Approximate Coherentism and Luck*†

Boris Babic

forthcoming in *Philosophy of Science*

Abstract

Approximate coherentism suggests that imperfectly rational agents should hold approximately coherent credences. This norm is intended as a generalization of ordinary coherence. I argue that it may be unable to play this role by considering its application under learning experiences. While it is unclear how imperfect agents should revise their beliefs, I suggest a plausible route is through Bayesian updating. However, Bayesian updating can take an incoherent agent from relatively more coherent credences to relatively less coherent credences, depending on the data observed. Thus, comparative rationality judgments among incoherent agents are unduly sensitive to luck.

---

*Department of Decision Sciences, INSEAD; e-mail: boris.babic@insead.edu.

†I would like to thank Glauber De Bona, Kenny Easwaran, Simon Huttegger, Calum McNamara and Julia Staffel for very helpful feedback. I have also benefited from conversations with the participants of the 34th Annual Conference on Chance and Probability in Science in Boulder, CO. Special thanks are also due to the anonymous reviewers from Philosophy of Science for supporting this paper and for their valuable comments.

1 Introduction

Glauber De Bona and Julia Staffel argue that the credences of non-ideal Bayesians 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 credence function (Staffel, 2015; De Bona and Staffel, 2017, 2018). I'll call this norm approximate coherence (and define it appropriately in Section 3). It says that from a (third person) evaluative perspective, if B's credences are closer to coherence than A's credences, then B's credences are to that extent more rational than A's credences. By third person evaluative perspective I simply mean, to paraphrase Pollock (1987), norms we use to evaluate the rationality of others' beliefs. This may be contrasted with first person norms, which tend to be described as action-guiding.

De Bona and Staffel show that for every incoherent credence function there exists an accuracy dominating less incoherent one, under fairly general conditions, whereas Joyce showed that for every incoherent credence function there exists an accuracy dominating coherent one (under similar conditions) (Joyce, 1998, 2009). Accordingly, we can think about De Bona and Staffel's notion of approximate coherence as extending Joyce's ordinary coherence norm by giving a graded judgment about rationality, rather than a categorical judgment. I will argue that such graded judgments about rationality can be misleading when we evaluate credences under updating.

Both coherence and approximate coherence should be distinguished from what some call final ends in epistemology, like truth or accuracy, which say that as between two beliefs or credence functions, the one that is in fact true or more accurate is better. To

understand why consider, in ordinary decision-making, a norm that requires an agent to maximize *actual* utility. Actual utility is analogous in many respects to actual accuracy in epistemology. It is possible to be unlucky, so to speak, because which option actually maximizes utility depends on which world turns out to be true. It may be that at time 1, my decision procedure for choosing an act is better; and at time 2 the same decision procedure, given the same acts and states, is worse, because an improbable state in which that act does not maximize utility occurs.

Now consider the norm of maximizing *expected* utility. This norm is analogous in some respects to the role that coherence and approximate coherence play in epistemology. 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 render an adverse judgment about their rationality due to factors outside their control. More precisely: we ordinarily do not think – at least as (epistemic) decision theorists – that such factors should adversely affect our assessment of the relative quality of their choices from an evaluative perspective. Similarly, we typically suppose (as a convenient fiction, perhaps) that an agent can choose their credences, and we evaluate their internal consistency or rationality – an assessment that should likewise not be susceptible to luck or misfortune. In this sense, coherence and approximate coherence are different from truth or accuracy norms.

I attempt to capture, for our purposes here, this notion of immunity to luck or misfortune, in the context of updating credences, through a principle I call comparative consistency. I then argue that while coherence is comparatively consistent, approximate coherence is not. Its verdicts regarding relative rationality between two agents (or,

rather, their credence functions) can depend on factors outside the agents' control. In particular, it is possible to have the following: 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 about the coin's bias. If the coin lands on tails, A becomes less coherent than B, whereas if it lands on heads, B remains less coherent. Therefore, it is possible to observe an 'unfortunate' sequence of data, so to speak, which adversely affects our evaluative assessment of the relative quality of an agent's doxastic state. A can go from doing relatively better, to doing relatively worse, solely due to the outcome of the coin toss.

The paper proceeds as follows. In Section 2, I explore the normative status of accuracy and coherence/ approximate coherence. In Section 3, I define both norms in the context of a simple example – credences about a coin's bias. In Section 4, I motivate and explain the notion of comparative consistency. I then show that while ordinary coherence is comparatively consistent, approximate coherence is not. In Section 5, I consider several potential objections. By way of conclusion, I suggest that what really matters under updating in ordinary Bayesian epistemology is the binary question of whether the posterior distribution can be identified. And since approximate coherentism likely requires us to give up updating along ordinary Bayesian lines, I suggest an important line of further research – namely, the nature of belief updating norms for imperfectly rational agents.

