Best Laid Plans: Idealization and the Rationality–Accuracy Bridge[†]

Brett Topey

Abstract

Hilary Greaves and David Wallace argue that conditionalization maximizes expected accuracy and so is a rational requirement, but their argument presupposes a particular picture of the bridge between rationality and accuracy: the Best-Plan-to-Follow picture. And theorists such as Miriam Schoenfield and Robert Steel argue that it's possible to motivate an alternative picture—the Best-Plan-to-Make picture—that does not vindicate conditionalization. I show that these theorists are mistaken: it turns out that, if an update procedure maximizes expected accuracy on the Best-Plan-to-Follow picture, it's guaranteed to maximize expected accuracy on the Best-Plan-to-Make picture as well, in which case moving from the former to the latter can't help us avoid the conclusion that conditionalization is a rational requirement. If there's a problem with Greaves and Wallace's argument, it must lie elsewhere.

1 Introduction

The accuracy-first programme in epistemology is, broadly speaking, the programme of giving decision-theoretic arguments for epistemic norms by appeal to the thesis that accuracy is what's epistemically good. Such arguments typically involve showing that, given that what an agent cares about is accuracy, it will be possible to convince her that shaping her beliefs in accordance with some particular norm is, by her own lights, the best course of action.

One such argument, the one I'll be discussing here, is Greaves and Wallace's ([2006]) accuracy-based argument for Bayesian conditionalization. Greaves and Wallace attempt to justify conditionalization by proving that, as long as the rule by which a probabilistically

[†] This is an author-produced version of an article accepted for publication in *The British Journal for the Philosophy of Science*. The version of record is available at <doi.org/10.1086/718275>.

coherent[1] agent assigns accuracy scores to credence functions in a given state of the world is strictly proper[2], it will always be the case, from the perspective of the agent's present credal state, that conditionalization is the update rule conforming to which will (uniquely) maximize her expected accuracy.

The proof provided by Greaves and Wallace certainly is valid; what philosophical conclusions can be extracted from it, though, is a further question. What the proof is intended to establish is that, given that accuracy is what's epistemically good, agents are rationally required to update their beliefs by conditionalization. But for reasons noted by Pettigrew ([2016], Chapter 15), it can establish this only given an additional premise, one that Greaves and Wallace don't make explicit:

> Follow-Through: If an agent, before receiving some evidence, is required to plan, on receiving that evidence, to update in accordance with a rule, she's also required, on actually receiving that evidence, to in fact update in accordance with that rule.

What the proof on its own shows is just that, before an agent receives some evidence E, she should expect updating on E by conditionalization to be the accuracy-maximizing response in the case where she receives E—it doesn't show that, after she receives E, she should still expect conditionalizing to be the accuracy-maximizing response. After all, when she receives E, she thereby learns that her present credal state is defective (since it doesn't respect all the evidence she now has), and so it's not obvious why she should still be bound by the accuracy judgments of that credal state. At best, then, what the proof can establish, on its own, is that, before receiving that evidence, she should plan to update by conditionalization. In order to get from here to the intended conclusion—that is, that she should in fact update by conditionalization—we need to appeal to Follow-Through.

Even if we grant Follow-Through, though, it's controversial whether Greaves and Wallace's proof establishes what it's intended to establish. As Schoenfield ([2015], [2018]) and Steel ([2018]) have recently pointed out, what the proof shows, strictly speaking, is only that agents should expect that conditionalization is the update rule successfully conforming to which will be accuracy-maximizing. To get from this result to the conclusion that conditionalizing is a rational requirement, Schoenfield and Steel argue, we must assume a particular picture of the bridge between rationality and accuracy, a picture on which the update rule to which an agent should plan to conform is that rule that she expects would maximize

[1] Greaves and Wallace assume that probabilistic coherence is a rational requirement, and so they assume that the agents they discuss have belief states that can be modelled as probability distributions. I'll be working under the same assumption.

[2] For an accuracy scoring rule to be strictly proper is for it to be such that every probabilistically coherent credence function will always expect itself to be more accurate than alternative credence functions, so that expected accuracy considerations always motivate the agent 'strictly to stick to [her] current credence distribution until and unless new evidence comes along' (Greaves and Wallace [2006], p. 626).

her accuracy were she to succeed in conforming to it.[3] That is, we need to assume the bridge principle Schoenfield ([2015], p. 641) calls 'Best-Plan-to-Follow' and states as follows:

> Best-Plan-to-Follow: 'The rational epistemic plan is the one that a rational agent would choose, a priori, if she were aiming to maximize the expected accuracy of the credences that an agent following the plan would adopt'.

And this, Schoenfield and Steel suggest, is a problematic assumption. The reason is that we often find ourselves in situations in which we have higher-order evidence to the effect that we're imperfect reasoners, evidence that makes it predictable that, if we plan to conform to some particular rule r, we'll fail to do so. And in such situations the expected accuracy of successfully conforming to r doesn't seem like it should be particularly relevant to us in our planning, since we expect that, even if we were to commit to conforming to r, we'd end up doing something else instead. Steel ([2018], p. 26) puts the point succinctly: 'When considering what we should believe […] we ought to take account of evidence that we will not succeed at doing what we try'. In particular, even if we expect that conforming to conditionalization would be accuracy-maximizing, we shouldn't commit to conditionalizing if we have good evidence that we'll fail to do so and so will end up worse off, from the point of view of accuracy, than we would've ended up had we committed to following some other rule.

Schoenfield's ([2015], p. 653) proposed fix, a version of which Steel endorses, is to replace Best-Plan-to-Follow with a different principle, which she calls 'Best-Plan-to-Make' and states as follows:

> Best-Plan-to-Make: 'The rational epistemic plan is the one that a rational agent would choose, a priori, if she were aiming to maximize the expected accuracy of the credences an agent would adopt as a result of *making* the plan'.

