Avoiding Polarization — *Adrian Liu, August 30, 2024*

Abstract: Recent authors (Elga 2013, 2007, Dorst et al. 2021, Dorst 2020) have explored a tension between diachronic reflection principles and the possibility of ambiguous evidence: evidence that warrants uncertainty in what opinions are warranted. Dorst (2023) extends this tension by arguing that asymmetrically-occuring ambiguous evidence can generate expectable polarization. This paper resolves the tension between ambiguous evidence and diachronic reflection by defending an account of ambiguous evidence in which it does not generate diachronic reflection failures. Following Schoenfield (2018), Gallow (2021), and Isaacs and Russell (2023), I suggest that a piece of evidence is ambiguous when epistemic agents expect themselves to have *fallible* responses to it. On a natural way of interpreting reflection principles for fallible agents, fallible rational responses to evidence can still satisfy diachronic reflection failures and avoid polarizing. I conclude that the burden is on those who wish to argue that even infallible agents can face evidential ambiguity; in the meantime, the threat of rational polarization needn't worry us too much.

Contents

1 | Introduction

We say people *polarize* when their opinions diverge from each other. Beyond the political problems arising from profound disagreement, polarization is also epistemically worrying: if there is some way the facts are, then people diverging from each other indicates that some are also diverging from the facts. Somes polarization results from bad evidence: cults, conspiracy theories, propaganda. But the more epistemically worrying variety of polarization involves divergence from what is *warranted by one's evidence*: polarization that is not only unlucky, but possibly irrational.

As epistemic agents always encountering new evidence, how do we avoid diverging from what is warranted by our evidence? Here's one answer: suppose I will encounter different pieces of evidence $e_1, e_2, \ldots$ with different likelihoods $l_1, l_2, \ldots$ and I will adopt the opinions $P_1, P_2, \ldots$ warranted by the evidence (where evidence *e warrants* an opinion P iff it is rational to hold P in response to e). Then I should now adopt a weighted average of those opinions proportioned by their likelihoods: $l_1 \cdot P_1 + l_2 \cdot P_2 + \cdots$. If I have applied to a program with a 25% acceptance rate and I will learn the result tomorrow, then I now know that tomorrow I will learn that I have been accepted (probability 0.25) or learn that I have been rejected (probability 0.75). So I should now think there is a 0.25 chance that I will be accepted.

This is a *diachronic reflection*-style constraint, and it suffices for avoiding polarization when one's evidence is veridical. But diachronic reflection constraints are in tension with another plausible constraint of epistemic rationality: that sometimes evidence can warrant uncertainty what the evidence warrants. Call such evidence *ambiguous*. The basic tension is as follows: rational responses to ambiguous evidence seem to generate diachronic reflection failures – and so ambiguous evidence threatens to polarize you away from the very opinions it warrants.

This paper resolves the tension between *ambiguous evidence* and *diachronic reflection* by defending an account of *ambiguous evidence* in which it does not generate diachronic reflection failures. I argue that a piece of evidence is ambiguous when epistemic agents expect themselves to have *fallible* responses to it. On a natural way of interpreting reflection principles for fallible agents, fallible rational responses to evidence can still satisfy diachronic reflection failures and avoid polarizing.

In §2 I outline the concept of ambiguous evidence. In §3 present an argument from Dorst (2023) that ambiguous evidence that occurs asymmetrically can generate reflection failures, and thus expectable polarization. In §4 I show that ambiguous evidence does generate polarization in any particular direction nor necessitate it at all: ambiguous evidence turns out not to have so tight a connection to polarization. §5 discusses the problem more generally in connection to constraints on epistemic rationality. §6 argues that evidence is ambiguous when responses to the evidence are fallible,and shows that on this picture, rational responses to ambiguity are compatible with diachronic reflection. There are two upshots: First, the threat (or hope) of rational polarization should not worry us too much. Second, the more general threat (or hope) of ambiguous evidence may also be more quotidian than we feared.

2 | Ambiguous Evidence

Our interest is in *rational responses to new information*: or, *rationally responding to your evidence*. Evidence, in this usage, is a theoretical term, defined together with epistemic rationality: your evidence at a time t is whatever information, together with your opinions at time t, determines how it is rational for you to respond (see Dorst (2020, 586–7, fn1),). One's evidence is often called the "strongest" information one learns. The idea here is that there is nothing more you could learn that would change how it is rational for you to respond.[1]

Defining evidence as a theoretical term allow us to proceed with very few stipulations about the nature of evidence: all we need to assume is that it plays the role of rationalizing opinions. So we will not assume that evidence is propositional or partitional, and I will speak informally of "responding to evidence e" without specifying what sort of entity e is: what matters is that upon "responding to evidence e", whatever that means, an agent comes to have some opinions, and we can evaluate the rationality of those opinions.

Ambiguous evidence, as we'll use the term here, is not just evidence that leaves us rationally uncertain, but rather evidence that leaves us rationally uncertain about how to rationally respond to the evidence. Here is an example of unambiguous evidence: if you see a fair die (that you are certain is fair) rolled onto a flat surface, you can be uncertain how it landed but certain in what probabilities to assign each outcome. Here is an example of possibly ambiguous evidence: suppose you are given a string with letters and blanks (Cf. Elga and Rayo 2022), like P_A_ET, and asked whether or not there is some way to fill in the blanks to complete the string into a word. After some time, either you find a word or you do not. But you also gain information beyond whether or not you've found a word. Suppose you don't find a word: sometimes the string will be obviously uncompletable (like with DX__M), but sometimes you'll get a funny feeling that there is a word even though you haven't found it (perhaps with R_Y__M).

When a fair six-sided die is rolled, you can be confident your evidence licenses a $1/6$ credence in each outcome. But when you haven't found a word in the word search and you have some subtle feeling, it could warrant some particular credence without you being certain or confident that it warrants that credence. Ambiguous evidence, then, is evidence where it is rational to respond to it in some way but it is also rational to be uncertain about how to respond to it. Put another way, ambiguous evidence is evidence that warrants uncertainty in what is warranted.[2]

Definition 1. Ambiguous Evidence: say that some evidence e is *ambiguous* if there is some proposition q such that for all $s \in [0, 1]$, $\mathscr{R}_e(\mathscr{R}(q) = s) < 1$, where $\mathscr{R}_e$ is the rational credence to adopt in response to e.

1. When evidence is assumed to be propositional, "strongest" is sometimes defined as "entailing all other propositions you learn and not entailed by any other proposition you learn." See Schoenfield (2017, pp. 1156–7) for discussion.
2. Elsewhere, uncertainty about what is rational to think is called *modesty* or *higher-order uncertainty*. I find both these terms confusing, since *modest* intuitively points to an underestimation of one's abilities (which is not implied here), and *higher-order uncertainty* could just as well mean uncertainty about what one's actual opinions are, not uncertainty about what the rational opinions are.

Core debates between *internalists* and *externalists* about epistemic rationality, epistemic justification, and evidence can be translated into a debate about whether evidence can ever be ambiguous (externalists say yes, internalists say no):

> ◊Ambiguous: evidence can be ambiguous: sometimes, whatever (together with prior opinions) determines the rational response also warrants uncertainty in what the rational response is.
>
> □Unambiguous: evidence can never be ambiguous: always, whatever (together with prior opinions) determines the rational response also warrants certainty in what the rational response is.

Which position is true is important to determining what information could even count as evidence: if evidence cannot be ambiguous, then any information that warrants rational-uncertainty in how to respond to it cannot be evidence. It also affects the question of how to rationally respond to one's evidence.

3 | A Case Study

3.1 | *The Case*

Dorst (2023) argues that *when evidence is ambiguous, rational updates can lead to expectable polarization*. Sometimes, when evidence is ambiguous, responding rationally to the evidence can mean responding in a way that will lead to reflection failures: your average opinions in the future will be ones that you don't now think are rational. Here is his mechanism, building on the word-search case:

> **Coin-Flip Word Search:** A subject, Haley is given a series of trials, in each of which a fair coin is tossed at t_0 and Haley is shown a word-search at t_1. If the coin lands Heads, Haley is shown a string that can be completed into a word ("completable"). If the coin lands Tails, Haley is shown a string that cannot be completed into a word ("not completable"). Then Haley is asked at for her credence that the coin landed Heads.

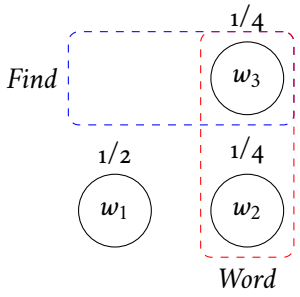
Let *Word* be the proposition that there is a word, *Find* be the proposition that Haley finds a word, and *Heads* be the proposition that the coin came up heads. Given the setup, *Word* = *Heads*: they define the same worlds. Now suppose Haley is certain about all the facts of the setup, and certain about the following further propositions:

1. She is 50% accurate at finding words in completable strings.
2. If she doesn't find a word, sometimes she will get a funny feeling that there is a word, even though she hasn't found it. Let *Feeling* be the proposition that Haley gets a funny feeling. In fact, she will get a funny feeling *if and only if* there is in fact a word. So *Feeling* = *Word* ∧ ¬*Find*.