2 Background: Accuracy and Coherence

I develop the argument to follow within the general framework of Bayesian epistemology or epistemic utility theory (Joyce, 1998, 2009). In this framework, minimizing expected inaccuracy plays a similar role that maximizing expected utility plays in ordinary decision theory. The main difference is that we replace the utility function with an appropriate measure of accuracy, or scoring rule. The relevant details of the scoring rule framework will be introduced below – fortunately, most can be omitted, because the conditions required for De Bona and Staffel (2017)'s results are similar to those required for Joyce (2009)'s results.

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. However, there remains a question about whether approximating coherence can be defended on non-pragmatic grounds. This is a natural question to ask with respect to norms in Bayesian epistemology. For example, Joyce (1998, 2009) gives a non-pragmatic defense of probabilism whereas Greaves and Wallace (2006) give non-pragmatic defenses of updating by Bayesian conditioning. Fittingly, DeBona and Staffel also offer a non-pragmatic defense for approximating coherence – namely, that approximating coherence improves one's accuracy outcomes. This is the issue I take up in this paper.

Joyce (1998) introduces the guiding ideal for epistemic utility theory, 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 better. Accuracy is ordinarily assessed using a scoring rule, which maps a pair of values – a credal 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. This score, known as the Brier score, maps credences to $[0, 1]$ 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, 2016). 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, when we evaluate the rationality of an agent's credences under conditions of uncertainty, we do not use graded accuracy on its own – we recognize it is possible for an agent to formulate her credences diligently but end up inaccurate due to misfortune.

For example, suppose Alice is making a forecast about the weather tomorrow in a rainy city during a rainy season – Seattle in November, say. She estimates the probability of rain to be 0.7. This is consistent with her evidence, including professional forecasts and the November benchmark for Seattle. 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 possible (albeit relatively improbable) worlds where it does not rain in Seattle tomorrow, Alice will be

substantially inaccurate. Ordinarily, we do not hold such misfortune against an agent's epistemic rationality.

Rather, we use graded accuracy, together with an appropriate decision rule, in order to identify certain evaluative norms for the assessment of an agent's credal state. For instance, Joyce (1998, 2009) defends coherence (the relevant norm) by showing that every incoherent credence function is dominated (the decision rule) by some coherent credence function with respect to the underlying measure of accuracy (the feature of primary epistemic value). If an agent is incoherent we say they are to that extent epistemically irrational. Likewise, Greaves and Wallace (2006) defend Bayesian updating (the relevant norm) by showing that under relatively general conditions, updating by Bayes' Rule maximizes the expected value (the decision rule) of the underlying measure of accuracy (the feature of primary epistemic value). Similarly, De Bona and Staffel defend approximate coherence (the relevant norm) by showing that every incoherent credence function is dominated by some less incoherent credence function (the decision rule) with respect to the underlying measure of accuracy (the feature of primary epistemic value).

3 Coherence and Approximate Coherence

In this section, I make the relevant notions of coherence and approximate coherence more precise and explain that approximate coherence, as De Bona and Staffel characterize it, is intended to be a graded generalization of its categorical relative – i.e., a notion that can help us make comparative rationality judgments on the basis of

accuracy considerations. I consider cases where we are formulating a credence about a single, continuous parameter – the unknown bias θ of a certain coin. One can also think about the stylized objective chance of rain tomorrow, the unknown mean of a normally distributed random variable, etc. I define the relevant concepts accordingly.

> **Coherence Norm.** Let Ω denote our parameter space containing all possible values of the true unknown quantity $\theta \in \mathbb{R}$. A credence function of θ, $p(\theta)$, is coherent if (1) $p(\theta) \geq 0$ and (2) $\int_{\Omega} p(\theta) d\theta = 1$.[1]
>
> A coherent credence function is (from an evaluative perspective) better than (more rational than) an incoherent one.

Coherence, as noted, is valuable because it guarantees improvements in accuracy. De Bona and Staffel propose the norm of approximate coherence as a suitable generalization to ordinary coherence which allows us to make comparative rationality judgments about Bayesian agents. Since it is unrealistic to expect ordinary agents to be coherent, we can evaluate them in terms of how closely they approximate that 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 Coherence Norm.** Let Ω denote our parameter space containing all possible values of the true unknown quantity $\theta \in \mathbb{R}$. Let $f(\theta)$

[1]The two conditions above commit us to countable additivity 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.

be a function that violates (1) or (2), from 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 (from an evaluative perspective) to that extent better than (more rational than) g.

This says that we should evaluate the degree of incoherence by looking at the divergence of the incoherent credence function from its closest coherent credence function, and that as between two incoherent credence functions we should judge the less incoherent one to be relatively more rational or, as De Bona and Staffel (2017) put it, epistemically better.

Note that for the purpose of this project, we will not look at violations of additivity, which add a further layer of complication. Similarly, De Bona and Staffel illustrate incoherence in terms of violations of normalization. For instance, in Section 3 of De Bona and Staffel (2017), the working example is of two agents, Clara and Jane, whose credences about a tripartite partition are incoherent in the sense of adding up to more than 1, and less than 1, respectively. For agents who are incoherent in the sense of violating additivity, **?** argue for a generalization of Bayes' Rule for updating such beliefs using the notion of Choquet expected utility – an approach to computing expectations with respect to non-additive functions.