The thought is that, when we calculate the expected accuracy of the credences that would result from making a plan, we don't simply assume that we'll succeed in executing that plan—instead, we think about what credences we'd in fact adopt as a result of making the plan, even if those are different from the credences the plan recommends. And that means that an account of the rationality–accuracy bridge based on Best-Plan-to-Make, unlike one based on Best-Plan-to-Follow, 'allows us to take into account, when evaluating an update procedure, our opinions concerning how successful we are likely to be at following it' (Schoenfield [2018], p. 711). As a result, the Best-Plan-to-Make picture, according to Schoenfield and Steel, allows us to avoid the verdict that we should always plan to

[3] In fact, according to Schoenfield ([2018], pp. 702–03), what the proof shows is even weaker, since it relies on an assumption about evidence that (she says) doesn't hold if we allow the agent's evidence to include self-locating propositions. Incidentally, I don't think Schoenfield's argument here works—as I discuss in my ([unpublished]), the formula she relies on to calculate the expected accuracy of update rules fails to capture what it's intended to capture when self-locating propositions are in play and so needs to be revised in order to return the correct results in cases of self-locating evidence. This is arguably the lesson of (Bradley [2020]).

conditionalize (and, more generally, allows us to return intuitively correct verdicts—that is, calibrationist verdicts—in cases of higher-order evidence).

What I'll be arguing here, though, is that this proposal is based on a misdiagnosis. What's motivating the adoption of the Best-Plan-to-Make picture is, again, the thought that the Best-Plan-to-Follow picture doesn't allow us to take into account evidence suggesting that, if we plan to conform to a given rule, we'll fail to do so. But this thought is mistaken. The Best-Plan-to-Follow picture does give us the resources to take this sort of higher-order evidence into account. In fact, the following is true:

> Best Plan Preservation: If any account based on Best-Plan-to-Follow returns the verdict that r is the rational rule, then the account that results from replacing Best-Plan-to-Follow with Best-Plan-to-Make (without making other unrelated changes) will also return the verdict that r is the rational rule.

That is, given any account based on Best-Plan-to-Follow, replacing that principle with Best-Plan-to-Make will make no difference to what update rule the account recommends.[4] In this sense, the Best-Plan-to-Follow picture and the Best-Plan-to-Make picture are simply equivalent. So, if there is indeed something wrong with Greaves and Wallace's argument for conditionalization, it's not a problem that can be resolved by replacing Best-Plan-to-Follow with Best-Plan-to-Make. The problem must lie elsewhere.

But where, exactly? The answer, I suggest, is that the problem, in so far as there is one, is simply that the agents Greaves and Wallace are interested in are idealized to an inappropriate degree. But further discussion of this diagnosis will have to wait until Section 7. First we must discuss some of the details of the Best-Plan-to-Follow picture and the Best-Plan-to-Make picture so that we can make clear just why it is that Best Plan Preservation is true.

2 Best-Plan-to-Follow, Availability, and Idealization

We begin with a definition. An update rule, as we'll be using the term, is what Greaves and Wallace ([2006], p. 612) call an 'epistemic act': a function r from world states to probability functions, where, if $r(s) = cr$, 'then cr is the probability function that an agent performing act r [that is, conforming to r] would adopt as his credence distribution if state s in fact obtained'.[5]

Notice, though, that if the set of rules from which the agent is choosing includes every function from world states to probability functions, then Best-Plan-to-Follow can't deliver the verdict that conditionalization is the rational rule. After all, one such function is the truth rule: the function that, for any world state s, returns a credence distribution cr such that $cr(p) = 1$ if p is true in s and $cr(p) = 0$ otherwise. And it's obvious that this rule is going to be the one conforming to which would maximize expected accuracy, in which case, by Best-Plan-to-Follow, this rule—not conditionalization—is going to be the rational rule.

---

[4] Here (and throughout the paper) I assume for simplicity that there's a unique available rule conforming to which would maximize expected accuracy—that is, that an account based on Best-Plan-to-Follow will recommend one update rule in particular as rational. All of the reasoning to follow can be adapted so as to eliminate this assumption.

[5] I've replaced the symbols here with the analogous ones from my own presentation.

The way Greaves and Wallace avoid this result is by restricting their attention to 'available' rules: rules to which the agent is, in a certain sense, able to conform. The truth rule, they say, isn't genuinely available—conforming to it is something the agent is 'not able' to do, since doing so 'would require the agent to respond to information that he does not have' ([2006], p. 612). Their result, then, is that, of the rules that are available, conditionalization is the one conforming to which would maximize expected accuracy.

Of course, this invites the question of what conditions a rule must meet in order to count as available to an agent. Greaves and Wallace's own answer is that the available rules are those that are consistent with evidentialism—that is, those that can be represented as functions from evidential states to credence distributions, so that, if some r is such that, for some world states s_1 and s_2, $r(s_1) \neq r(s_2)$ despite the agent's evidence being the same in both world states, then r isn't an available rule. But this answer is offered without much in the way of explicit justification—it's simply assumed that agents are able to conform to a rule unless conforming to it requires responding to information one doesn't have. What's motivating this assumption?

The answer, as far as I can tell, is simply that the agents Greaves and Wallace are interested in are highly idealized agents, ones that, aside from their limited access to information about the world, are subject to no cognitive limitations whatsoever: only on the assumption that an agent is ideal in this way is it reasonable to assume that every evidentialist update rule is one to which she's able to conform. Greaves and Wallace, then, are working with a highly idealized notion of what rationality requires.[6] This isn't a problem, necessarily. But it is a particular choice Greaves and Wallace have made. It's worth taking a moment to think about what the Best-Plan-to-Follow picture would look like were we to choose to consider agents idealized to a different degree.

Here the first thing we need to do is to say something about what it means, in general, to say that an agent is able to conform to a rule. And it's reasonably clear that what's relevant here isn't simply whether it's possible to so conform: it's logically, metaphysically, and even physically possible for an agent to happen, by accident, to have perfectly accurate credences (and so to conform to the truth rule) even if that agent is highly cognitively limited, but such an agent certainly isn't able to conform to the truth rule in any sense that's of interest to us. What's relevant must instead be whether, given that the agent commits to conforming to the rule, success in so conforming is in some sense the expected result.

