin that is heavily biased toward tails – this is unlikely, but possible.

SCI implies that an agent cannot get 'lucky' and receive a more favorable assessment of their credences, and it implies that an agents cannot get 'unlucky' and receive a less favorable assessment of their credences, as a result of the way the world shapes up to be. SCI is not implausible. At first blush at least, this sounds like the sort of condition an internalist standard of evaluation should meet. Indeed, the passage from Wedgwood suggests something of this sort. However, even Joyce's Coherence norm cannot satisfy this high standard. As

we will see below, it is possible for an agent to start with an incoherent credence function, update by Bayes' Rule, and end up with a coherent one, if the priors they start with and the evidence they observe are both just right. This is a rare case, to be sure, but it is possible. As a result, we might consider a weaker standard, one where we allow for the possibility that an agent gets lucky, but where we disallow punishing agents for misfortune. This is a policy suggesting that unjustifiably rewarding someone can be tolerated but wrongfully punishing them will not be. The idea is that at a minimum we will not hold errors about rationality against an agent, but we will tolerate if we must those made in their favor. This is weak credal internalism.

Weak Credal Internalism.

> As above, let X be a random variable representing an uncertain event, whose possible realizations are denoted by an integer from the finite set $\{1, 2, 3, ...N\}$. Suppose further we have an agent who has a credence function defined on X, $c(X)$. Suppose we observe $X = i$.
>
> An assessment regarding the quality of $c(X)$ should not produce a verdict that renders an agent less rational than they would otherwise have been had some other value $X = j$ in fact been realized.

Joyce's Coherence satisfies this constraint. If an agent has a coherent credence function, and updates by Bayesian conditioning, she will never end up with an incoherent credence function. However, the requirement to be as accurate as possible would not because, again, an agent might encounter misleading evidence that renders the agent less rational than she would have been had she observed more favorable evidence. Indeed, as (Fallis and Lewis, 2016) point out, an agent's actual accuracy can go down after conditioning, *even if* the evidence is not misleading (Fallis and Lewis, 2016). In the next section, we will show that Approximate violates WCI.

4 Weak Credal Internalism Applied to Approximate

The example to follow involves a credence function regarding a coin's bias. As a result, our agents have credences about a continuous unknown rate parameter for an observable binary process. There is nothing unusual about this setup. Indeed, it is quite typical. In Bayesian inference, binomial and multinomial data, count data, queuing, and failure times are all processes which are typically modeled in terms of finite discrete sample spaces given an unknown continuous (e.g. location or scale) parameter. It is that parameter that is of interest to Bayesians, because its posterior probability distribution forms the basis for point estimates, hypothesis tests, credible regions and, of course, predictions. Indeed, Bayesian inferences are often simply summary statistics of the unknown parameter's posterior distribution. The categorical Coherence norm extends quite naturally to this context.[4]

[4]Whether or not a credence function is discrete or continuous does not (and should not) matter. It could matter from a different perspective: we have defined coherence in terms of countable rather than finite additivity, but I am comfortable with proceeding on this assumption for now and would like to set aside that debate.

De Bona and Staffel's Approximate norm does not. Suppose we have two agents, A and B, who have opinions about the unknown bias of a certain coin, which may take any value $\theta \in [0, 1]$. These opinions are expressed by their credence functions for θ, which we will denote as $p_A(\theta)$ and $p_B(\theta)$, respectively. A and B may not be coherent, which means that $p_A(\theta)$ and $p_B(\theta)$ can fail to be probability distributions. Now, let us suppose for illustration that A and B will perform a very simple experiment. In particular, they are going to toss the coin once. The assumption that our agents will toss the coin only once is adopted for ease of exposition. It will be clear that the same problem can arise with any finite number of tosses. We will denote the result of the coin toss with the random variable X which can take the value 0 (for heads) or 1 (for tails). After observing the result they will update their beliefs by applying Bayes' Rule to get the posterior credence functions $p_A(\theta|x)$ and $p_B(\theta|x)$. That is, for each agent, their posterior credences are given by

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

where

$$p(x) = \int_0^1 p(x|\theta)p(\theta)d\theta$$

is the pseudo marginal credence for $X = x$ over all possible values of the unknown quantity θ.[5] WCI requires that our normative assessment of the quality of A and B's new credence functions should not depend on whether the coin lands on heads or tails. After all, both A and B now plan to update by Bayesian conditioning, and we may assume they will in fact do so, both in the possible world where the coin lands on heads and in the possible world where it lands on tails. Therefore, since their mental states are the same in all relevant respects in both the heads world and the tails world, our normative assessment about the rationality of their credences should not punish them, so to speak, depending on the actual outcome that is observed.