De Bona and Staffel defend approximate coherence 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 measure of divergence (Proposition 2). In general, a divergence measure is suitable if its associated scoring rule satisfies the conditions required by Joyce (2009) (continuity, truth-directedness, and strict propriety) and it is additive.[2] A scoring rule is additive if we evaluate the inaccuracy of a credal assignment by adding up (not necessarily in equal weight) the inaccuracies of the credences assigned to every possible outcome in the relevant partition or algebra of events.

4 Coherence and Comparative Consistency

To begin, I want to highlight a basic insight that I think is generally accepted in epistemic utility theory, insofar as the field is itself a type of decision theory, where the rationality of epistemic acts is evaluated in terms of their expectations under conditions of uncertainty. The insight, to paraphrase Wedgwood (2002), is that the rationality, from an evaluative perspective, of a credence function should generally not depend on facts about the world outside the agent's control. The clearest illustration of this guideline is

[2]A scoring rule $s(I_v, p)$ that measure the inaccuracy of credence p given that the true state is I_v is a measure of divergence $D(q, p)$ where q is a credence function that puts all probability mass/density on the true outcome – i.e., q is the omniscient credence function. That is, a scoring rule is a special kind of divergence; thus, when we say 'a measure of divergence and its associated scoring rule', what is meant is a measure of divergence between two credence functions and that same measure of divergence between a credence function and the omniscient credence function. For a recent discussion of the structure of scoring rules, and their relationship to divergence measures see Babic (2019).

the focus on expected (as opposed to actual) epistemic utility in evaluating the rationality of a credence function.[3]

4.1 A Metanormative Principle for Epistemic Norms

The above general insight is underspecified for our purposes, however. To apply it here, we will articulate a metanormative principle that captures its spirit in the context of evaluating the rationality of credences as the agent obtains new evidence. I am interested in whether our relative assessment about a credence function is appropriately consistent, either over time or under equivalent representations of the agent's doxastic state. In particular, consider the following: if we render a verdict about the rationality of A's credences, relative to B's credences, at time 1, and between time 1 and time 2 the only relevant aspect of the situation that changes is that a new piece of evidence is obtained, and A and B both update their beliefs diligently to account for the new evidence, then this much seems clear: at time 2, it does not seem appropriate for our assessment of the relative rationality of A and B's credences to shift. In other words, since A and B have both diligently updated their beliefs, the verdict rendered at time 2 should be consistent with the verdict rendered at time 1. For lack of a better term, I will call this principle the requirement of comparative consistency.

Comparative Consistency.

Suppose A and B have credences about an unknown quantity θ. Evidence

[3]Perhaps one can put this in terms of the distinction between internalist and externalist norms. I have tried to avoid this distinction as it may distract from what is a fairly straightforward point about norms in decision-theoretic epistemology.

about θ comes from experiments whose outcomes are denoted by X. If A's priors about θ are less incoherent than B's then, for any value $X = x$ that might be observed, A's posteriors about θ should not be more incoherent than B's.

In short, if A's priors are less incoherent than B's then, for any observation, A's posteriors should not be more incoherent than B's.[4] In some respects, comparative consistency is like a stability condition on norms of epistemic rationality. It says that if you now believe A's credences to be more rational than B's credences, you should not change that assessment unless you have an appropriate reason to do so – such as a new piece of evidence suggesting that one of them acted irrationally (or less rationally) in some way. In other words, the verdict about comparative rationality should not shift arbitrarily. While this is only a bare bones sketch of the notion, the goal of this project is not to fully articulate and argue for a metanormative theory. Rather, I hope the principle is sufficiently plausible for now.

4.2 Updating Incoherent Credences

I will show through a simple example that approximate coherence is not comparatively consistent. The updating situation is ordinary. There are no issues of uncertain evidence that would trigger Jeffreys' Rule or some alternative to it; or of misleading evidence, etc.

[4]In Section 5, I take up potential objections to the Bayesian updating assumption. I also explain, in Section 4.3, that under some fairly modest assumptions, the only updating strategy for incoherent agents that satisfies reflection and the martingale principle is Bayesian.

Further, our agents will indeed diligently update. The notion of comparative consistency prohibits a verdict which deems one of them more rational before the update and less so following the update.

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, denoted by $p_A(\theta)$ 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. However, in the case we will consider, these functions remain additive. This is similar to the kind of incoherence that De Bona and Staffel use to describe graded incoherence (violations of normalization). Suppose for illustration that A and B will perform a simple experiment. In particular, they are going to toss the coin once. The assumption that our agents will toss the coin once is adopted for ease of exposition. It will be clear that the problem can arise for 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 by Bayes' Rule to get the posterior credence functions $p_A(\theta|x)$ and $p_B(\theta|x)$. For ordinary coherent priors, updating would be straightforwardly accomplished through the familiar expression of Bayes' Rule, as follows:

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{m(x)} \tag{1}$$

where

$$m(x) = \int_0^1 f(x|\theta)p(\theta)d\theta \tag{2}$$

is the marginal credence function for $X = x$ over all possible values of the unknown

quantity θ and $f(x|\theta)$ is the data generating distribution.