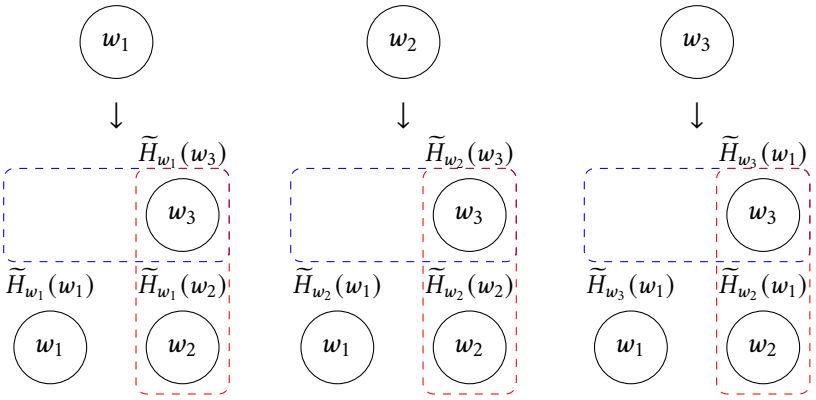
Let H be Haley's rational prior credence (prior to seeing the string). Given the setup as described, the prior credence H can be modeled as follows. We index the relevant epistemically possible worlds (our worlds will be coarse-grained, rather than maximally-specific descriptions): let w_1 be the world where there is no word (and so Haley does not find a word, nor get a funny feeling), w_2 be the world where there *is* a word but Haley does not find it, and let w_3 be the world where there is a word and

Haley *does* find it.[3] Then we can create a diagram by drawing a node for each w_i and arranging the nodes so that the top row Haley finds a word while in the bottom row she doesn't, and in the right row there is a word while in the left row she doesn't. We can further label each node with Haley's prior rational credence that she is in that world.

Prior Haley is certain that half the time there will be no word, so $H(w_1) = 1/2$, and half the time there is a word (which itself happens half the time), she will find it, so $H(w_2) = 1/2 \cdot 1/2 = 1/4$ and $H(w_3) = 1/2 \cdot 1/2 = 1/4$.



How should Haley respond to the evidence she has after she tries to complete a word? We seek an *update plan* for Haley that prescribes a set of opinions at t_1, when she sees the string, for each possibility. A complete rational update plan for Haley would involve a rational posterior credence function $\widetilde{H}_{w_i}$ at t_1 for each world w_i she might be in. We can think of this as a map from worlds to credence functions, as below:

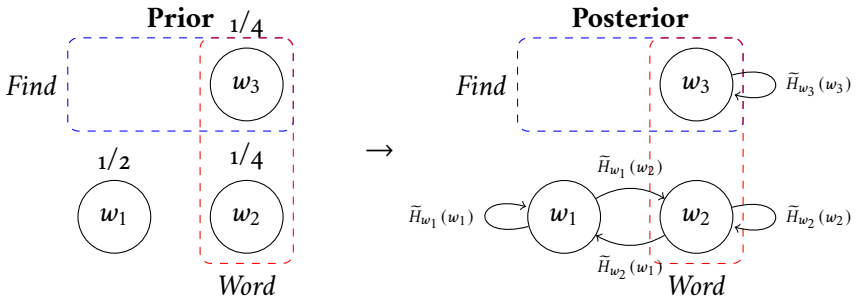


This information can be condensed: instead of writing out the entire diagram for each possible w_i, we can just have one copy of the diagram and draw an arrow $w_i \longrightarrow w_j$ with a label s if $H_{w_i}(w_j) = s$. This way of diagramming is *slightly* more confusing but *much* more space-efficient. I will adopt this way of drawing posteriors, and omit arrows whenever $H_{w_i}(w_j) = 0$. Using this diagramming, we can describe an *update* as a transition from a prior at t_0 (left diagram) to a *set of* posteriors at t_1 (right diagram), with the specific posterior depending on what evidence Haley gets:

3. More specifically: [1] w_1 is the world where at t_0 the coin comes up **tails** (¬*Heads*), and so Haley sees a string that is **not completeable** (¬*Word*), and so Haley **does not** get a funny feeling (¬*Feeling*) and **does not** find a word (¬*Find*).

[2] w_2 is the world where at t_0 the coin comes up **heads** (*Heads*), and so Haley sees a string that **is completeable** (*Word*), so Haley **does** get a funny feeling (¬*Feeling*) but **does not** find the word (¬*Find*).

[3] w_3 is the world where at t_0 the coin comes up **heads** (*Heads*), and so Haley sees a string that **is completeable** (*Word*), so Haley **does** get a funny feeling (¬*Feeling*) and **does** find the word (*Find*).

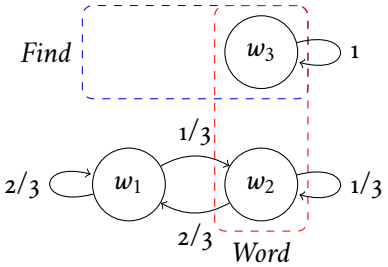


3.2 | Two unambiguous strategies and an ambiguous one

What update plan is most rational for Haley to adopt? This subsection considers three possibilities. The Bayesian conditionalizer says that Haley should conditionalize on the total body of information she becomes certain in. If the funny feeling is not something Haley is certain about, then all Haley is certain about after being shown the word is (1) that the coin is fair, (2) that she can find words 50% of the time when they are there, and (3) whether or not she found a word. So there is probability 1/4 that there is a word and Haley finds the word, probability 1/4 that there is a word and Haley does not find it, and probability 1/2 that there is no word to find. Adding the last two, there is probability 3/4 that Haley does not find a word, and so conditional on not finding a word, her credence should be 1/3. Call this **Bayesian Update.**

Bayesian Update: if Haley finds a word, she is certain there is a word. If she doesn't find a word, she conditionalizes:

$$\begin{aligned}\widetilde{H}(Word) &= H(Word \mid \neg Find)\\ &= \frac{H(Word\&\neg Find)}{H(\neg Find)} = \frac{1/4}{3/4} = \frac{1}{3}.\end{aligned}$$

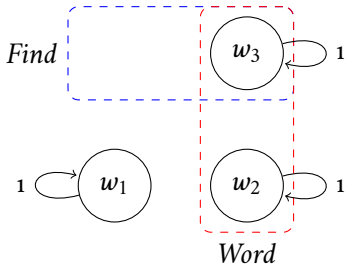


Bayesian Update ignores the funny feeling. What if Haley were to take the funny feeling into account? Assuming that Haley should not ignore the *Feeling*, the defender of □Unambiguous and the Bayesian agree on how to respond to it. The Bayesian says that, if there is any rational response to *Feeling* other than ignoring it, that response is to conditionalize on it. The defender of □Unambiguous thinks that if *Feeling* is evidence, then it is unambiguous, and thus the rational response is to conditionalize on it. Now, Haley knows (by stipulation of the case) that she gets the funny feeling at t_1 if and only if there is a word (that she doesn't find). So if Haley conditionalizes, she becomes certain there is a word if and only if there is a word, and becomes certain there is no word if and only if there is no word. This delivers us **Certain Update**.

Neither Bayesian Update nor Certain Update give due credit to the intuitions around *Feeling*: Bayesian Update doesn't respect that *Feeling* seems to tell us *something*, rather than nothing, about whether there is a word. Certain Update is overconfident in *Feeling*'s ability to tell Haley whether there is a word. Yet, if *Feeling* is unambiguous, these are basically the only two choices.

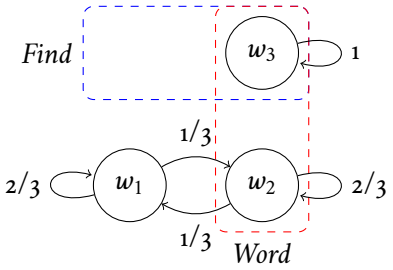
Dorst argues that *Feeling* is ambiguous evidence and warrants increasing cre-

Certain Update: if Haley finds a word, she is certain there is a word. If she doesn't, then if she gets a funny feeling, she is certain there is a word, and if she doesn't get a funny feeling, she is certain there is no word.



dence in *Word* somewhat, but not all the way up to certainty. He proposes an update wherein Haley increases her credence in *Word* to $2/3$ if she has funny feeling, and leaves it unchanged from the Bayesian Update otherwise (The $2/3$ number is not crucial: all we need is that it be $> 1/3$).

Ambiguous Update (Heads): if Haley finds a word, she is certain there is a word. If she doesn't find a word, then if she has the funny feeling, she has credence $2/3$ that there is a word; otherwise she has credence $1/3$ that there is a word.



3.3 | Asymmetric ambiguity allows expectable polarization

In Ambiguous Update (Heads), Haley *expectably polarizes*. Haley can predict that her rational credence in *Word* and thus *Heads* will, on average, be $7/12$, even though she is certain the coin is fair. That is, she predicts that she will think a fair coin will come up heads about 58% of the time!

Definition 2. Expected Credence: If H is a credence function over W and $\widetilde{H}$ is a function from W to credence functions, then H's expectation of $\widetilde{H}$, written $\mathbb{E}_H(\widetilde{H}(q))$, is given by the following equation:[4]

$$\mathbb{E}_H(\widetilde{H}(q)) := \sum_{x \in [0,1]} H(\widetilde{H}(q) = x) \cdot x \qquad \text{(Expected Credence)}$$

Then we can say that Haley's rational prior *expectation-reflects* her rational posterior on a proposition p if her prior rational credence $H(p)$ in p equals her prior rational credence's expectation of her posterior rational credence in p:

4. Note that

$$\sum_{x \in [0,1]} H(\widetilde{H}(q) = x) \cdot x = \sum_{x \in [0,1]} \sum_{w : \widetilde{H}_w(q) = x} H(w) \cdot \widetilde{H}_w(q) = \sum_{w \in W} H(w) \cdot \widetilde{H}_w(q). \qquad (1)$$

We will sometimes use the last form of the expectation to do calculations.