Furthermore, we can say something plausible about what, in the present context, this sort of expectation comes to. Recall that we're working in a decision-theoretic paradigm, one on which determining what's rational involves determining what is the best course of action by an agent's own lights. So, for some rule r to be the rational rule, it must be the case that the agent herself should take r to be the available rule conforming to which would maximize expected accuracy. And that means that whether a rule counts as available, in the sense relevant here, must be a question that can be answered from the agent's own perspective: availability must be a matter of whether the agent herself should regard conforming to

[6] Note that they're explicit about this: they begin their paper by explaining that what they're attempting to justify is the Bayesian claim that an 'ideal epistemic agent' will update by conditionalization ([2006], pp. 607–08).

that rule as an option that's genuinely open to her.[7] So, in the present context, whether some agent is able to conform to a rule must be a matter of whether she herself is reasonable in taking the expected outcome of committing to conforming to that rule to be success in so conforming—that is, is reasonable in being entirely sure, on the assumption that she commits to conforming to the rule, that she'll succeed in doing so.[8]

With this general picture in place, it's easy to see the way in which the availability of rules depends on how and to what degree we choose to idealize. If, for instance, the idealized agents in question are subject to no cognitive limitations (and are reasonable in being certain of this fact), then they should expect that they'll be successful in their attempts to conform to any evidentialist rule whatsoever.[9] But we might also choose to idealize in a less extreme way. We might, for instance, consider agents who are reasonable in expecting themselves to be perfect calculators only up to a certain degree of complexity, so that there are evidentialist rules to which these agents should not expect themselves to succeed in conforming. This might be a way of generating an account of rationality that's a bit more human-scaled than Greaves and Wallace's account. Or we might go in the other direction, choosing to make our idealization more extreme—if we're interested in an externalist account of rationality, for instance, we might consider idealized agents that are reasonable in taking themselves to have some magical mechanism by which to respond to information they don't have, in which case they should expect themselves to succeed in their attempts to conform to at least some nonevidentialist rules. We might even choose not to idealize at all, in which case whether a rule counts as available will depend on whether we ourselves should expect, on the assumption that we commit to conforming to it, that we'll succeed.[10]

---

[7] Bronfman ([2014], pp. 887–88) argues, much as I have, that it isn't mandatory to follow Greaves and Wallace in taking all evidentialist update rules to be available, but he takes what's relevant to a rule's availability to be not what the options the agent should regard as open to her but 'what the agent's abilities are', in some more objective sense. Schoenfield ([2017], p. 1177) points out, though, that, since Bronfman is still working in a decision-theoretic paradigm, 'it seems against the spirit of [his] proposal to demand that the agent update in accord with the update procedure that maximizes expected accuracy relative to her *actual* abilities when she has no way of knowing which update procedure this is'.

[8] Or, at least, is reasonable in being sure of this if she also assumes that she'll remain rationally evaluable at all. (The possibility that she'll fall into a coma, for instance, is obviously not going to be relevant here.) If this seems too strict a requirement, consider: our decision-theoretic paradigm is, again, one on which the relevant question is what is the best course of action by the agent's own lights, and if the agent couldn't be certain that committing to conforming to a given rule would result in actually conforming to it, then whether conforming to that rule maximizes expected accuracy couldn't on its own settle whether committing to conforming to it is, from the point of view of accuracy, the best course of action. See Section 6 for some discussion of how to model the expectations of agents who don't have this kind of certainty.

[9] The parenthetical is crucial. If an agent is subject to no cognitive limitations but can't reasonably be sure of this fact, then there will be evidentialist rules to which the agent can't reasonably expect to succeed in conforming. Greaves and Wallace's choice of idealization, then, is one on which the agent is subject to no cognitive limitations and knows this about herself.

[10] Incidentally, Bronfman's view appears to be that we shouldn't idealize at all: what's relevant, for him, is what the actual agent's abilities are. But he's not very explicit about this.

The point: Best-Plan-to-Follow doesn't on its own tell us anything at all about what the available rules are. To answer that question, we must first choose how and to what degree to idealize. Greaves and Wallace make a particular choice here, but accepting the general Best-Plan-to-Follow picture doesn't commit us to this choice—that picture is compatible with a wide variety of other choices as well, with different choices corresponding to different verdicts about what rules are available. This fact is going to play a critical role in our argument.

3 A Generalization of the Notion of a Rule

I've followed Greaves and Wallace in taking an update rule to be a function from world states to (probabilistic) credence distributions. But it's worth noting that such rules don't represent the only possible ways for an agent to set her credences. An agent might, after all, randomize: she might update in a nondeterministic way. And a function from world states to probability functions can't model this behaviour.[11]

There's a simple fix. We can allow that an update rule, rather than being a function from world states to credence distributions, is a probability distribution over functions from world states to credence distributions, where this probability distribution represents how likely it is, for each function from world states to credence distributions, that an agent conforming to the rule will update in accordance with that function. On this generalized definition, an update rule is simply a probability distribution over deterministic rules—that is, rules as previously defined. Furthermore, any deterministic rule can itself be represented as an update rule in our generalized sense: the degenerate probability distribution that assigns probability 1 to the relevant function from world states to credence distributions and probability 0 to all other such functions.

It's also reasonably easy to see that generalizing in this way won't interfere with Greaves and Wallace's proof that conditionalization is the evidentialist rule conforming to which would maximize expected accuracy: it's well known that randomization, though it's often strategically beneficial in games against intelligent opponents, can be of no help at all in straightforward decision problems of the sort we're discussing here.[12] To see why, note first that the expected accuracy (relative to a credence distribution cr) of conforming to a probabilistic update rule—that is, a probability distribution p over deterministic rules

[11] Strictly speaking, this claim is too strong: if an agent updates in a nondeterministic way by letting her credences be influenced by the behaviour of some worldly source of randomness such as a radioisotope, she can be modelled as conforming to a function from world states to credence distributions in which this worldly source of randomness plays the role of a random number generator. So, in so far as it's reasonable to suppose that all nondeterministic updating has this sort of worldly source, there's no real need to generalize Greaves and Wallace's notion of a rule. Even so, though, there's no harm in generalizing in the way suggested in this section, and doing so turns out to be convenient for presenting my argument. Thanks to an anonymous referee for bringing this issue to my attention.