For incoherent agents, however, $m(x)$ may or may not be a valid probability function. In ordinary Bayesian inference, this is usually not an issue, because the function m is not a function of the unknown quantity of interest, θ, but is instead a function of the data, x. Accordingly, we would proceed as usual, treating this function as a pseudo marginal distribution, focusing on the components that are functions of θ, and taking the posterior to be proportional to the prior times likelihood:

$$p(\theta|x) \propto f(x|\theta)p(\theta). \tag{3}$$

This is indeed what I will do.

But whereas for ordinary credences we might, if desired, normalize the posterior so as to rescale the resulting distribution into $[0, 1]$, for our purposes here doing so is not innocuous. Our main subject of interest is the incoherence in the agent's credence function. Accordingly, I do not want to automatically rescale because I am taking it for granted, as De Bona and Staffel do, that there is something noteworthy about the agent's imperfect doxastic state, and I want to preserve that aspect of the situation under updating. Accordingly, I want to take the prior distribution to a posterior distribution as faithfully as possible without arbitrarily wiping out the initial incoherence and thereby changing the agent's subjective representation of uncertainty. Since the agent's beliefs remain additive, the Bayesian posterior is a faithful posterior that maintains prior proportions of probabilities while reflecting the new evidence. Further, inferences on θ and predictions about X remain the same whether we normalize or not. After updating,

I want to compare relative incoherence under different possible observations $X = x$ to see if the ranking is stable. That will be the strategy in the next section.

Ultimately, however, we will see what it takes to normalize each credence function. Rather than hiding the ball, so to speak, we will do things one step at a time. Recall that for a fixed value of $X = x$, the quantity $m(x)$ is a constant. The core of the problem will be that for credence functions that can be normalized, it is a finite normalizing constant whereas for credence functions that cannot be normalized, the problem we identify persists for any real number assigned to this quantity.

Comparative consistency 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. Since their mental states are the same in all relevant respects in both the heads world and the tails world, and their updating behavior is the same, and neither has done anything to suggest a change in our verdict regarding their relative rationality, our post-experiment normative assessment about the rationality of their credences should not depend on the observed outcome. It should be consistent with our pre-experiment verdict.

4.3 Reflection and Expectation

The updating strategy in Equation 3 is far from arbitrary. Rather, it is a more general expression of the ordinary statement of Bayes' Rule as given in Equation 1. Huttegger (2014, 2017) argues that in order for an update rule to count as a reasonable response to a learning experience it must satisfy the principle of reflection or, more specifically, the

martingale principle. And drawing on Zabell (1982), he further shows that, given some fairly modest assumptions (called exchangeability, regularity, and sufficientness), the general expression in Equation 3 is the only one that does so for the kind of case we are considering (single parameter inference for an unknown proportion) (See Huttegger, 2017, Chs. 5-6). This remains true whether the agent is coherent or not, provided their credences remain additive.

In our context, the martingale principle requires that the conditional expected value of the probability of heads/tails on the $n+1$th toss, given the result on the nth toss, is equal to the probability of heads/tails on the nth toss. We can show that the update rule in Equation 3 indeed satisfies the martingale principle, as follows. To simplify, I will express the point using the first and second toss.

First, we need to define the predictive distribution of an as yet unobserved outcome of X, which we will denote as $\widetilde{X}$, given the outcomes of X that have so far been observed, which we will denote with the lowercase x. This distribution takes the following form:

$$\Pr(\widetilde{X} = 1|x) = \int_0^1 \Pr(\widetilde{X} = 1|x, \theta)p(\theta|x)d\theta = \int_0^1 \theta p(\theta|x)d\theta = \mathrm{E}[\theta|x] \tag{4}$$

That is, the probability that $\widetilde{X} = 1$ (the coin lands on tails) is given by the posterior mean of θ after conditioning on the tosses that have been observed. Before any tosses have been observed, we would simply use the prior mean for θ, given by $\mathrm{E}[\theta]$. This would be the predicted probability that $X_1 = 1$. Under the above assumptions, both the prior and the posterior mean can be computed from the kernel of the prior/posterior

distribution – i.e., we do not need to determine the normalizing constant. Now we want to show that the expected value of the probability that the second toss lands on tails, given the result of the first toss, is equal to the probability that the first toss lands on tails. That is,

$$\mathrm{E}[\Pr(X_2 = 1)|X_1 = 1] = \Pr(X_1 = 1) \tag{5}$$

We know from Equation 4 that the quantity $\Pr(X_2 = 1)|X_1 = 1]$ is given by the posterior mean, that is, $\mathrm{E}[\theta|X_1 = 1]$. And by the law of iterated expectation, we know that $\mathrm{E}\left[\mathrm{E}\left[\theta|X_1 = 1\right]\right] = \mathrm{E}[\theta]$. The same point holds for any sequence of outcomes $X_1, ...X_n$.