Definition 3. Expectation Reflection: Say that H expectation-reflects $\widetilde{H}$ on q if

$$H(q) = \mathbb{E}_H(\widetilde{H}(q)) \qquad \text{(Expectation Reflection)}$$

In Bayesian Update, Haley's prior expectation-reflects the posterior. We calculate by asking how likely Haley thinks it is that she ends up with posterior credence 1 in heads (1/4) and how likely Haley thinks it is that she ends up with posterior credence 1/3 in heads (3/4). Then we add up the posterior credences weighted by how likely Haley's prior thinks they are:

$$\begin{aligned} \mathbb{E}_H(\widetilde{H}(Word)) &= 1 \cdot H(\widetilde{H}(Word) = 1) + \tfrac{1}{3} \cdot H(\widetilde{H}(Word) = \tfrac{1}{3}) \\ &= (1 \cdot \tfrac{1}{4}) + (\tfrac{1}{3} \cdot \tfrac{3}{4}) = \tfrac{1}{2}. \end{aligned} \tag{2}$$

The same result can be checked, more trivially, for Certain Update. But in Ambiguous Update (Heads), Haley's prior does not expectation-match its posterior – her expected posterior is higher than her prior: $H(Heads) = 0.5$ but $\mathbb{E}_H\widetilde{H}(Heads) = 0.58$.[5] Dorst argues that this coin-flip word-search game corresponds to certain cases of political polarization. Take some question, like whether lax gun laws increase death rates. Now let completable strings correspond to articles or studies that provide evidence against your current opinion, and incompletable strings correspond to articles or studies that provide evidence for your current opinion. Suppose that we are better at finding problems in articles we disagree with than those we agree with.When we don't find a problem in an article, that evidence is ambiguous (just like when we don't find a word in a string): it could be that there is a problem and we haven't found it, but it could be that there are no relevant flaws. The asymmetry in when evidence is ambiguous means that we are more likely to discount articles we disagree with in virtue of finding problems in them, even if we start out thinking that articles on both sides are equally likely to have flaws.

4 | Asymmetric Ambiguity Does Not Generate Polarization

Does the Ambiguous Update demonstrate that political polarization can be rational? We've seen that the Ambiguous Update illustrates that if ambiguous evidence like *Feeling* is possible, then there exist updates that polarize. But for the general structure of the Word Search to be a good story of *rational* political polarization, we would need to establish that the ambiguity of Haley's evidence generates Haley's polarization – that the evidence *warrants* Haley's expectably polarizing updates, and thus that Haley is rational despite expectably polarizing.[6] In this section, I argue that the ambiguity of Haley's evidence does not generates Haley's polarization – that the evidence does

5.
$$\begin{aligned} \mathbb{E}_H(\widetilde{H}(Word)) =& 1 \cdot H(\widetilde{H}(Word) = 1) + \tfrac{1}{3} \cdot H(\widetilde{H}(Word) = \tfrac{1}{3}) + \tfrac{2}{3} \cdot H(\widetilde{H}(Word) = \tfrac{2}{3}) \\ =& (1 \cdot \tfrac{1}{4}) + (\tfrac{1}{3} \cdot \tfrac{1}{2}) + (\tfrac{2}{3} \cdot \tfrac{1}{4}) = \tfrac{7}{12} > \tfrac{1}{2}. \end{aligned}$$

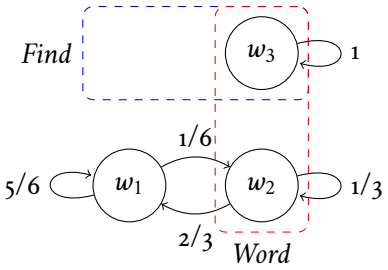
6. Dorst in fact argues for a stronger claim: that ambiguous evidence can generate *predictable* rational polarization. I do not discuss this part of his argumentation.

not warrant an expectably polarizing update.

4.1 | Asymmetric ambiguity does not asymmetrically generate polarization

Does the ambiguity of Haley's feeling push Haley's credences toward a bias in *Heads*? This is not clear. We could reason that when Haley has the funny feeling, she should increase her confidence in there being a word, and when she does not, she should keep her confidence steady. But what if many of the uncompletable strings are obviously uncompletable, like _YXH__? Then it seems Haley should keep her confidence steady at $\frac{1}{3}$ unless she doesn't get a funny feeling, in which case she should lower it. This makes Haley polarize in the *opposite direction* from Model 2:[7]

Ambiguous Update (Tails): if Haley finds a word, she is certain there is a word. If she doesn't find a word, then if she has the funny feeling, she hascredence 1/3 that there is a word; otherwise she has credence 1/6 that there is a word.



Ambiguous Update (Tails) polarizes in the opposite direction as Ambiguous Update (Heads), and to the same degree: $H(Heads) = \mathbb{E}_H\widetilde{H}(Heads) = 5/12$.[8] So even if asymmetric ambiguity generates polarization, it's doesn't generate polarization in any particular direction.

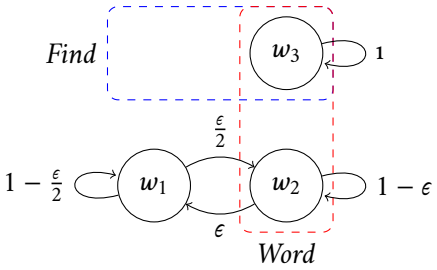
4.2 | Asymmetric ambiguity does not necessitate polarization

Even if asymmetric ambiguity did not asymmetric generate polarization, it would be an interesting enough result if Haley were forced to polarize in some direction or another. This would establish, in some sense, that it generates polarization (just not asymmetrically). But she isn't. For instance, suppose she adopts the credences from Ambiguous Update (Heads) when she has the feeling and the credences from Ambiguous Update (Tails) when she doesn't have the feeling. Then she doesn't polarize. $H(Heads) = \mathbb{E}_H(\widetilde{H}(Heads)) = 0.5$. In fact, if she adopts credence $1-\epsilon$ for some $\epsilon \in (0,1)$ when she has the feeling, and credence $\epsilon/2$ when she doesn't have the feeling, she will not polarize, and as long as $\epsilon \leq 2/3$, it is at least as accurate as

7. Thanks to Yong Xin Hui for suggesting this model.
8.
$$\begin{aligned}\mathbb{E}_H(\widetilde{H}(Word)) =& 1\cdot H(\widetilde{H}(Word)=1) + \tfrac{1}{6}\cdot H(\widetilde{H}(Word)=\tfrac{1}{6}) + \tfrac{1}{3}\cdot H(\widetilde{H}(Word)=\tfrac{1}{3})\\ =& (1\cdot\tfrac{1}{4}) + (\tfrac{1}{6}\cdot\tfrac{1}{2}) + (\tfrac{1}{3}\cdot\tfrac{1}{4}) = \tfrac{5}{12} < \tfrac{1}{2}.\end{aligned}$$

the Standard Bayesian update.[9]

Ambiguous Update (Reflecting): if Haley finds a word, she is certain there is a word. If she doesn't find a word, then if she has the feeling, she has credence $1-\epsilon$ that there is a word; otherwise she has credence $\epsilon/2$ that there is a word.



4.3 | *Asymmetric ambiguity allows almost-certain updates*

Dorst motivated the heads-polarizing update by saying it was more accurate than conditionalizing. But the tails-polarizing update is just as accurate, the nonpolarizing one is even more accurate. And in fact, if accuracy is our metric, we can go further. Recall that we assumed a word is wordlike if and only if there is a word, that Haley in fact distinguishes wordlike and non-wordlike strings with perfect reliability, and that Haley is certain that she does so. But if this is the case, shouldn't Haley adopt an update that not only does not expectably diverge, but also maximizes accuracy? In other words, why doesn't Haley adopt the Certain Update?[10]

Intuitively, it seems the Certain Update misses what we wanted to model out of the uncertainty. If the funny feeling is *ambiguous*, if Haley is never sure whether she has it, and thus whether it is rational to update on it, then it seems irrationally overconfident to adopt the Certain Model. But we need to say more than just that the Certain Update is not warranted because the evidence is ambiguous, if "ambiguous evidence" is defined as not warranting higher-order certain updates — we will have merely stipulated that *Feeling* is evidence that does not warrant the Certain Update, without explaining why, and to what extent, higher-order uncertainty is warranted. Furthermore, it is easy to obey the letter and not the spirit of ambiguous evidence: in Ambiguous Update (Nonpolarizing), we can let ϵ be arbitrarily small ($\epsilon \to 0$), and thus let Haley become arbitrarily close to certain. Call this **Almost-Certain Update**. Almost Certain Update is not very uncertain; nor does it expectably polarize. Clearly this doesn't capture the thought that ambiguous evidence warrants modesty to a degree commensurate with how ambiguous the evidence is, and the definitional approach does not provide such a link. Instead, we want a story that explains why *Feeling* does not warrant Certain Update in terms other than what is warranted or not by *Feeling*.