Suppose that A and B's credence functions before the experiment are given by,

$$p_A(\theta) = \frac{1}{1-\theta} \qquad\qquad p_B(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} \tag{1}$$

These credences are not coherent. It is not entirely clear how we should interpret incoherent credence functions. If we assume that they indeed encode an agent's partial beliefs, as De Bona and Staffel do, then the interpretation of agent A would be that they are very confident the coin is tails-biased whereas the interpretation of agent B would be that they are just as confident that the coin is biased but indifferent as to which direction it is biased toward. However, for any plausible measure of divergence B is better because A's credence function is a non-finite measure over the parameter space $[0, 1]$ whereas B's credence function is an

[5]One might worry that this raises thorny issues about how incoherent agents evaluate moments. Fortunately, the counter-example to follow will sidestep this issue. For a fixed value of $X = x$, the quantity $p(x)$ is a constant. For credence functions that can be normalized, it is the normalizing constant. For credence functions that cannot be normalized, the problem we identify persists for any real number assigned to this quantity.

unnormalized probability. Specifically,

$$\int_0^1 \frac{1}{1-\theta} d\theta = +\infty \qquad\qquad \int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}} d\theta = \pi \tag{2}$$

The left-hand equality obtains because the function $1/(1-x)$ does not converge on our parameter space, whereas the right-hand equality obtains because $(x(1-x))^{-1/2}$ is the kernel of the arcsin distribution whose cumulative probability is given by $1/2\pi \arcsin(\sqrt{x})$.

Suppose we assess the degree of A and B's incoherence by applying the well-known Kullback-Leibler divergence. Between $p_i(\theta)$ and $p_j(\theta)$ it is defined as,

$$KL(p_i||p_j) = \int_\Omega p_i(\theta) \log \frac{p_i(\theta)}{p_j(\theta)} d\theta \tag{3}$$

Nothing in this argument depends on KL divergence being the 'right' measure of divergence. I use this measure for illustration because it is among the most well-known measures of divergence. It corresponds to the Shannon measure of information entropy as well as the strictly proper logarithmic scoring rule for evaluating the inaccuracy of a credence function. Applying this expression to the bias of the coin and rearranging, we have,

$$KL(p_i||p_j) = \int_0^1 p_i \log p_i d\theta - \int_0^1 p_i \log p_j d\theta \tag{4}$$

$KL(p_A||p_C)$ is either undefined or ∞ for any coherent candidate C since the left-hand side of the difference is ∞ and the right-hand side is either ∞ or 0 (if C's credences are uniform on θ, $p_C = 1$). Meanwhile, $KL(p_B||p_C)$ is a finite number. For example, if C's credences are uniform then $KL(p_B||p_C) = 4.35$. Therefore, our initial assessment must be that B is doing better. A has a distribution that is beyond repair. It is not coherent, and there is no normalizing constant that would turn it into a coherent distribution. If we use simple absolute value distance (or squared distance, for that matter) then the divergence between B and the closest coherent distribution (indeed, any coherent distribution) is likewise ∞. Since $x/(1-x)$ is not bounded above on our domain, the absolute divergence would be $\infty - k$ for $k < 1$. To summarize,

Verdict according to Approximate:	
At time-0,	
A is Awful, and B is Better.	In short, B > A

Now let us consider what happens after our agents perform a simple experiment.

Case 1: The coin is tossed and it lands on heads.