Thus, both expressions of Bayes' Rule (Equation 1 and Equation 3) remain faithful to the martingale principle and more generally reflection. This has to be true because they are proportional to each other and as a result inferences on θ and associated predictions about X remain unchanged. The constant, $c = m(x)$, rescales the distribution of θ, but the scale is irrelevant to inference. What matters is the relative density assigned to different regions of the parameter space. But, if we move to some other non-Bayesian update rule, it is not clear how we could satisfy the martingale principle. I discuss the Bayesian updating assumption further in Section 5.2, and the normalization point in Section 5.3.

4.4 A Counterexample

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

$$p_A(\theta) = \frac{1}{1-\theta} \qquad\qquad p_B(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} \tag{6}$$

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 suggest, 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's credence function is more rational (better) because A's credence function is a non-finite measure over the parameter space $[0, 1]$ whereas B's credence function is an unnormalized probability. Specifically,

$$\int_0^1 \frac{1}{1-\theta} d\theta = +\infty \qquad\qquad \int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}} d\theta = \pi \tag{7}$$

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,

$$KL(p_i||p_j) = \int_\Omega p_i(\theta) \log \frac{p_i(\theta)}{p_j(\theta)} d\theta \tag{8}$$

Nothing in this argument depends on KL divergence being the 'right' measure of divergence. I use this measure for illustration, as it is well-known, corresponds to the

Shannon measure of information entropy (Shannon, 1948a,b), and to the strictly proper logarithmic scoring rule (Gneiting and Raftery, 2007). Applying this expression to the bias of the coin and rearranging, we have,

$$KL(p_i||p_j) = \int_0^1 p_i \log p_i d\theta - \int_0^1 p_i \log p_j d\theta \tag{9}$$

For any coherent candidate credence function p_C, $KL(p_A||p_C)$ is either undefined or ∞, 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's credences are better. A's credences are not coherent and they cannot be made coherent – there is no finite normalizing constant that would turn A's credence function into a valid probability distribution. If instead we use simple absolute value distance (or squared distance, for that matter) then the divergence between A and the closest coherent distribution (indeed, any coherent distribution) is likewise ∞ since $x/(1-x)$ is not bounded above on $[0, 1]$. To summarize,

Verdict, according to approximate coherence, of the rationality of A and B's credences:	
At time-1,	
A is awful, and B is better.	In short, $B > A$

Consider, next, what happens after our agents perform a simple experiment. The

distribution of the coin is given by $\theta^x(1-\theta)^{1-x}$. Accordingly, when the coin lands on tails, the posterior density is proportional to the prior density multiplied by θ, and when the coin lands on heads, the posterior density is proportional to the prior density multiplied by $1-\theta$.

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

If the coin lands on tails, we obtain the following posterior credences,

$$p_A(\theta|X=1) = \frac{\theta}{1-\theta} \qquad\qquad p_B(\theta|X=1) = \sqrt{\frac{\theta}{1-\theta}} \tag{10}$$

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

$$\int_0^1 \frac{\theta}{1-\theta} d\theta = +\infty \qquad\qquad \int_0^1 \sqrt{\frac{\theta}{1-\theta}} d\theta = \pi/2 \tag{11}$$

Approximate coherence gives the verdict that A remains awful and B remains better. To use the shorthand from above: $B > A$. This is consistent with the verdict rendered before the coin was tossed. So far so good – approximate coherence is comparatively consistent. Now consider what happens if the coin lands on heads.

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

If the coin lands on heads, we obtain the following posterior credences,

$$p_A(\theta|X=0) = 1 \qquad\qquad p_B(\theta|X=0) = \sqrt{\frac{1-\theta}{\theta}} \tag{12}$$

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,

$$\int_0^1 1 d\theta = 1 \qquad\qquad \int_0^1 \sqrt{\frac{1-\theta}{\theta}} d\theta = \pi/2 \tag{13}$$

A has become perfectly coherent, whereas B remains as incoherent as they would be if the coin landed on tails. The verdict according to approximate coherence is now the reverse: A's credences are better and B's credences are worse. Approximate coherence is no longer comparatively consistent. Insofar as approximate coherence is a standard for making comparative rationality judgments between agents – or between their credence functions – then B is in effect punished for being unlucky. The following box summarizes the possible results after updating on a single coin toss.

Verdicts, according to approximate coherence, of the rationality of A and B's credences:

At time-2,

If the coin lands on tails, then B is better, and A is awful. (B > A)	If the coin lands on heads, then A is perfect, and B is worse. (A > B)

The verdict according to approximate coherence, as between whether A has more rational credences than B, 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.

In short, we have the following: if credences that sum to 1 are less incoherent than those that do not, and posteriors that can be normalized by a (finite) constant are less incoherent than those that cannot be, then approximate coherence violates comparative consistency. This problem is guaranteed to occur for any finite number n of tosses if we observe a sequence of n heads or tails. In particular, if we observe a sequence of n tails, then approximate coherence will give the verdict that B is better and A is awful, whereas if we observe a sequence of n heads, approximate coherence will give the verdict that A is better (though not necessarily perfect) and B is worse. So while the problem is increasingly improbable as n increases, it is a problem nonetheless and can occur for any finite n.