9.

$$\begin{aligned}\mathbb{E}_H(\widetilde{H}(Word)) =& 1 \cdot H(\widetilde{H}(Word)=1) + \tfrac{1}{6} \cdot H(\widetilde{H}(Word)=\tfrac{1}{6}) + \tfrac{2}{3} \cdot H(\widetilde{H}(Word)=\tfrac{2}{3}) \\ =& (1 \cdot \tfrac{1}{4}) + (\tfrac{1}{6} \cdot \tfrac{1}{2}) + (\tfrac{2}{3} \cdot \tfrac{1}{4}) = \tfrac{1}{2}. \\ \mathbb{E}_H(\widetilde{H}(Word)) =& 1 \cdot H(\widetilde{H}(Word)=1) + \tfrac{\epsilon}{2} \cdot H(\widetilde{H}(Word)=\tfrac{\epsilon}{2}) + (1-\epsilon) \cdot H(\widetilde{H}(Word)=1-\epsilon) \\ =& (1 \cdot \tfrac{1}{4}) + (\tfrac{\epsilon}{2} \cdot \tfrac{1}{2}) + ((1-\epsilon) \cdot \tfrac{1}{4}) = \tfrac{1}{4} + \tfrac{\epsilon}{4} + \tfrac{1-\epsilon}{4} = \tfrac{1}{2}.\end{aligned}$$

10. Thanks to Dmitri Gallow for raising this point.

5 | Stakes of ambiguity

The updates in the last section varied the update value for ambiguous evidence in different ways all compatible with the hypothesis that *Feeling* was ambiguous evidence. Ambiguous Update (Tails) polarized in the opposite direction as Ambiguous Update (Heads); Ambiguous Update (Reflecting) did not polarize at all, and Almost Certain Update didn't polarize, was basically certain, and basically maximized expected accuracy (by coming arbitrarily close to Certain Update). We were able to generate so many different updates because in the absence of constraints like diachronic reflection principles and □Unambiguous, we had no account of how the ambiguity of *Feeling* should constrain the rational response to it.

The issue is as follows: when evidence is unambiguous, multiple independently-plausible constraints on epistemic rationality are extensionally equivalent. But in the presence of ambiguous evidence, they come apart. I'll focus on four, with unfortunately brief discussions of each.

1. First is *Expected-accuracy Maximization*:

   Definition 4. $\widetilde{P}$ maximizes expected accuracy according to π and an accuracy scoring rule $\mathscr{A}$ if out of all available credence distributions $\widetilde{P} : w \longrightarrow \widetilde{P}_w$[11], $\widetilde{P}$ maximizes the quantity

   $$\sum_{W \in W} \sum_{w_i \in W} \pi(w) \cdot \widetilde{P}_w(w_i).$$

   An accuracy scoring rule is a function that takes as arguments a world $w \in W$ and a credence function defined over σW and returns an accuracy "score" (typically a real number). I will have to decide how much exposition to do about accuracy functions.
2. Second is *Conditionalization*:

   Definition 5. $\widetilde{P}$ is a conditionalizing update from π if there is some partition B over the worlds, and whenever in B_i, $\widetilde{P}_{B_i}(q) = \pi(q \mid B_i) := \frac{\pi(q \wedge B_i)}{\pi(B_i)}$.

   Conditionalization is not a desideratum of epistemic rationality so much as a method of updating that Bayesians defend as rationally required. There are good reasons to like conditionalization as an updating norm: when evidence is unambiguous, it not only maximizes accuracy, satisfies expectation-reflection, and satisfies total trust (below), but also accuracy-dominates other update rules (Briggs and Pettigrew 2020).
3. Third is *Total Trust* (Dorst et al. 2021), which formalizes and idea of *deference* to opinions as deferring decisions to those made on the basis of other opinions:

   Definition 6. $\widetilde{P}$ is totally trusted by π if for any random variable $X : W \longrightarrow \mathbb{R}$,

   $$\mathbb{E}_\pi(X \mid \{w \mid \mathbb{E}_{\widetilde{P}_w}(X) \geq t\}) \geq t. \qquad \text{(Total Trust)}$$

   (Dorst et al. 2021, 107).

11. (TODO FIX functions vs families/rules/distributions)

Intuitively, $\widetilde{P}$ is totally trusted by π if, whenever $\widetilde{P}$ has an estimate of some value (represented by the variable X that is at least at some level of confidence), π adopts that level of confidence in its own estimate. It is a weaker cousin of standard reflection principles, but does not imply *Expectation-reflection* (as we'll see below). Dorst argues for the rationality of the Ambiguous Update (Heads) by showing that it satisfies Total Trust and is more accurate than the Bayesian update at every world. Even if it is not a sufficient condition on epistemic rationality, it is plausibly a necessary one.

4. Finally, *Expectation-reflection*, which we have already discussed.

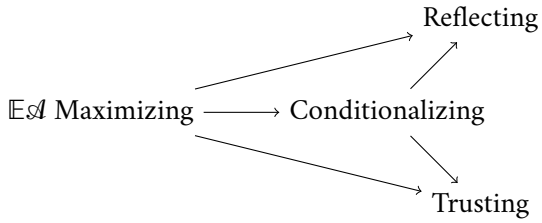
   Definition 7. $\widetilde{P}$ is expectation-reflected by π if $\pi(q) = \mathbb{E}_\pi(\widetilde{P}(q))$ (this is a restatement of (Expectation Reflection) above).

Now the equivalences and non-equivalences:

Theorem 1. *Equivalences given Unambiguous Evidence If $\pi \longrightarrow \widetilde{P}$ is an update such that the prior is π and the posterior satisfies $\mathcal{R}(q) = s \longrightarrow \mathcal{R}_e(\mathcal{R}(q) = s) = 1$ for any possible e, then the following are extensionally equivalent: (1) P maximizes accuracy according to π; (2) P is a conditionalizing update from π, derived from conditioning on e; (3) P is expectation-reflected by π; (4) P is totally trusted by π.*

The various equivalences are established by [authors cited in Dorst fn 62], Schoenfield (2017), and Dorst (2023): I discuss this in §a.1.[12] If we do not assume unambiguous evidence, some of the implications drop away, and leave us with the following interesting structure:

Theorem 2. *Equivalences given Ambiguous Evidence If $\pi \longrightarrow \widetilde{P}$ is an update, but possibly $\mathcal{R}(q) = s \wedge \mathcal{R}_e(\mathcal{R}(q) = s) < 1$, then the following implications hold, but no others do:*



Proof. It suffices to provide examples of (1) conditionalizing updates, reflected updates, and totally-trusted updates that does not maximize accuracy and (2) expectation-reflected updates that are not totally trusted and totally-trusted updates that are not expectation reflected. For (1): $\pi \longrightarrow \widetilde{P}$ can be a conditionalizing, reflected, and totally-trusted update without maximizing accuracy (witness: Standard Bayesian Update). For (2): π can expectation-reflect $\widetilde{P}$ without totally-trusting it (witness: Ambiguous Update (Reflecting) with $\epsilon > 2/3$), and can totally-trust $\widetilde{P}$ without expectation-reflecting it (witness: Ambiguous Update (Heads) and (Tails)). □

12. Note: I need to be more careful here about a distinction between maximizing accuracy, not maximizing accuracy, and getting arbitrarily close to maximizing accuracy (e.g. with the Ambiguous Updates taking $\epsilon \longrightarrow 0$). The sense in which an Almost-Certain update strategy maximizes accuracy needs to be given in an epsilon-delta way: it maximizes accuracy in the sense that for any other fixed update, we can choose an ϵ that is more accurate than that update.

So here, more carefully, is the tension that I think the previous section reveals. When evidence is unambiguous, accuracy maximization, reflection, conditionalization, and trust all align, so in some sense there is no need to choose between them. When evidence is ambiguous, the four come apart, and we risk being forced to choose. And when we ask what we should choose, the tension arises: on one hand, accuracy maximizing seems optimal, because it allows us to obtain the other three constraints. On the other hand, accuracy maximizing has two problems: First, it is technically incompatible with ambiguity, because it treats ambiguous evidence as unambiguous. Second, it doesn't respect the intuition behind ambiguity, which is that a higher amount of ambiguity warrants a higher degree of uncertainty about what opinions are warranted.

In the rest of this paper I will argue that we can vindicate the hope of going with accuracy maximization and obtaining the other constraints.[13] We will adopt an idea from Schoenfield that fallible agents cannot always follow the plans they plan to follow. Accounting for fallibility motivates adopting modified versions of accuracy-maximization, conditionalization, reflection, and trust. Using the modified versions, the diagram above holds once more.[14]

6 | Avoiding Polarization

6.1 | *Fallibility*

Here is a story fo why evidence like *Feeling* does not warrant a higher-order certain update: it is because epistemic agents have update processes that are *fallible* with respect to *Feeling*. This is the picture Schoenfield (2015b, 2015a) argues for, the picture Gallow (2021) adopts when he models ambiguous evidence as evidence on which an update process might *misfire*. It is the picture Isaacs and Russell (2023) adopt when they include "guesses" that have an indeterministic relationship to evidence. Schoenfield and Isaacs & Russell suggest, evocatively, that on this picture choosing update rules is like choosing rules to program into a robot (with hardware limitations, presumably). It underlies perceptual-discrimination arguments for antiluminosity. It also appears to be the general background picture Dorst adopts (2023, p. 4).

On this story, the reason *Feeling* does not warrant a higher-order certain update is that if Haley plans to adopt a certain posterior credence $\widetilde{H}(\textit{Feeling})$ when *Feeling* is true, she will in fact adopt that credence only some of the times when *Feeling* is true: other times, her update process will *misfire* and she will adopt some other credence[15] (in Gallow's terminology), or she will incorrectly guess that *Feeling* is false, and thus adopt her plan for $\neg$*Feeling* (in Isaacs and Russell's terminology).