[12] In fact, the only single-agent decision problems in which randomization can be of any help are extensive decision problems with absentmindedness—that is, problems in which the agent faces what looks subjectively like the same decision twice, without remembering whether she's faced it before. As Piccione and Rubinstein ([1997], Section 4) discuss, this was proved by Isbell ([1957]).

$r_1, r_2, \ldots, r_n$—can be characterized in terms of the expected accuracies of conforming to each of those deterministic rules, as follows:

$$EA^{cr}(p) = \sum_{i=1}^{n} p(r_i) \times EA^{cr}(r_i)\,.$$

That is, the expected accuracy of conforming to p is just the probability-weighted sum of the expected accuracies of conforming to the various deterministic rules over which p ranges. Suppose, then, that, in a given class C of deterministic rules, r is the one conforming to which maximizes expected accuracy. Can there be a probability distribution p over the rules in C (other than the degenerate one that assigns probability 1 to r) such that $EA^{cr}(p) \geq EA^{cr}(r)$? No: in so far as p assigns any nonzero probability to a rule other than r, it gives weight to some $r_i \in C$ such that $EA^{cr}(r_i) < EA^{cr}(r)$, and so, since there can be no compensating $r_j \in C$ such that $EA^{cr}(r_j) > EA^{cr}(r)$, it will be the case that $EA^{cr}(p) < EA^{cr}(r)$.[13]

In short, generalizing so as to allow probabilistic update rules changes very little, on the Best-Plan-to-Follow picture. This, I suspect, explains why Greaves and Wallace didn't see a need to consider randomization in the first place. That said, the generalized notion of an update rule will turn out to make presenting our argument much simpler. So this is the notion we'll be working with.

4 Best-Plan-to-Make, Failure to Conform, and Idealization Again

Now: according to Best-Plan-to-Make, what it takes for an update rule to be the rational rule is not for it to be the available rule successfully conforming to which would maximize expected accuracy. Instead, the rational rule is that rule planning to conform to which would maximize expected accuracy. And what motivates adopting Best-Plan-to-Make, Schoenfield and Steel suggest, is that an account of the rationality–accuracy bridge based on this principle, unlike one based on Best-Plan-to-Follow, has the resources to allow agents to acknowledge the gap between making an epistemic plan and executing that plan, between committing to conforming to an update rule and actually conforming to that rule, and so has the resources to allow agents to take into account, when determining which rule maximizes expected accuracy, the possibility of trying and failing to traverse that gap.

Here alarm bells should be going off already. Again, on accounts based on Best-Plan-to-Follow, the set of rules that are under consideration isn't the set containing all update rules—it's a restricted set containing just those rules that are available, in the sense discussed above. And the point of this restriction is precisely to ensure that Best-Plan-to-Follow doesn't require agents to consider rules to which they're unable to conform. So it appears that Best-Plan-to-Follow, despite what Schoenfield and Steel suggest, does give agents a way to take into account the possibility of trying and failing to conform to a rule. It's just that it does so by allowing agents to eliminate certain rules from consideration altogether rather than by having agents incorporate possible cases of failure to conform to a rule into their expected accuracy calculations.

[13] Compare Theorem 11 in (Pettigrew [2020]).

Of course, to observe that this is so isn't yet to establish that our two principles do an equally good job of allowing agents to take into account the possibility of trying and failing to conform to a rule. The two approaches just described, after all, are on the surface quite different from one another. Might advocates of Best-Plan-to-Make acknowledge our observation and yet continue to insist that the former approach leads to incorrect verdicts about particular cases, in which case the latter approach—that is, the Best-Plan-to-Make approach—is to be preferred?

No. There isn't any logical space for such a claim, for the simple reason that the two approaches here, despite appearances, don't turn out to be substantively different at all: as I claimed above, Best Plan Preservation is true. I'll argue for this claim in a moment; first, we need to discuss the role of idealization in the Best-Plan-to-Make picture.

According to Best-Plan-to-Make, recall, the rational rule is the one that a rational agent would take to be such that making a plan to conform to it would maximize expected accuracy. This, notice, leaves entirely open the question of in what sense the agent in question is rational—that is, the question of what this agent's cognitive resources are, of how and to what degree this agent is idealized. And it's perfectly clear that what credences an agent will adopt as the result of making a given plan is going to depend on what the agent's cognitive resources are, since those resources are going to determine whether she'll succeed in conforming to the rule to which she plans to conform.

In short, Best-Plan-to-Make, like Best-Plan-to-Follow, is compatible with a variety of choices about how and to what degree to idealize, with different choices corresponding to different verdicts about what the expected outcome would be of planning to conform to a given rule. We might, for instance, take what's relevant to whether a rule is rational to be whether it's the rule that we ourselves, with all our cognitive failings, would be rational in taking to be such that planning to conform to it would maximize expected accuracy—this would amount to choosing not to idealize at all. But we might instead take what's relevant to be what an agent reasonable in taking herself to be subject to no cognitive limitations would be rational in expecting—this would amount to choosing to idealize in the same extreme way that Greaves and Wallace do.

Incidentally, neither Schoenfield nor Steel explicitly discusses idealization in this context. As far as I can tell, though, Steel is of the opinion that we shouldn't idealize at all. He claims, for instance, that, in certain cases, 'I have excellent evidence both that I will not succeed at conditionalization and that the results will be bad', and from this claim he infers that, in the cases in question, the right thing to do is to 'avoid trying to conditionalize on that evidence' ([2018], p. 26). And this suggests that, on his view, what's relevant to whether conditionalization is the rational rule is what we ourselves, limited as we are, should expect will happen if we plan to conditionalize.