5 Objections

In this section, I try to preempt certain potential misconceptions and consider several more general objections.

5.1 Misleading Evidence

First, it's worth emphasizing that the issue is not about misleading evidence – i.e., a situation where the agent receives some evidence about the outcome of the coin toss which may or may not correspond to what actually happened. The outcomes of the toss in either case are by hypothesis accurately observed and recorded. Rather, the concern is that if a credence function about the bias of a coin is said to be relatively more rational/irrational conditional on heads, it should be equally rational/irrational conditional on tails. There is no relevant distinction to be made between the heads state

and the tails state that would motivate a shift in the evaluative verdict regarding the agents' comparative rationality.

5.2 Bayesian Updating

Second, and more importantly, one might question whether incoherent agents should update along Bayesian lines. The updating strategy adopted here is a standard application of Bayes' Rule for single parameter inference in the case of an unknown proportion with an improper prior (e.g., Lindley and Phillips, 1976). I presuppose that the agents' update should be a reasonable response to a learning experience and thus the update rule should satisfy the martingale principle (Huttegger, 2017). It is possible, of course, to insist that for subideally rational agents, everything goes out the window – they cannot update along Bayesian lines and more generally they violate the martingale principle and reflection. One might defend this on the ground that when you deviate from rationality in one respect, it is overall better to deviate from rationality in other respects as well, in the spirit of the theory of the second best (Lipsey and Lancaster, 1956). But the very idea of approximate coherence suggests that when one is not ideal, it is better to be closer to the ideal than to be further away from it.

Further, the prior I have used is one with a distinguished history in Bayesian statistics – it is the Jeffreys' Prior (for person B) and its transformation (for person A) (Jeffreys, 1946). Both are closely related to the uninformative (and incoherent) Haldane prior (Haldane, 1932). It is fair to say, then, that this project starts from the observation that incoherent priors are an important part of the Bayesian paradigm. Indeed, in many types of problems, both invariant (in the sense of Jeffreys) and uninformative priors are

often incoherent. For example, a conventional uninformative prior for the mean (μ) and variance (σ^2) of a normal likelihood is a prior that is uniform on $(\mu, \log \sigma)$ (Gelman et al., 2013, Eq. 3.2). This prior is incoherent.

However, one might insist that an incoherent agent ought not update before making the prior coherent. While this might seem like a reasonable request, it would arbitrarily restructure the agent's representation of uncertainty in the absence of new evidence and significantly limit the scope of Bayesian inference.

In some cases, the request might be feasible. In particular, when the prior integrates to a finite constant, the position would be that we must normalize before updating. Even here, the demand is not innocuous. It would create substantial impediments to inference. Often, we normalize after updating in order to avoid computing the marginal distribution of the data (in our case x), which is not a function of the parameter of interest (in our case θ). This request would make inference less tractable in difficult problems. But suppose we grant that this is what rationality requires of imperfect agents, so to speak. There is a further problem still.

Sometimes, the prior is incoherent in such a way that its integral does not converge – i.e., is improper. For example, the prior I used for person A. If we demand that this prior must be made coherent, the request is impossible to meet. As a result, if we took this position, it would commit us to prohibiting the use of improper priors, such as the Jeffreys' and Haldane priors for inference involving proportions. Suppose we grant this too. There is yet a further problem, which is that proper priors on one space are often improper on simple transformations of that space. So it is not clear how we should

understand such a request – i.e., the request that improper priors must not be used in Bayesian inference.

For example, instead of reasoning about our problem using a prior for the unknown parameter θ, it is often useful in Bayesian inference to reason about such a problem using a prior for the natural parameter of the exponential family representation of the data-generating process. In our case, that would be the log odds for θ, $\varphi = \log \frac{\theta}{1-\theta}$, because $\theta^x(1-\theta)^{1-x} = \exp[\log(\theta/(1-\theta))x - \log(1-\theta)]$ which is of the form $\exp[\varphi x - A(\varphi)]$ for $A = \log(1+e^{\varphi})$. In the absence of information about φ, a natural starting point would be a prior that is uniform for all its possible values. However, using the change of variable formula, a uniform prior for φ would lead to a prior for θ that is proportional to $1/(\theta(1-\theta))$ which is not integrable – i.e., is improper. Yet the two distributions should be equivalent in terms of corresponding to the same representation of uncertainty. In other words, if we insist that no updating can be done until the prior is made proper, then making the prior proper on the transformed space will often imply an improper prior on the original space. Thus, whether or not one is permitted to apply Bayes' Rule will then be sensitive to the parameterization of the problem. Yet it is the same underlying inference problem. Consistent with these remarks, Robert justifies the use of improper priors as follows: "the inclusion of improper distributions in the Bayesian paradigm allows for a closure of the inferential scope (figuratively as well as topologically)" (Robert, 2007, pg. 28).