Suppose Haley has some rational prior estimation of her fallibility. To formalize this, let $\widehat{H}$ be Haley's *actual* posterior credences, whatever they are (to recap: H is Haley's *rational* prior, $\widetilde{H}$ is Haley's *rational* posterior, and $\widehat{H}$ is Haley's *actual* pos-

13. Draft Note: This doesn't mean going accuracy-first; since in the end we get all the epistemic goods above. I make no attempt to explain the other epistemic values in terms of accuracy.

14. (Draft Note: conditionalization seems like an outlier, since it is instrumental to epistemic goods and not a good itself.)

15. or she might adopt the same credence but not through the intended process.

terior). Then Haley's rational prior estimation of her overall fallibility is given by $H(\widehat{H} = \widetilde{H})$, her prior estimation that her actual posterior will be rational, and her prior expectation of her posterior actual credence can be given by

$$\mathbb{E}_H(\widehat{H}) = \sum_{[\widehat{H}=\widetilde{H}]} (\widetilde{H} \cdot H(w)) + \sum_{[\widehat{H}\neq\widetilde{H}]} (\widehat{H} \cdot H(w)). \tag{3}$$

If Haley thinks her update processes are infallible, i.e. $H(\widehat{H} = \widetilde{H}) = 1$, then the second term drops out, so $\mathbb{E}_H(\widehat{H}) = \mathbb{E}_H(\widetilde{H})$. But if Haley thinks her update processes might *misfire* or she might *guess wrong* with respect to *Feeling*, i.e. if $H(\widehat{H}(\textit{Feeling}) = \widetilde{H}(\textit{Feeling})) < 1$, then $\mathbb{E}_H(\widehat{H}) \neq \mathbb{E}_H(\widetilde{H})$.

6.2 | Fallibilist updating

Suppose that the possibility of fallibility is the *only* source of rational modesty. Then whenever $H(\widehat{H} = \widetilde{H}) = 1$, $\widetilde{H}$ must be immodest, and the Bayesian Update is optimal (it satisfies Reflection, obeys Value, and maximizes expected accuracy, as discussed bove.) What is the optimal update when $H(\widehat{H} = \widetilde{H}) < 1$?

Here is a thought: if we can satisfy value and reflection and maximize accuracy, we should do that. If we cannot, then we can consider tradeoffs between the different constraints and battle out intuitions about epistemic rationality. Our strategy will be to maximize accuracy first (without being committed to accuracy-first) and then check if value and reflection are satisfied.

Schoenfield (2017) shows, in a generalization of a result from Greaves and Wallace (2007),[16] that if an agent with prior π updates deterministically, and there is a partition of worlds $[w_i]$ such that the agent adopts some update $U(P_i)$ whenever P_i obtains, then the update with maximum expected accuracy is the one that assigns to each P_i the credence function $\pi(\cdot \mid P_i)$.

Our fallible agents update indeterministically, so this theorem does not apply. But it provides a guide: Schoenfield (2018) proposes to distinguish between the expected accuracy of planning to update in a certain way and the expected accuracy of conforming to that plan. Suppose that when an agent plans to update by adopting some update plan U assigning a credence function $U(w)$ to each possibility, she ends up adopting credences $\widehat{U}(w)$ at each w, which (for different worlds) may or may not be the credences she had planned to adopt. Schoenfield gives the following example:

> Suppose that I am planning my vacation and I am considering two possibilities: spending my vacation camping in the woods or spending my vacation on the moon. Clearly, vacationing on the moon would be more exciting than camping. Nonetheless, the moon plan is worse than the camping plan. Why is this?

Schoenfield's explanation is that even if I can make this plan, I can't expect anything good to come of it (since I will almost certainly not be able to conform to the plan and actually get to the moon). Thus, "while I can expect that the result of conforming to the moon plan will be better than the result of conforming to the camping plan, I

16. TODO: import reference.

can also expect that the result of making the moon plan (again, assuming I can make such a plan) will be worse than the result of making the camping plan."

If fallibility is a good explanation of the ambiguity of *Feeling*, then the reasoning above extends to an argument against Certain Update: it doesn't make sense for Haley to plan to adopt credence 1 in *Word* whenever she has *Feeling*, because she cannot conform this plan with sufficient reliability to justify certainty: sometimes she will get it wrong on whether she has *Feeling*, and her credences should take this into account. When agents are fallible in this way, the expected accuracy of *planning* to update in a certain way can be calculated by taking the accuracy of $\widehat{U}(w)$ at each w, and taking a weighted sum based on the probability of w:

$$\mathbb{PLE}_\pi \mathscr{A} U := \mathbb{E}_\pi \mathscr{A} \widehat{U} = \sum_{w \in W} \pi(w) \mathscr{A}(\widehat{U}(w), w). \tag{4}$$

Since we are focused on how ambiguous evidence could be explained by fallibility, I will assume that agents may be fallible in their responses to evidence, but not in other actions they take. In particular, they don't make errors on actions like "adopt credence t in q" *simpliciter* – there won't be situations in which, when they plan to adopt credence t in q, sometimes they adopt credence t' in q or credence t in q' instead. Rather, their errors manifest in responding to evidence: so given an action like "adopt credence t in q in response to evidence e", they may be fallible in virtue of not being able to perfectly respond to evidence e (they may mistake it for some other evidence e'.

If we consider only responding-to-evidence fallibility, and not fallibility in the ability to perform an action like adopting a certain credence,[17] then if Haley has a doxastic plan that covers every situation, any credence function she adopts will be the correct credence function (according to the plan) for *some* possible world. We can then write a more specific version of equation (4). Letting $[w_i] := \{w_i \in W \mid \widehat{U} = \widehat{U}(w_i)\}$:

$$\begin{aligned} \mathbb{PLE}_\pi \mathscr{A} U &:= \mathbb{E}_\pi \mathscr{A} \widehat{U} \\ &= \sum_{w \in W} \pi(w) \sum_{[w_i]} \pi([w_i] \mid w) \mathscr{A}(\widehat{U}(w_i), w). \end{aligned} \tag{Plan-Expected Accuracy}$$

This new way of evaluating accuracy, together with Schoenfield's SD-CondMax, suggests a way forward: modify the conditionalizing rule in an analogous way to the way we modified the expected-accuracy formulation. For agents certain that they would update with $U(P_i)$ whenever P_i obtained, the accuracy-maximizing update was conditionalizing on P_i. For an agent who updates with a range of credences $\{\widehat{U}(P_j)\}$ with some probability $\{\epsilon_{ij}\}$, the "closest thing" to Schoenfield's conditionalizing rule is the rule that conditionalizes on each P_j, and sums the results, weighted by the probability ϵ_{ij} of updating with $\widehat{U}(P_j)$ when P_i obtains. In other words, update $U(P_i)$ takes the *expectation*, given P_i, of conditionalizing on whatever P_j corresponds

17. See Isaacs and Russell (2023), §4 for a discussion of fallibility in the ability to adopt a certain credence.

to the agent's actual update $\widehat{U}(P_j)$:

$$U(P_i) = \mathbb{E}_\pi(\pi(\cdot \mid U) \mid P_i) := \sum_j \pi(\cdot \mid P_j)\pi(U(P_j) \mid P_i). \qquad \text{(Fallibilist Updating)}$$

Fallibilist Updating maximizes Plan-Expected Accuracy. This can be shown in a few different ways: Gallow (2021, Appendix B) proves an equivalent result directly, while Isaacs and Russell (2023, Appendix C) prove another equivalent result as an extension of Greaves and Wallace. I prove the result in §a.2 using the notation of this paper. [18]

Suppose Haley has a *complete* estimate about about how fallible she is: not only is $H(\widehat{H} = \widetilde{H})$ well defined, but the conditional probabilities $\widehat{H} = \widetilde{H}_{w_j} \mid \widetilde{H} = \widetilde{H}_{w_i}$ for all w_i, w_j are well defined (and thus the reverse conditional probabilities are well defined as well). Suppose also that Haley is certain her actual posterior credences are always the rational credences for *some* world ($H(\exists w.(\widehat{H} = \widetilde{H}_w)) = 1$). Writing $\widetilde{w}_i = \{w \mid \widetilde{H}_w = \widetilde{H}_{w_i}\}$ and $\widehat{w}_i = \{w \mid \widehat{H}_w = \widetilde{H}_{w_i}\}$, the following update is an implementation of Fallibilist Updating.

$$\widetilde{H}_w(q) = \mathbb{E}_H(H(q \mid \widetilde{w}_i) \mid \widehat{w}) := \sum_{\widetilde{w}_i} H(q \mid \widetilde{w}_i)H(\widetilde{w}_i \mid \widehat{w}) \quad \text{(Fallibilist Updating V2)}$$

For example, suppose that Haley's estimations are as follows: (1) At w_1, where she doesn't get the feeling, she will adopt the posterior $\widetilde{H}_{w_1}(\neg \textit{Feeling})$ with probability $5/6$ and adopt the posterior $\widetilde{H}_{w_2}(\neg \textit{Feeling})$ with probability $1/6$. (2) At w_2, where she does get the feeling, she will adopt the posterior $\widetilde{H}_{w_2}(\textit{Feeling})$ with probability $2/3$ and adopt the posterior $\widetilde{H}_{w_1}(\textit{Feeling})$ with probability $1/3$. That is, her *true positive* rate is $2/3$ and her *true negative* rate is $5/6$.According to the formula, the rational credences are given by $\widetilde{H}_{w_2}(\textit{Feeling}) = 2/3$ and $\widetilde{H}_{w_1}(\textit{Feeling}) = 1/6$, as in the Non-polarizing Update. If her true positive rate and true negative rate are both $2/3$, then her rational credences are given by $\widetilde{H}_{w_2}(\textit{Feeling}) = 1/2$ and $\widetilde{H}_{w_1}(\textit{Feeling}) = 1/5$. In fact, any pair of priors and error values uniquely determines the update according to the equation above. And none of the updates polarize.