As for Schoenfield: I suspect that her package of views commits her to at least some degree of idealization. She claims, for instance, that 'beliefs that are unsupported by the evidence due to wishful thinking or fear of disappointment are irrational *even if* the wishful thinkers or fearers can't help themselves *and know they can't*' ([2018], p. 693). And this suggests that, even in a case where, for example, the actual (nonidealized) agent knows that she can do no better, from the point of view of expected accuracy, than to conform to a rule r_w that sanctions some degree of wishful thinking (and knows also that, if she plans to

conform to r_w, she'll succeed), r_w is nevertheless not rational; the rational rule must instead be some rule that sanctions no wishful thinking at all. But this result can be reconciled with Best-Plan-to-Make only by idealizing, only by allowing that what's relevant to what rule is rational isn't what the actual agent should expect that she'd do on making various plans but what an agent idealized so as to be capable of doing better than conforming to r_w (and to know that she's so capable) should expect that she'd do on making various plans.[14]

Ultimately, though, the commitments of particular Best-Plan-to-Make advocates don't matter very much, for our purposes. The point here is just that we have a choice to make about how and to what degree to idealize, and the general Best-Plan-to-Make picture is like the general Best-Plan-to-Follow picture in that it doesn't commit us to any particular choice. To put it plainly: the question of what sort of idealization is appropriate is separable from, and independent of, the question of which bridge principle to adopt.

5 Comparing the Principles I: The Schoenfield–Steel Strategy

This brings us to our crucial observation: given the separability of choice of idealization from choice of bridge principle, the correct way to determine what difference our choice of bridge principle makes to our overall account of rationality is to consider what would change were we to replace one of these principles with the other while holding our choice of idealization fixed. To fail to hold fixed our choice of idealization would, after all, be to introduce a confounding variable into our investigation. And this, I submit, is where Schoenfield and Steel have gone wrong.

Recall that what motivates the move from Best-Plan-to-Follow to Best-Plan-to-Make, for both Schoenfield and Steel, is the claim that the latter, unlike the former, allows us to avoid the verdict that conditionalization is the rational rule. And each of their arguments for this claim depends crucially on the premise that, in some evidential situations, a rational agent will expect, on the assumption that she makes a plan to conditionalize, that she'll fail to execute that plan—in particular, will fail to execute that plan in such a way that she'll end up worse off, from the point of view of expected accuracy, than she'd have ended up had she planned from the start to conform to some less demanding rule, some rule to which she could expect herself to succeed in conforming. This is why it's supposed to be the case that, on the Best-Plan-to-Make picture, the rational rule is some rule that's less demanding than conditionalization.[15] But notice: on Greaves and Wallace's account—in particular, on their

[14] The reason, very briefly, is that, by Best-Plan-to-Make, for a rule to be such that planning to conform to it would maximize expected accuracy is both necessary and sufficient for that rule to be rational. Given this fact, the only way to avoid the conclusion that r_w is rational is to insist that planning to conform to it wouldn't maximize expected accuracy—that is, that the agent whose expected behaviour is relevant isn't the actual (nonidealized) agent.

[15] Steel's ([2018], p. 26) argument, remember, is simply that 'I should avoid trying to conditionalize' in cases in which 'I have excellent evidence both that I will not succeed at conditionalization and that the results will be bad'. As for Schoenfield, she suggests both that conditionalizing involves steadfastly apportioning one's beliefs to one's first-order evidence and that, in the face of certain sorts of higher-order evidence, an agent should not expect that she'll succeed in steadfastly apportioning her beliefs in this way. This is why it can be the case that '*planning* to calibrate does better than *planning*

choice of idealization—conditionalization is an available rule. And this is just to say that, from a rational agent's perspective, the expected result of planning to conform to that rule is that she'll succeed in doing so. So, if what we're doing is taking Greaves and Wallace's account and replacing Best-Plan-to-Follow with Best-Plan-to-Make, the premise on which Schoenfield's and Steel's arguments depend is simply false.

More generally (and more formally): in so far as a rule counts as available, the expected accuracy of a rational agent's planning to conform to the rule will be exactly the same as the expected accuracy of successfully conforming to the rule, since the agent will expect that, were she to plan to conform to the rule, she'd succeed in doing so. That is, for any available rule r,

$$EA^{cr}(r) = EA_p^{cr}(r), \tag{1}$$

where $EA^{cr}(r)$ is the expected accuracy (relative to the agent's credence distribution cr) of the credences that would result from conforming to r and where $EA_p^{cr}(r)$ is the expected accuracy of the credences that the agent expects would result from her planning to conform to r. And this fact entails a restricted version of Best Plan Preservation, which we can state as follows:

> Best Available Plan Preservation: If an account based on Best-Plan-to-Follow returns the verdict that r is the rational rule, then, given any other available rule r', replacing Best-Plan-to-Follow with Best-Plan-to-Make while holding fixed our choice of idealization will always result in an account that recommends r over r'.

Proof: Suppose an account based on Best-Plan-to-Follow returns the verdict that r is the rational rule. This is just to say that r is available and that, for any other available rule r', $EA^{cr}(r) > EA^{cr}(r')$. Furthermore, given that both r and r' are available, we know by Equation 1 both that $EA^{cr}(r) = EA_p^{cr}(r)$ and that $EA^{cr}(r') = EA_p^{cr}(r')$. So $EA_p^{cr}(r) > EA_p^{cr}(r')$. And this is just to say that, if we adopt Best-Plan-to-Make while holding fixed how and to what degree the relevant agent is idealized, the result will be an account that recommends r over r'. □

It follows immediately from Best Available Plan Preservation that, if we start with an account based on Best-Plan-to-Follow and then replace that principle with Best-Plan-to-Make while holding fixed our choice of idealization, our new account can never recommend as rational a rule less demanding than the rule r recommended by the old one—all of these less-demanding rules will, after all, be available rules other than r. And this is enough to show that the motivations offered by Schoenfield and Steel for replacing Best-Plan-to-Follow with Best-Plan-to-Make are fundamentally misguided.