Perhaps it will be helpful to explain by analogy the strategy adopted in this paper. Suppose we have an imperfectly rational agent, not in the sense we are currently considering (imperfect coherence) but in the sense of imperfect precision. When an agent

is imprecise, we can identify their upper and lower probabilities for an event and thus compute their interval estimate for the event. Seidenfeld and Wasserman (1993) describe a phenomenon where an agent starts with an interval estimate for an event, obtains some evidence, updates the initial set of priors by Bayesian conditioning, and ends up with a posterior interval that strictly contains the prior interval. This is the well-known phenomenon of probabilistic dilation. By identifying this phenomenon, Seidenfeld and Wasserman (1993), among others, suggested there is a problem either with imprecise probability models, or with the assumption that one should update them by Bayes' Rule because, intuitively, one's interval should become narrower – or, at least, not wider – after undergoing a learning experience. By framing the problem in terms of this potential incompatibility – made explicit by the occurrence of dilation – they encourage us to think carefully about how the imprecise framework fits with the standard Bayesian updating picture. Since then, there have been many proposed solutions of the apparent conflict. My goal in this paper is to highlight something similar for approximate coherence – i.e., to identify a potential conflict between the approximate coherence framework and standard Bayesian updating, but whereas Seidenfeld and Wasserman (1993)'s underlying normative principle for interval estimates is that they should not dilate under updating, my underlying normative principle is that comparative judgments about rationality should be consistent under updating.

5.3 To Normalize or Not?

Third, one might question why I don't automatically normalize the agent's credences as part of the update. Instead, I break up the analysis as follows. First, I derive the

posterior, taken to be as prior times likelihood, then I show what it would take to normalize it. For example, in Equation 10, p_A is A's non-normalized posterior and P_B is B's non-normalized posterior. In the next step, Equation 11, I show just how incoherent they are by highlighting what it would take to normalize them – p_B sums to $\pi/2$, and p_A is non-convergent. Thus, p_B can be normalized by dividing by the marginal probability, which is equivalent here to the integral of the unnormalized posterior treated as a function of x. That is, $1/(\pi/2) \times (\pi/2) = 1$. But for p_A, there is no finite normalizing constant, which is what makes it epistemically "worse".

I do this in order to be faithful to De Bona and Staffel's project and create room for evaluating deviations from coherence across updates. In other words, I don't normalize because our starting point is that there is something interesting about the doxastic state of an agent whose credences violate normalization. Another way to put the point is as follows: If we take incoherent priors and normalize them automatically as part of the update rule, then we erase the agent's incoherence by the same stroke. No matter how incoherent one is, the update rule wipes it all out. In that case, approximate coherentism under updating would only apply to violations of additivity, and one could only violate the normalization axiom in one's ur-prior. Moreover, normalizing as part of the update rule would make my own conclusion somewhat circular. I would suggest that there is not much room for approximate coherence judgments under updating because my update rule eliminates any incoherence.

Now, we could normalize right away and draw a similar overall conclusion. The conclusion would be that credences which cannot be normalized by a finite constant are beyond repair, whereas those that can be normalized are all essentially on the same

footing from the perspective of epistemic rationality. The fact that one normalizable credence is closer to coherence than another normalizable credence is epistemically insignificant. Thus, the spirit of the conclusion would be the same: what really matters under updating is the binary question of whether the posterior can be normalized, not the graded question of how far from coherence it happens to be. This is what I have tried to show.

One might still object: couldn't the same type of problem arise for ordinary coherence if we don't normalize as part of the update? It could, but this would be the result of forcing the approximate coherence model into the ordinary coherence framework. In other words, it is not a problem for the ordinary coherentist. The ordinary coherentist is not interested in ranking credence functions in terms of their degrees of incoherence. As a result, it is not problematic for the ordinary coherentist to normalize as part of the update – normalizing does not jeopardize the relevance of coherence as a normative principle for evaluating the epistemic rationality of a person's credences. The same is not true for approximate coherence because, as mentioned, normalizing wipes out precisely that feature of the person's doxastic state that approximate coherence is designed to help us model and evaluate.

For the ordinary coherentist, being incoherent is epistemically bad because one's credence function is then accuracy dominated. But the ordinary coherentist is not in the business of telling us *how* to move to a coherent credence function. Thus, if we ask the ordinary coherentist should the incoherent agent update and normalize, the answer ought to be yes. And if we ask, should the coherent agent update and normalize, the answer should likewise be yes. And neither answer undermines the role of ordinary

coherence as an evaluative norm for the rationality of an agent's credences.

But suppose that in the ordinary coherence framework we have one coherent agent, who updates and normalizes, and one incoherent agent, who updates but does not normalize. Is this a problem for the ordinary coherentist? It is not because in this case they are doing different things. Still, we can take note that the coherent agent will remain coherent. And the incoherent agent will be sometimes coherent and sometimes not. Thus, the non-ideal agent will never become epistemically better, and the ideal one will remain ideal. It is the graded nature of the approximate coherence framework that suggests itself to making ongoing comparisons under updating, but as I hope to have shown, these ongoing comparisons can be misleading.