So what of Dorst's Ambiguous Update (Heads)? It can't be modeled by Fallibilist Updating: there are no values for Haley's error rates and her priors that produce it. The result follows from equation 5 below. For Ambiguous Update (Heads), we can show that if we hold fixed the prior and the posterior, we get that $H(\widehat{w}_2) = 0$, which implies that the only way to interpret Ambiguous Update (Heads) as deriving from Fallibilist Updating is one in which the result of adopting is that Haley actually

18. Notice that I have been characterizing the various equations in terms of certainty of fallibility, not actual fallibility. The results do not require that rational agents be *actually* infallible, just that they are certain they are infallible (and in fact, if an agent is not certain they are infallible, then they are certain they are fallible (an agent is not certain they are infallible if $\pi(\widehat{U} = U) < 1$; but then $\pi(\exists w.[\widehat{U}(w) \neq U(w)]) = 1$, so $\pi(\widehat{U} \neq U) = 1$). This is also a sensible constraint: even for externalists, rationality is a matter of what an agent should do "by their own lights". However, we can disconnect the error estimations from the agent's prior credences without damage to any of the results: we just have to be careful about the interpretation of these quantifies. I show some of these generalizations in §a.2, and use generalized forms (which are more elegant) in some proofs in footnotes. But for now, let us proceed by considering Haley's own estimations of her fallibility.

follows Bayesian Update.[19]

6.3 | Evaluating fallibilist updating

Fallibilist Updating is not conditionalization. But, as I argued, it is *sort of* conditionalization, and follows the logic of Schoenfield's conditionalization rules (see Schoenfield 2017, §5 and Isaacs and Russell 2023, Appendix C for more discussion of this connection). So with a suitably fallibilist notion of expected accuracy, we obtained a suitably fallibilist substitute for conditionalization. How about Expectation Reflection and Total Trust? Plan-Expectation, following the logic of Plan-Expected Accuracy, is rather straightforward:

$$\mathbb{PLE}_H(\widetilde{H}(q)) := \mathbb{E}_H(\widehat{H}(q)) = \sum_{\widehat{w}_j} H(\widehat{w}_j)\widetilde{H}_{w_j}(q) \qquad \text{(Plan-Expectation)}$$

This allows us to say that Haley's prior Plan-Expectation-Reflects the posterior if

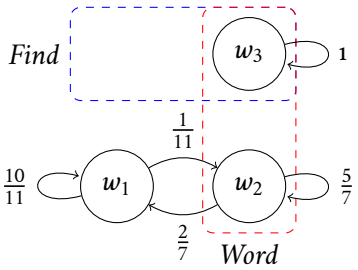
$$H(q) = \mathbb{PLE}_H(\widetilde{H}(q)) = \mathbb{E}_H(\widehat{H}(q)) \qquad \text{(Plan-Expectation Reflection)}$$

Then Fallibilist Updating satisfies Plan-Expectation Reflection:

$$\begin{aligned}\mathbb{E}_H(\widehat{H}(q)) &= \sum_{\widehat{w}_j} H(\widehat{w}_j)\widetilde{H}_{w_j}(q) = \sum_{\widehat{w}_j} H(\widehat{w}_j)\sum_{\widetilde{w}_i} H(q \mid \widetilde{w}_i)H(\widetilde{w}_i \mid \widehat{w}_j)\\ &= \sum_{\widetilde{w}_i} H(q \mid \widetilde{w}_i)\sum_{\widehat{w}_j} H(\widetilde{w}_i \mid \widehat{w}_j)H(\widehat{w}_j) = \sum_{\widetilde{w}_i} H(q \mid \widetilde{w}_i)H(\widetilde{w}_i) = H(q).\end{aligned} \tag{5}$$

But it doesn't satisfy Expectation Reflection. For consider **Ambiguous Update (**$\mathbb{PLE}$**-Reflecting but not** $\mathbb{E}$**-Reflecting)**, which is the result of Fallibilist Updating with $H(\widehat{w}_2 \mid \widetilde{w}_2) = H(\widehat{w}_1 \mid \widetilde{w}_1) = 5/6$.[20]

Ambiguous Update ($\mathbb{PLE}$**-Reflecting but not** $\mathbb{E}$**-Reflecting):** if Haley finds a word, she is certain there is a word. If she doesn't find a word, then if she has the feeling, she has credence $5/7$ that there is a word; otherwise she has credence $1/11$ that there is a word.



Note that this is not an instance of Ambiguous Update (Reflecting): no value of ϵ delivers these credences. In fact, every instance of Ambiguous Update (Reflecting)

19. By Fallibilist Updating, $2/3 = \widetilde{H}_{w_2}(\textit{Feeling}) = H(\widetilde{w}_2 \mid \widehat{w}_2)$ and $1/3 = \widetilde{H}_{w_1}(\textit{Feeling}) = H(\widetilde{w}_2 \mid \widehat{w}_1)$. So we have
$$1/4 = H(\widetilde{w}_2) = \sum_{\widehat{w}_i} H(\widetilde{w}_2 \mid \widehat{w}_1)H(\widehat{w}_i) = \tfrac{1}{3}H(\widehat{w}_1) + \tfrac{2}{3}H(\widehat{w}_2).$$
On the other hand, $H(\widehat{w}_1) + H(\widehat{w}_2) = 3/4$, which gives us the system of equations $3/4 = H(\widehat{w}_1) + 2H(\widehat{w}_2)$ and $3/4 = H(\widehat{w}_1) + H(\widehat{w}_2)$, which has a solution only at $H(\widehat{w}_2) = 0, H(\widehat{w}_1) = 3/4$.

20. TODO: include calculations for this proof.

(i.e., for each value of ϵ) satisfies both Expectation-Reflection and Plan-Expectation Reflection.[21] Why do they line up? A sufficient condition on Expectation-Reflection being satisfied iff Plan-Expectation Reflection is satisfied is that $H(\widetilde{w}) = H(\widehat{w})$ for all w.[22] Say that Haley **Updates Proportionally** if this is the case. Note that Haley can update proportionally even if it isn't the case that $\widetilde{H} = \widetilde{H}_w \leftrightarrow \widehat{H} = \widetilde{H}_w$ (which would imply infallibility): she can have false negatives and false positives, as long as they average out.[23]

When Expectation Reflection and Plan-Expectation Reflection come apart, which one is a better guide to polarization? If we are interested in when actual agents polarize, then we should consider how their opinions will actually evolve. So if we are interested in agents with uncertainty about what their evidence warrants, and if this uncertainty derives from their epistemic fallibility, then we should follow Plan-Expectation. Another way to put it: *any* opinions Φ about how likely Haley is to be in any particular world and how likely Haley is to adopt a certain opinion in that world, given facts about her epistemic fallibility, will give an estimate of Haley's credences that takes her fallibility into account: it will use Plan-Expectation. In particular, an probabilistically omniscient oracle $\nabla\Phi$ (who knows objective probabilities but not the future) will evidently give estimates for Haley's future credences that line up with Plan-Expectation. So if we return to Haley's perspective, it seems she should use Plan-Expectation Reflection and not Expectation Reflection to avoid polarization.

What about Total Trust? The idea of Total Trust was that a prior H totally trusts a posterior distribution $\widetilde{H}$ when they raise their expectation of any random variable X, conditional on $\mathbb{E}_{\widetilde{H}}(X) \geq t$, to at least t. What if Haley knows she is fallible, and thus that her posterior distribution will involve some amount of error? If Total Trust-style is a plausible constraint on epistemic rationality, it seems that Haley does better to plan to adopt posteriors that she totally trusts *given how she expects the plan will turn out*, rather than to plan to adopt posteriors that she totally trusts *if they turn out well*. She should adopt posteriors that she "Plan-Trusts."

Definition 8. $\widetilde{H}$ is totally plan-trusted by π if for any function $X : W \to \mathbb{R}$,

$$\mathbb{E}_H(X \mid \{w \mid \mathbb{PLE}_H(\mathbb{E}_{\widetilde{H}_w}(X)) \geq t\}) \geq t, \qquad \text{(Total Plan-Trust)}$$

where by Plan-Expectation, $\mathbb{PLE}_H(\mathbb{E}_{\widetilde{H}_w}(X)) := \mathbb{E}_H(\mathbb{E}_{\widehat{H}_w}(X))$.

Again, it seems like a constraint on rationality is that you satisfy total trust toward your expected actual posterior, not the one you would have if you were infallible. To take Schoenfield's practical example again, Haley should not expect good outcomes to come from planning to do the moon-vacation, even if the expected value of conforming to the moon-vacation plan is very high. So she should not Plan-Trust the moon plan, and thus she should not adopt it.