---

to be steadfast [from the point of view of expected accuracy], even though steadfasting does better than calibrating' ([2018], p. 710).

6 Comparing the Principles II: Proving Best Plan Preservation

That said, there remains a question here. Best Available Plan Preservation tells us that, if an account based on Best-Plan-to-Follow recommends as rational some rule r, the account that results from replacing Best-Plan-to-Follow with Best-Plan-to-Make will recommend r over any other available rule. But what about rules that are not available? We know that, if a rule isn't available, a rational agent will expect that, on planning to conform to it, she won't succeed—she'll end up doing something else instead. And nothing I've said up to this point has ruled out the possibility that, among the unavailable rules, there's some r' that, first, presents a genuine alternative to r (that is, is such that the agent expects that, on planning to conform to r', she'll form credences that are different from those that would result from planning to conform to r), and second, is such that the agent expects that planning to conform to it would result in credences at least as accurate as those that would result from planning to conform to r. So: can we rule out this possibility?

It's worth emphasizing that we are at this point merely exploring logical space. No one, as far as I know, has ever seriously advanced a view on which the reason for preferring Best-Plan-to-Make to Best-Plan-to-Follow is that there may be some unavailable rule planning to conform to which does at least as well, from the point of view of expected accuracy, as planning to conform to any available rule. But it's logical space worth exploring. I claimed above that replacing Best-Plan-to-Follow with Best-Plan-to-Make can never make any difference to what rule an account recommends—that is, that Best Plan Preservation is true. And Best Available Plan Preservation on its own doesn't entail that this is correct; in order to show that it's correct, we must also rule out the possibility of an unavailable r' with the features just described. That is, we must demonstrate the truth of the following:

> Unavailable Plan Exclusion: If an account based on Best-Plan-to-Follow returns the verdict that r is the rational rule, then, given any unavailable rule r' that presents a genuine alternative to r, replacing Best-Plan-to-Follow with Best-Plan-to-Make while holding fixed our choice of idealization will always result in an account that recommends r over r'.

Only then will we have shown definitively that there's no substantive difference between the Best-Plan-to-Follow picture and the Best-Plan-to-Make picture.

As it turns out, we can show that Unavailable Plan Exclusion is true. To see how, consider first that, for a rule r' to be a counterexample to that thesis, it must be the case that $EA_p^{cr}(r') \geq EA_p^{cr}(r)$. So, in order for r' even to be a candidate for being a counterexample, $EA_p^{cr}(r')$ must be defined, which means the agent must have some expectation about what credences she'll in fact form in various situations, given a plan to conform to r'. That is, while there may be some unavailable rules so unimaginably complex or bizarre that the agent has no expectation whatsoever about what credences she'll form on planning to conform to those rules, no such rule is a candidate for being a counterexample to Unavailable Plan Exclusion. So, if r' is indeed a candidate, the agent does have some expectation here. And in so far as the agent has some expectation, it can be represented as the expectation that, on planning to conform to r', she'll in fact conform to some other rule r''.

Notice: the agent's expectation can be represented in this way even if she's less than certain about what she'll do on planning to conform to r'. We simply need to make use of our generalization of the notion of a rule from Section 3. Suppose, for instance, that she's 70% sure that she'll in fact conform to some deterministic rule r_1 and 30% sure that she'll in fact conform to some other deterministic rule r_2. Then her expectation can be represented as the expectation that she'll conform to r'', where this is a probabilistic rule such that $r''(r_1) = 0.7$ and $r''(r_2) = 0.3$.[16] And similarly for other cases of uncertainty.[17]

Now, there are two important implications of the fact that the agent expects that, on planning to conform to r', she'll in fact conform to r''. The first is immediate: in so far as what the agent expects here is that she'll conform to r'', the expected accuracy of planning to conform to r' is exactly the same as the expected accuracy of actually conforming to r''. That is,

$$EA_p^{cr}(r') = EA^{cr}(r''). \tag{2}$$

And the second, though it's not immediate in the same way, is, I take it, still highly plausible: in so far as what the agent expects here is that she'll conform to r'', r'' is an available rule. Or, in other words, the following thesis is true:

> Expected Plan Availability: If r'' is some rule to which a rational agent expects she'll in fact conform on planning to conform to some rule r', then r'' is available.

The reason this is plausible is simply that the agent has a strategy available that, by her own lights, is expected to result in successfully conforming to r''—namely, the strategy of doing her best to conform to r'. Since she knows this strategy is available, it seems clear enough that she can reasonably expect, given that she commits to conforming to r'', that she'll succeed in doing so. And this is all that's required in order for r'' to count as available.

With these two implications in hand, we're in a position to provide a proof of Unavailable Plan Exclusion.

---

[16] This is a bit of an abuse of notation, but it's not really a problem. What will be important for our purposes is the expected accuracy of the agent's planning to conform to r', and this is calculated in just the same way regardless of which of the following is true:

(i) the agent is certain that, on planning to conform to r', she'll in fact conform to some genuinely probabilistic rule r''; or

(ii) the agent is certain that, on planning to conform to r', she'll in fact conform to some deterministic rule, but she's uncertain which one—in particular, is such that, for any deterministic rule d, her credence that she'll conform to d is $r''(d)$.

[17] Here I've assumed for simplicity that, even if the agent can't be sure what rule she'll in fact conform to, she can be sure that she'll conform to some deterministic rule. But this assumption can be relaxed: if some of the rules to which the agent thinks she might conform are themselves probabilistic, then her expectation can be represented as a probability distribution over probability distributions over deterministic rules, and this is in the end just equivalent to a probability distribution over deterministic rules.

Proof: Suppose some account based on Best-Plan-to-Follow returns the verdict that r is the rational rule. Then r is the available rule conforming to which would maximize expected accuracy. Suppose for *reductio*, then, that r' is some unavailable rule that, first, presents a genuine alternative to r, and second, is such that $EA_p^{cr}(r') \geq EA_p^{cr}(r)$. Then there's some rule that's the rule to which the relevant agent expects she'll actually conform on planning to conform to r'—call this rule r''. By Equation 2, $EA_p^{cr}(r') = EA^{cr}(r'')$. And since r is available, we know, by Equation 1, that $EA^{cr}(r) = EA_p^{cr}(r)$. So $EA^{cr}(r'') \geq EA^{cr}(r)$.