5.4 Synchronic or Diachronic Norms?

One might further object to the concerns articulated here on the ground that De Bona and Staffel only assess the rationality of a credence function at a time. To what extent, then, should they be persuaded by this notion of comparative consistency over time? Note, as I flagged in Section 4, that while I articulate comparative consistency in terms of what happens before, and after, a learning experience, the temporal framing is a heuristic for illustrating the problem. The notion of comparative consistency is at a minimum a relation among a person's doxastic attitudes. In the preceding paragraphs, I suggested that it would be difficult to banish improper priors from the Bayesian landscape because a proper prior under one representation of a problem commits an agent to an improper prior under a slightly different representation of the same problem. Here, a similar suggestion is appropriate: comparative consistency applies between a

prior (unconditional) credence function and a posterior (conditional) credence function. While the prior/posterior language suggests a temporal perspective, we can also understand the notion in terms of the agent's conditional commitments. My current credence function commits me to a certain credence function in a heads world and to a certain credence function in a tails world. That is, our commitments can be understood from a time-slice centric perspective of rationality, to borrow Hedden (2015) and Moss (2015)'s expression. From an evaluative perspective, the heads and tails worlds, or the estimates conditional on heads and the estimates conditional on tails, are indistinguishable. Yet approximate coherentism's assessment differs among them.

In the large literature pertaining to the proper scope of the requirements of rationality, Broome (2007) argues that they should be understood as state requirements. In the epistemic context, state requirements require the agent's doxastic attitudes to be a certain way at a given time (by comparison, Kolodny (2005) argues for a process interpretation). The notion of comparative consistency suggested here can be understood within a state-based picture of the requirements of rationality. It is a notion of consistency between unconditional doxastic attitudes and the conditional doxastic attitudes one is thereby committed to. The concern on this framing would be that comparative judgments about relative rationality at a time are sensitive to normatively innocuous features of a state (i.e., one's conditional credence given the event Heads vs. one's conditional credence given the event Tails). Note that this is a minimalist conception of comparative consistency. It is of course possible to understand the concept diachronically, in terms of the temporal framing. I only wish to highlight that such a conception is not necessary in order to find the idea compelling.

6 Concluding Remarks

The Bayesian version of the likelihood principle suggests that the only relevant ingredient for drawing inferences should be the posterior distribution. Therefore, whether or not the prior is approximately coherent or, indeed, whether it converges is not especially 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 inferences because the posterior distribution is not identifiable. Therefore, while it is tempting to take a graded approach to evaluating 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 luck.

The lessons of this article, therefore, can be summarized as follows: in the standard Bayesian framework, approximate coherence may not be a helpful guide for evaluating comparative rationality, and in the partial coherence framework, updating should not proceed along Bayesian lines. It remains to be seen what a non-Bayesian updating policy for incoherent agents might look like, how it would compare to Bayesian updating, and on what grounds it might be justified. It would be especially interesting to see future work in this area – i.e., on belief updating norms for imperfectly rational agents.

References

Babic, B. (2019). A Theory of Epistemic Risk. *Philosophy of Science 86*(3), 522–550.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review 78*(1), 1–3.

Broome, J. (2007). Wide or Narrow Scope? *Mind 116*(462), 359–370.

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 78*(3), 405–415.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). New York: CRC Press (Taylor & Francis).

Gneiting, T. and A. E. Raftery (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association 102*(477), 359–378.

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.

Haldane, J. (1932). A Note on Inverse Probability. *Mathematical Proceedings of the Cambridge Philosophical Society 28*(1), 55–61.

Hedden, B. (2015). Time Slice Rationality. *Mind 124*(494), 449–491.

Huttegger, S. M. (2014). Learning Experiences and the Value of Knowledge. *Philosophical Studies 171*(2), 279–288.

Huttegger, S. M. (2017). *The Probabilistic Foundations of Rational Learning*. Cambridge: Cambridge University Press.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London: Series A 186*(1007), 453–461.

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.

Kolodny, N. (2005). Why be rational? *Mind 114*(455), 509–563.

Lindley, D. V. and L. Phillips (1976). Inference for a Bernoulli Process (A Bayesian View). *The American Statistician 30*(3), 112–119.

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

Moss, S. (2015). Time-Slice Epistemology and Action Under Indeterminacy. In J. Hawthorne and T. S. Gendler (Eds.), *Oxford Studies in Epistemology*, Volume 5. Oxford: Oxford University Press.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Pollock, J. L. (1987). Epistemic Norms. *Synthese 71*(1), 61–95.

Robert, C. P. (2007). *The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation*. Springer.

Seidenfeld, T. and L. Wasserman (1993). Dilation for sets of probabilities. *The Annals of Statistics 21*(3), 1139–1154.

Shannon, C. E. (1948a). A Mathematical Theory of Communication. *Bell Systems Technical Journal 27*(3), 379–423.

Shannon, C. E. (1948b). A Mathematical Theory of Communication. *Bell Systems Technical Journal 27*(4), 623–666.

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.

Zabell, S. L. (1982). W.E. Johnson's Sufficientness Postulate. *The Annals of Statistics 10*(4), 1090–1099.