Does Fallibilist Updating satisfy Total Plan-Trust? On the limited case where the random variables are indicator variables of propositions, e.g. $\mathbb{1}_q$ (which equals 1 if

21. Can we find an example of the other way around – expectation reflection but not plan-expectation reflection?

22. Proof: If $H(\widetilde{w}) = H(\widehat{w})$ for all w, then $\mathbb{E}_H(\widetilde{H}(q)) = \sum_{\widetilde{w}_j} H(\widetilde{w}_j)\widetilde{H}_{w_j}(q) = \sum_{\widehat{w}_j} H(\widehat{w}_j)\widetilde{H}_{w_j}(q)$. This is the RHS of the first equality of (5), so we can follow the proof from there.

23. TODO: Is it a necessary condition? I don't know but should check.

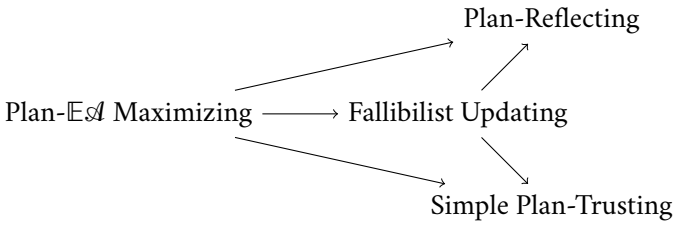
q is true and 0 otherwise: see Dorst 2020, Dorst et al. 2021, p. 107), Total Plan-Trust reduces to

$$\begin{aligned} & \mathbb{E}_H(\mathbb{1}_q \mid \{w \mid \mathbb{E}_H(\mathbb{E}_{\widehat{H}_w}(\mathbb{1}_q) \geq t\})) \geq t. \\ \Longrightarrow\ & H(q \mid \{w \mid \mathbb{E}_H(\widehat{H}_w(q) \geq t\})) \geq t. \qquad \text{(Simple Plan-Trust)} \\ :\Longrightarrow\ & H(q \mid \mathbb{E}_H\widehat{H}(q) \geq t) \geq t. \end{aligned}$$

But since Fallibilist Updating satisfies Plan-Expectation Reflection, we have $\mathbb{E}_H\widehat{H}(q) = H(q)$, Total Plan-Trust reduces to $H(q \mid H(q) \geq t) \geq t$, which is satisfied by any credence function. Note: I haven't had the time to figure out how to establish (or disprove) the case for random variables in general. But I suspect Total-Plan Trust holds.

The results in the previous subsection establish the following result, analogous to Theorem (1):

Theorem 3. *More Equivalences given Ambiguous Evidence If $\pi \rightarrow \widetilde{P}$ is an update, but possibly $\mathcal{R}(q) = s \wedge \mathcal{R}_e(\mathcal{R}(q) = s) < 1$, then the following implications hold.*



Under the assumption that fallibility is the source of evidential ambiguity, then, we were able to make use of a way of evaluating the expected results of making plans to recover an accuracy-maximizing update plan in the spirit of conditionalization that satisfied plan-theoretic versions of Expectation Reflection and Total Trust. Fallibilist Updating gives a satisfying answer to the question of how our levels of uncertainty should correspond with the ambiguity of our evidence: if ambiguity is a function of fallibility, then the more likely we think we are to make mistakes in response to our evidence, the more uncertain we should be in our updates. Although this does not establish that fallibility is the *only* source of ambiguity, it suggests a straightforward way to deal with ambiguity when it does arise from fallibility.

7 | Conclusion

I've argued that the fallibility of epistemic agents can account for ambiguous evidence, and can do so in a way that does not generate polarization. To the extent that ambiguous evidence is a matter of fallible responses, Fallibilist Updating is a good candidate for the rational update rule and actual-expectation-reflection is a good candidate for measuring polarization. And if Fallibilist Updating is the rational update rule, then rational agents will never polarize according to actual-expectation-reflection. So to the extent that ambiguous evidence is a matter of fallible responses, we have good reason to think that rational agents never polarize.

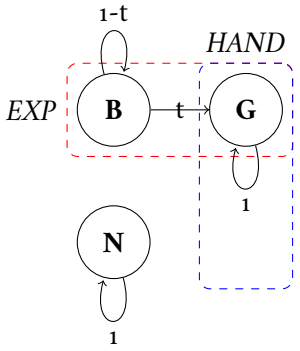
Let me end with a way to resist the above conclusions. I've argued that the fallibility of epistemic agents can account for ambiguous evidence and drawn out implications of the fallibility interpretation. But I haven't argued that fallibility is the *only* way to account for ambiguous evidence. Perhaps in some cases ambiguity has other explanations that work in tandem with fallibility to explain ambiguous evidence. Or perhaps sometimes the ambiguity of evidence is solely an *exogenous* factor: it's a feature of the evidence itself that it presents itself ambiguously to us and does not warrant rational certainty, *even for infallible agents.* Zendejas Medina (2024) suggests this possibility when he defends a rule for updating on ambiguous evidence which he says "is about *ideal* rationality, which is only sensitive to the agent's lack of information, and not to the possibility that they are cognitively limited." So Zendejas Medina argues that the rational update rule is conditionalization. He stipulates that ideally rational agents adopt the rational update rule and are *plan-coherent*: they carry out plans made at earlier times. This is equivalent to infallibility.

I am not sure if fallibility is the *only* source of ambiguity, but I think the burden is on those who think there could be another source. Consider the standard xternalist story of the ideal agent without cognitive limitations who faces a skeptical scenario. Here is his case:

> *Here's a Hand:* You're about to open your eyes, and know that you will either have an experience as of a hand (*EXP*) or not. You know that, if you don't have the hand-experience, then you don't have a hand. However, if you do have the hand-experience, you may in fact have a hand (*HAND*), or it could be that you're being deceived by a malicious demon, and there is no hand (¬*HAND*) (Zendejas Medina 2024, p. 4, following Williamson 2002, Chapter 8).

If, as externalists argue, one's possible evidence in this case is $\{EXP, HAND, \neg EXP \wedge \neg HAND\}$ and the ideally rational plan is to conditionalize on the evidence one has in the world one ends up in, then the rational posterior is given by the diagram below, where $t = \widetilde{H}_{\mathbf{B}}(\mathbf{G}) = H(HAND \mid EXP) < 1$ is the prior probability of being in the good case.

Bad Case: *EXP* ∧ ¬*HAND*.
→ Evidence: *EXP*.
Good Case: *EXP* ∧ *HAND*.
→Evidence: *EXP* ∧ *HAND*.
No Experience: ¬*EXP* ∧ ¬*HAND*.
→ Evidence: ¬*EXP* ∧ ¬*HAND*.



If the ideally rational plan is to conditionalize on *EXP*, then the ideally rational agent in the bad case conditionalizes on *EXP*. But let's put Haley in the skeptical scenario and now suppose she is certain she is ideally rational, and thus infallible in responding to her evidence. Then, because agents become certain about what the conditionalize

on $\pi(q \mid q) = 1$, Haley becomes certain in *EXP*. She does not become certain in *HAND*, since $\widetilde{H}_{\mathbf{B}}(HAND) = H(HAND \mid EXP) = t < 1$. But if Haley is certain she is infallible, then she is certain that if she were in the good case and her evidence were $EXP \wedge HAND$, she would conditionalize on it and thus be certain in *HAND*. Because she is not certain in *HAND*, she can infer that her evidence is not $EXP \wedge HAND$. By similar reasoning, she knows her evidence is not $\neg EXP \wedge \neg HAND$. By process of elimination, she can be certain her evidence is *EXP*. But then she can be certain that she is in the bad case: $\widetilde{H}_{\mathbf{B}}(\mathbf{B}) = 1$.[24] And this is not conditionalization on what was defined as the possible evidence above, but rather the accuracy-maximizing conditionalization plan from Schoenfield above. It seems that in order to reject this conclusion, Zendejas Medina would have to reject the infallibility assumptions. But this would be to reject plan-coherence, which is needed in order for his argument to apply to *ideally* rational agents, as he conceives of them.

In fact, complete skeptical cases seem to be limit cases of fallibility: cases of *complete* unreliability. If antiluminosity arguments based on perceptual discrimination are fallibilty arguments, skeptical scenarios, involving complete inability to perceptually discrimnate, are limit cases.[25]

I am sympathetic to the intuition that evidence could be *exogenously* ambiguous: warranting uncertainty in what it warrants even for infallible epistemic agents. But given the story of ambiguity as falliblity that I have defended, more work is needed to defend the existence of ambiguous evidence for infallible epistemic agents.

24. This requires that when agents are certain of propositions, they are certain they are certain; and thus when they are uncertain, they are certain they are uncertain. Any agent who satisfies Total Trust satisfies this (Dorst 2020, p. 594). But it does not require agents to have access to their credences in general.

25. (Greco (2021, 2019) argues through fragmentation. But it is in conjunction with fallibility. Indeed, fragmentation without fallibility only delivers ambiguous evidence when fragments cannot access each other. Logical information case?)

A | Proofs

A.1 | *Proof of Equivalence Theorem*

To come:

1. Conditionalization maximizes expected accuracy: Greaves and Wallace, Schoenfield
2. Expectation-Reflection, Total Trust, and Conditionalization are equivalent: various authors, but a complete proof in Dorst 2023.