If the coin lands on heads and the agents update by Bayesian conditioning, we end up with the following posterior credences,

$$p_A(\theta|X=0) = \frac{\theta}{1-\theta} \qquad\qquad p_B(\theta|X=0) = \sqrt{\frac{\theta}{1-\theta}} \tag{5}$$

In this case, A's credences are now given by the odds for heads whereas B's credences correspond to their square root. As a result,

$$\int_0^1 \frac{\theta}{1-\theta} d\theta = +\infty \qquad\qquad \int_0^1 \sqrt{\frac{\theta}{1-\theta}} d\theta = \pi/2 \tag{6}$$

Approximate gives the verdict that A remains awful and B remains better. This is consistent with the verdict rendered before the coin was tossed. So far so good. Now let us consider what happens if the coin lands on tails.

Case 2: The coin is tossed and it lands on tails.

If the coin lands on tails and the agents update by Bayesian conditioning, we end up with the following posterior credences,

$$p_A(\theta|X=1) = 1 \qquad\qquad p_B(\theta|X=1) = \sqrt{\frac{1-\theta}{\theta}} \tag{7}$$

In this case, A's credences are uniform. The density is a straight line at $y = 1$ for every value of θ between 0 and 1. Meanwhile, B's credences correspond to the odds for tails – i.e., they are reciprocal to B's credences when the coin lands on heads. As a result,

$$\int_0^1 1 d\theta = 1 \qquad\qquad \int_0^1 \sqrt{\frac{1-\theta}{\theta}} d\theta = \pi/2 \tag{8}$$

A has become perfectly coherent, whereas B remains as incoherent as they would be if the coin landed on heads. The verdict according to Approximate is now the reverse: A is doing better and B is doing worse. A has gotten lucky and is now coherent. This illustrates that not even the ordinary Coherence norm satisfies strong credal internalism. Further, B is now doing worse. Since Approximate is a standard for making relative judgments, A is in effect punished for being unlucky. The following table summarizes the possible results after updating on a single coin toss.

Verdicts according to Approximate:	
At time-1,	
1. If the coin lands on heads, then B is Better, and A is Awful. (B > A)	2. If the coin lands on tails, then A is Perfect, and B is Worse. (A > B)

Therefore, the verdict given according to Approximate, as between whether A or B is doing better, depends on whether the coin lands on heads or tails, a feature of the problem that is external to both agents and over which they have no control. Therefore, an agent can be unlucky to be irrational, in violation of WCI.

This problem is guaranteed to occur for any finite number of n tosses if we observe a sequence of n heads or tails. In particular, at time-n if we observe a sequence of n heads, then Approximate will give the verdict that B is better and A is awful, whereas if we observe a sequence of n tails, Approximate will give the verdict that A is better (though not necessarily perfect) and B is worse.

5 Incoherence and the Theory of Second Best

I have assumed that our non-ideal imperfectly coherent agents perfectly update their beliefs by Bayesian conditionalization. The justification for this is that I did not want to make further assumptions about how irrational our agents are. In other words, I started with an ideal agent, perturbed their coherence, and proceeded without further deviations from the ideal state. However, one might object that this is unfair. Imperfectly coherent agents should also be thought of as imperfect conditionalizers. This line of reasoning suggests that when A sees the coin land on heads, she should nonetheless put some weight on tails for good measure. This would ensure that her posterior converges. I do not find this especially compelling. The example given with a single coin toss is simple enough that it does not seem to be especially burdensome to apply Bayes' Rule. Indeed, it seems quite odd to put more weight on tails after seeing the coin land on heads. But suppose we do indeed go with this response. It is not clear that this helps to salvage De Bona and Staffel's approach. Rather, it seems to reinforce the general argument suggested in this paper: when one is imperfectly rational, it does not necessarily pay to approximate rationality. The example we have given suggests this lesson for coherence. The considerations in this paragraph suggest something similar regarding conditionalization: when an agent is imperfectly coherent, she might want to be an imperfect updater as well. The argument presented in this paper is in some respects similar to Lipsey and Lancaster (1956) General Theory of the Second Best for approximating (economically ideal) Pareto efficient outcomes.