Furthermore, since r' presents a genuine alternative to r, the agent does not expect planning to conform to r' to result in actually conforming to r. That is, $r'' \neq r$. In addition, Expected Plan Availability guarantees that r'' is available. So r'' is an available rule other than r such that $EA^{cr}(r'') \geq EA^{cr}(r)$. So r is not the available rule conforming to which would maximize expected accuracy. Contradiction. □

With this proof of Unavailable Plan Exclusion, we've completed the task of showing that Best Plan Preservation is true, since, again, the latter thesis follows immediately from Best Available Plan Preservation and Unavailable Plan Exclusion. What this tells us is the following: not only are the particular motivations offered by Schoenfield and Steel for moving from Best-Plan-to-Follow to Best-Plan-to-Make misguided; any motivation for this move, if it goes via the claim that an account based on the latter makes intuitively correct recommendations in cases in which an account based on the former does not, is necessarily misguided, for the simple reason that, by Best Plan Preservation, replacing Best-Plan-to-Follow with Best-Plan-to-Make never makes any difference to what rule an account recommends.

7 Diagnosis and Philosophical Upshots

Assuming all this is right, two questions immediately arise. First, if which principle we adopt can never make a difference to what rule an account recommends, why is it that both Schoenfield and Steel think it can make a difference? And second, if the problem with Greaves and Wallace's argument for conditionalization isn't that it presupposes the Best-Plan-to-Follow picture, what is the problem (if indeed there is one)? We'll close with an attempt to answer these questions.

To answer the first question: Schoenfield and Steel's mistake, I want to suggest, arises from the fact that they begin with a different notion of an update rule than that with which Greaves and Wallace begin. In particular, Schoenfield and Steel characterize a rule as a function from evidential states to credence distributions. Greaves and Wallace, on the other hand, begin with a broader characterization, one on which a rule is a function from world states to credence distributions, and then argue that only a subset of those rules are genuinely up for consideration: those that are consistent with evidentialism—that is, are representable as functions from evidential states to credence distributions. In other words, Greaves and Wallace endorse a restriction on what rules are legitimate, but where they end up, given this restriction, is where Schoenfield and Steel begin. As a result, Schoenfield and Steel don't see this as a restriction at all—when they evaluate Greaves and Wallace's picture, they do so as though no restriction has been introduced.

The reason this is significant is that, as discussed in Section 4, the point of Greaves and Wallace's restriction is precisely to take into account the possibility of an agent's trying and

failing to conform to certain rules—the rules they allow, remember, are just those that are genuinely available to the agent. So, in neglecting the fact that a restriction has been introduced at all, Schoenfield and Steel thereby also neglect the fact that the Best-Plan-to-Follow picture provides a mechanism for taking this possibility into account. This, I suspect, is what explains why they think moving from Best-Plan-to-Follow to Best-Plan-to-Make can make a difference: they see that move, incorrectly, as a way of introducing such a mechanism where none existed before.

This, incidentally, brings us to a possible objection to my argument that's worth pausing to consider. I've claimed that Schoenfield neglects the fact that Greaves and Wallace have introduced a restriction, but it's certainly not the case that she ignores this fact altogether. Indeed, at one point she explicitly considers the view that we can motivate the restriction by appeal to the thought that agents should be allowed to eliminate from consideration rules to which they don't expect to succeed in conforming. But she rejects this way of motivating the restriction as 'unpromising' on the grounds that it's not compatible with the claim about wishful and fearful thinking mentioned above—that is, that these kinds of thinking are irrational even in cases in which the agent knows she's unable to avoid them ([2018], p. 693). The reason Schoenfield doesn't see Greaves and Wallace's restriction as a mechanism for taking into account the fact that some rules are unavailable to the agent, then, seems to be that she thinks these cases of wishful and fearful thinking are counterexamples to the view that mere unavailability gives us good reason to eliminate rules from consideration in the first place, in which case the decision to consider only rules consistent with evidentialism must have some entirely different motivation. And if this is correct, there's something fundamentally wrong with the conception of the Best-Plan-to-Follow picture on which my argument depends. So: is it correct?

No. Which rules count as available, recall, is a function of how and to what degree we choose to idealize. And it's clear enough that, on a wide range of choices of idealization, Schoenfield's cases of wishful and fearful thinking aren't genuine counterexamples to Greaves and Wallace's availability-based motivation for eliminating certain rules from consideration. On Greaves and Wallace's own choice of idealization, for instance, the agents relevant to availability are certain that they're subject to no cognitive limitations whatsoever and so are certain that they're capable of avoiding wishful and fearful thinking, which means that any case in which an agent knows she can't avoid these kinds of thinking is thereby a case in which that agent isn't relevantly idealized. Similarly for the sort of idealization to which I argued in Section 4 that Schoenfield herself is committed: here, too, the relevant agents are certain that they can avoid the kinds of thinking that Schoenfield is suggesting are irrational, which means those agents can never find themselves in the situations she describes. Indeed, a more general result is available: the very same reasoning I used in arguing that this sort of idealization is a commitment of Schoenfield's also shows that, in so far as a kind of thinking really is irrational despite the fact that the actual agent knows she can't avoid it, this can only be because the actual agent isn't the one whose expected behaviour is relevant—the appropriate choice of idealization must be one on which relevantly idealized agents can indeed avoid that kind of thinking (and know that they can). And if that's right, cases in which an agent knows she's unable to avoid that kind of thinking simply

can't be genuine counterexamples to Greaves and Wallace's availability-based motivation for introducing a restriction on what rules are up for consideration. [18]

Back to the main thread: attending to the restriction introduced by Greaves and Wallace can also help us to answer our second question—that is, the question of what, if anything, is wrong with their argument for conditionalization. What bears emphasizing here is that their endorsing the particular restriction they do, as opposed to some other restriction, is a direct result of their choice of idealization; again, what rules count as available depends on how and to what degree the relevant agent is idealized. It's clear enough, then, that Greaves and Wallace's choice of idealization, not their choice of bridge principle, is what determines the particular verdicts delivered by their picture. (The connection between choice of restriction and choice of idealization, incidentally, also helps to explain why Schoenfield and Steel misdiagnose the source of their disagreement with Greaves and Wallace: their neglect of the fact that a restriction has been introduced at all leads them to regard the set of rules up for consideration as fixed, and this makes invisible to them the role choice of idealization plays in determining what rules Greaves and Wallace take to be legitimate.[19]) So, in so far as there's a problem with Greaves and Wallace's argument for conditionalization, it must in the end be a problem with their choice of idealization.

When this has been made explicit, it's relatively easy to see that Schoenfield's and Steel's worries about conditionalization can be understood as worries about idealization. The source of the purported problem, recall, is that an agent might gain evidence that indicates that, if she plans to conditionalize, she'll fail. But on Greaves and Wallace's choice of idealization, a rational agent can never gain such evidence, for reasons discussed in Section 2: their choice of idealization is one on which a rational agent is subject to no cognitive limitations and is certain of this fact, in which case, if a rule is consistent with evidentialism, there's no way for the agent to come to expect that the result of planning to conform to that rule will be anything but success.

Incidentally, in so far as the concern here can indeed be understood as a concern about Greaves and Wallace's choice of idealization, I'm inclined to share this concern. My view is that the kind of idealization relevant to epistemology is idealization of reasoning abilities, and I take it that, even if an agent is a perfect reasoner, she might be less than certain of her own perfection, for reasons explained by (for example) Christensen ([2007]).[20] And if that's right, even a perfect reasoner might gain evidence—misleading evidence—that she isn't cognitively perfect, in which case some rules consistent with evidentialism might turn out to be unavailable.[21]

---

[18] Thanks to an anonymous referee for pressing me to clarify my response here.

[19] Interestingly, Schoenfield does acknowledge the connection between availability and idealization in her ([2017], Section 4.2). But this isn't a paper in which she discusses the relative merits of Best-Plan-to-Follow and Best-Plan-to-Make. In the two papers where she does that—her ([2015]) and her ([2018])—the connection between availability and idealization is never drawn.

[20] Though see (Titelbaum [2015]) for an opposing view.

[21] That said, whether conditionalization is one of the rules that turns out to be unavailable on an appropriate choice of idealization is a further question. I'm inclined to think it's not one of those rules, for reasons I discuss elsewhere (again, see my [unpublished]).

Ultimately, though, my own opinions about what sort of idealization is suitable aren't relevant to the central point of this paper. That point is this: the disagreement between Greaves and Wallace, on the one hand, and Schoenfield and Steel, on the other, is not, despite appearances, a disagreement about what rationality–accuracy bridge principle is most suitable for the purposes of accuracy-first epistemology. It's just a disagreement about how to answer a much older question: the question of how idealized a notion of rationality we're interested in having.

Acknowledgements

I'd like to thank Lorenzo Rossi, Julien Murzi, Zach Barnett, David Christensen, Mary Renaud, Jonathan Courtney, Louis Gularte, William Peden, Matthew Hewson, Miriam Schoenfield, audiences at the 2019 Bayes by the Sea conference and the 2019 Conference for Philosophy of Science and Formal Methods in Philosophy of the Polish Association for Logic and Philosophy of Science, and several anonymous referees for helpful discussion of the ideas in this paper. I'm also grateful to the Austrian Science Fund (grant no. P29716-G24) for their generous financial support as I completed this research.

Department of Philosophy (KGW)
University of Salzburg
Salzburg, Austria
brett.topey@plus.ac.at

References

Bradley, D. [2020]: 'Self-Locating Belief and Updating on Learning', *Mind*, **129**, pp. 579–84.

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

Christensen, D. [2007]: 'Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals', in T. S. Gendler and J. Hawthorne (*eds*), *Oxford Studies in Epistemology*, Vol. 2, Oxford: Oxford University Press, pp. 3–31.

Greaves, H. and Wallace, D. [2006]: 'Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility', *Mind*, **115**, pp. 607–32.

Isbell, J. R. [1957]: 'Finitary Games', in M. Dresher, A. W. Tucker, and P. Wolfe (*eds*), *Contributions to the Theory of Games*, Vol. 3, Princeton: Princeton University Press, pp. 79–96.

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

Pettigrew, R. [2020]: 'What Is Conditionalization, and Why Should We Do It?', *Philosophical Studies*, **177**, pp. 3427–63.

Piccione, M. and Rubinstein, A. [1997]: 'On the Interpretation of Decision Problems with Imperfect Recall', *Games and Economic Behavior*, **20**, pp. 3–24.

Schoenfield, M. [2015]: 'Bridging Rationality and Accuracy', *Journal of Philosophy*, **112**, pp. 633–57.

Schoenfield, M. [2017]: 'Conditionalization Does Not (In General) Maximize Expected Accuracy', *Mind*, **126**, pp. 1155–87.

Schoenfield, M. [2018]: 'An Accuracy Based Approach to Higher Order Evidence', *Philosophy and Phenomenological Research*, **96**, pp. 690–715.

Steel, R. [2018]: 'Anticipating Failure and Avoiding It', *Philosophers' Imprint*, **18**, pp. 1–28.

Titelbaum, M. G. [2015]: 'Rationality's Fixed Point (Or: In Defense of Right Reason)', in T. S. Gendler and J. Hawthorne (*eds*), *Oxford Studies in Epistemology*, Vol. 5, Oxford: Oxford University Press, pp. 253–94.

Topey, B. [unpublished]: 'Higher-Order Evidence and the Dynamics of Self-Location: An Accuracy-Based Argument for Calibrationism', available at <philpapers.org/rec/TOPHEA>.