6 Concluding Remarks

According to the Bayesian version of the likelihood principle, the only thing that matters for drawing inferences is the posterior distribution. Therefore, whether or not the prior is approximately coherent or, indeed, whether it converges is not too important. What really matters is a categorical question: can the posterior be normalized? If it can be, then Bayesian inference proceeds as usual. If it cannot be, then one cannot make Bayesian inferences because the posterior distribution does not exist. Therefore, while it is tempting to take a graded approach to evaluating non-ideal Bayesian agents, perhaps the question that really matters is categorical. A Bayesian agent's behavior is either rational, or not. If it is not, it can be misleading to rank degrees of irrationality because any such ranking can depend on misfortune.

NOTE: One more objection to consider. Agent A's bad luck is agent B's good luck. So what my example shows is that A is 'punished' to the extent 'B' is rewarded, since the comparisons being made are relative. Perhaps De Bona an Staffel should resist the use to which I have put their approach – i.e., one of ranking a group of people (or, their credence functions) relative to each other, and expecting this ranking to remain consistent over time. They could point out that if the comparison is from old A to new A, then new A remains as incoherent in both a heads world and a tails world. In absolute terms, new A under heads is just as incoherent as new A under tails. So from the perspective of 'how perfect am I?', in isolation, the answer remains the same. From the perspective of 'who's more perfect, me or you?', the answer can depend on what seem to be arbitrary features that should perhaps be irrelevant to our rationality. I do think an especially attractive feature of the notion of

graded coherence is that it promises to enable us to make these relative judgments. That is, to take a group of imperfect agents, and say something about how they are doing from the perspective of ideal rationality – things like, A is doing better than B, is doing better than C, etc. If we want to preserve this feature, then we want to be able to make relative statements, as I have done, and (this is another point De Bona and Staffel might want to push back on) it does seem that these statements should be stable under updating, so to speak. In other words, if A is better than B now, at time 1, and if between time 1 and time 2 they do nothing wrong, so to speak, then A should be better than B at time 2. If the order flips, something in our normative assessment has gone wrong. A did not do anything that evidence further irrationality, B did not do anything that evidences improved rationality, yet our ranking has changed. This is the problem I have tried to identify. Perhaps there is a better way to characterize this concern than I have done in terms of so-called 'weak credal internalism'.

Summary of potential replies by DeBona and Staffel:

– Deny that imperfect agents should update by perfect Bayesian conditioning.
– Push back on the distinction between strong and weak credal internalism. Since ordinary coherence violates the strong version, then ordinary and graded coherence are in the same boat.
– Deny that the relative graded ranking needs to be stable under updating.

References

Babic, B. (2018). A Theory of Epistemic Risk. *Philosophy of Science 1*(2), 2–3.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association 57*(298), 269–306.

Blackwell, K. and D. Drucker (2018). When Propriety is Improper. *Philosophical Studies Forthcoming*.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*(1), 1–3.

Cox, D. (1958). Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics 29*(2), 357–372.

De Bona, G. and J. Staffel (2017). Graded Incoherence for Accuracy-Firsters. *Philosophy of Science 84*(2), 189–213.

De Bona, G. and J. Staffel (2018). Why be (Approximately) Coherent? *Analysis (*Forthcoming).

Fallis, D. and P. Lewis (2016). The Brier Rule Is not a Good Measure of Epistemic Utility (and Other Useful Facts about Epistemic Betterness). *Australasian Journal of Philosophy 94*(3), 576–590.

Goldman, A. I. (2002). *Pathways to Knowledge: Private and Public*. Oxford: Oxford University Press.

Greaves, H. and D. Wallace (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind 115*(459), 607–632.

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science 65*, 575–603.

Joyce, J. M. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In F. Huber and C. Shmidt-Petri (Eds.), *Degrees of Belief*, pp. 263–300. Springer.

Lipsey, R. and K. Lancaster (1956). The General Theory of Second Best. *The Review of Economic Studies 24*(1), 11–32.

Pettigrew, R. (2017). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Staffel, J. (2015). Measuring the Overall Incoherence of Credence Functions. *Synthese 192*(5), 1467–1493.

Wedgwood, R. (2002). Internalism Explained. *Philosophy and Phenomenological Research 65*(2), 349–369.