A.2 | *Accuracy Proofs*

This is the statement of Schoenfield's most generalized result:

Theorem 4. *Super-Duper (SD) Generalized CondMax. Consider any partition of propositions P_i over a set of states Ω. Let U be a function from P_i to credence functions with the intended interpretation that an agent adopts $U(P_i)$ whenever P_i obtains. The U that maximizes expected accuracy [of $U(P_i)$] is the one that assigns to each P_i the credence function that results from conditionalizing on P_i .*

The generalization of this result from from §6.2 can be stated as follows:

Theorem 5. *Indeterministic Super-Duper Generalized CondMax (ISD CondMax). Consider any partition of propositions P_i over a set of states Ω. Let U be a function from P_i to credence functions and ϵ be a function from P_i to probability distributions over P_i with the intended interpretation that an agent with prior π adopts $U(P_j)$ with probability $\pi(U(P_j) \mid P_j)$ whenever P_i obtains. The U that maximizes the expected accuracy of $\mathbb{E}_{\epsilon(P_i)}U$is the one that assigns to each P_i the credence function that results from conditionalizing on P_j for each P_j and weighting that credence function by $\pi(U(P_j) \mid P_j)$.*

We can also detach the error estimations from the prior, and further generalize Fallibilist Updating to

$$\widetilde{P}_{s_j}(s) = \frac{\pi(s)\epsilon_s(s_j)}{\sum_{s_k} \pi(s_k)\epsilon_{s_k}(s_j)}, \qquad \text{(Generalized Fallibilist Updating)}$$

where $\epsilon_{s_i}(s_j)$ is interpreted as an agent's chances of adopting $\widetilde{P}_{s_j}$ when they are in state s_i. These chances need not be part of the agent's prior, as it was with Haley. The equation is otherwise identical. This generalization allows a further generalization of ISD CondMax:

Theorem 6. *Independent-Error Indeterministic Super-Duper Generalized CondMax (IE-ISD CondMax). Consider any partition of propositions P_i over a set of states Ω. Let U be a function from P_i to credence functions and ϵ be a function from P_i to probability distributions over P_i with the intended interpretation that an agent adopts $U(P_j)$ with probability $\epsilon(P_i)(P_j)$ whenever P_i obtains. The U that maximizes the expected accuracy of $\mathbb{E}_{\epsilon(P_i)}U$is the one that assigns to each P_i the credence function that results from conditionalizing on P_j for each P_j and weighting that credence function by $\epsilon(P_i)(P_j)$.*

Notational Variants are not reconciled right now, but IE-ISD CondMax should be rewritten as below:

Theorem 7. *Generalized Plan-Accuracy Maximization. Given any strictly proper, additive, and extensional measure of accuracy $\mathscr{A}$, the update rule that maximizes Plan-Expected Accuracy is the one that assigns each state s the credence function $\widetilde{P}_s(\cdot) = \sum_{s_i \in S} \pi(\cdot \mid s_i)\epsilon s(s_i)$.*

Proof. This proof basically follows the proof in Gallow (2021, Appendix B). As I noted above, Isaacs and Russell (2023, Appendix C) prove another equivalent result as an extension of

Greaves and Wallace. I think that other result is more conceptually edifying but is too complicated (they have too much notation). I am trying to figure out a way to simplify it using my notation. In the mean time, below is a rough outline of a straightforward proof.

Since $\mathscr{A}$ is additive, we can write $\mathscr{A}(\widetilde{P}_{s_i}, s)$ as $\sum_q \lambda_q \cdot \mathscr{A}_q(\widetilde{P}_{s_i}(q), s)$ for some set of accuracy functions $\mathscr{A}_q(Cr(q), q, w)$ that evaluate accuracy of credence functions *on specific propositions at specific worlds*. So the RHS of Plan-Expected Accuracy can be rewritten

$$\sum_{s\in S} \pi(s) \cdot \sum_{s_i \in S} \epsilon_s(s_i) \cdot \sum_q \lambda_q \cdot \mathscr{A}_q(\widetilde{P}_{s_i}(q), s). \tag{6}$$

Now, we want to find $\widetilde{P}_{s_i}(q)$ that maximizes (6), and $\widetilde{P}_{s_i}(q)$ is a function of q and s_i (we keep the prior π fixed). Now, for a given q and s_i, (6) is maximized when (7) is maximized:

$$\sum_{s\in S} \pi(s) \cdot \epsilon_s(s_i) \cdot \mathscr{A}_q(\widetilde{P}_{s_i}(q), s). \tag{7}$$

Since $\mathscr{A}$ is extensional, meaning that the value of $\mathscr{A}_q(Cr(q), q, w)$ only depends upon $Cr(q)$ and the truth value of q at w, then there exist functions $\mathscr{A}_1, \mathscr{A}_0 : \{Cr\} \to \mathbb{R}$ that satisfy

$$\mathscr{A}_q(Cr(q), q, w) \begin{cases} \mathscr{A}^1_q(Cr(q)) & \text{if } w \in q \\ \mathscr{A}^0_q(Cr(q)) & \text{if } w \notin q. \end{cases} \tag{8}$$

Then the choice of $\widetilde{P}_{s_i}(q)$ that maximizes (7) is the choice that maximizes

$$\sum_{s\in q} \pi(s \wedge q) \cdot \epsilon_s(s_i) \cdot \mathscr{A}^1_q(\widetilde{P}_{s_i}(q)) + \sum_{s\notin q} \pi(s \wedge \neg q) \cdot \epsilon_s(s_i) \cdot \mathscr{A}^0_q(\widetilde{P}_{s_i}(q)) \tag{9}$$

The value of (9) is still maximized if we divide it by a positive constant

$$\mathscr{A}^1_q(\widetilde{P}_{s_i}(q)) \cdot \sum_{s\in q} \frac{\pi(s \wedge q) \cdot \epsilon_s(s_i)}{\sum_s \pi(s) \cdot \epsilon_s(s_i)} + \mathscr{A}^0_q(\widetilde{P}_{s_i}(q)) \cdot \sum_{s\notin q} \frac{\pi(s \wedge \neg q) \cdot \epsilon_s(s_i)}{\sum_s \pi(s) \cdot \epsilon_s(s_i)} \tag{10}$$

Since $\cdot \sum_{s\in q} \frac{\pi(s\wedge q)\cdot\epsilon_s(s_i)}{\sum_s \pi(s)\cdot\epsilon_s(s_i)}$ is a probability function, we can rewrite the above as

$$\mathscr{A}^1_q(\widetilde{P}_{s_i}(q)) \cdot x + \mathscr{A}^0_q(\widetilde{P}_{s_i}(q))(1 - x) \tag{11}$$

with $x = \cdot \sum_{s\in q} \frac{\pi(s\wedge q)\cdot\epsilon_s(s_i)}{\sum_s \pi(s)\cdot\epsilon_s(s_i)}$. Since $\mathscr{A}$ is strictly proper, this value of x maximizes the value of $\mathscr{A}$. □

Theorem 8. *Generalized Plan-Expectation Reflection. An agent with error ϵ and prior π who adopts $\widetilde{P}_{s_j}(s) = \frac{\pi(s)\epsilon_s(s_j)}{\sum_{s_k} \pi(s_k)\epsilon)s_k(s_j)}$ satisfies*

$$\pi(q) = \mathbb{E}_\pi \left[\mathbb{E}_{\epsilon_{[s_i]}} \left[\widetilde{P}(q) \right] \right] := \sum_{s_i \in S} \pi(s_i) \cdot \sum_{s_j \in S} \epsilon_{s_i}(s_j) \cdot \widetilde{P}_{s_j}(q). \tag{12}$$

Proof. To simplify notation, let $\epsilon\pi(s_1, s_2) := \pi(s_1)\epsilon_{s_1}(s_2)$. Intuitively, $\epsilon\pi(s_1, s_2)$ is the probability that the agent is in s_1 and performs the response for s_2. Assuming a finite set of worlds, it suffices to show the result on propositions that are single worlds. We want to show that,

given $\widetilde{P}_{s_j}(s) = \frac{\pi(s)\epsilon_s(s_j)}{\sum_{s_k}\pi(s_k)\epsilon)s_k(s_j)}$,

$$\pi(s) = \sum_{s_i,s_j\in S}\pi(s_i)\cdot\epsilon s_i(s_j)\cdot\widetilde{P}_{s_j}(s) \coloneqq \sum_{s_i,s_j\in S}\epsilon\pi(s_i,s_j)\cdot\widetilde{P}_{s_j}(s) = \sum_{s_i,s_j\in S}\epsilon\pi(s_i,s_j)\cdot\frac{\epsilon\pi(s,s_j)}{\sum_{s_k}\epsilon\pi(s_k,s_j)}. \tag{13}$$

Now, $\pi(s) = \sum_{s_j}\pi(s)\epsilon_s(s_j) \coloneqq \sum_{s_j}\epsilon\pi(s,s_j)$, so we have

$$\sum_{s_i,s_j\in S}\epsilon\pi(s_i,s_j)\cdot\frac{\epsilon\pi(s,s_j)}{\sum_{s_k}\epsilon\pi(s_k,s_j)} = \frac{\sum_{s_j}\epsilon\pi(s,s_j)\sum_{s_i}\epsilon\pi(s_i,s_j)}{\sum_{s_k}\epsilon\pi(s_k,s_j)} = \sum_{s_j}\epsilon\pi(s,s_j) =: \pi(s). \quad \square \tag{14}$$

□

References

Briggs, R.A. and Richard Pettigrew (Mar. 2020). "An Accuracy-Dominance Argument for Conditionalization". In: *Noûs* 54.1, pp. 162–181. ISSN: 0029-4624, 1468-0068. DOI: 10.1111/nous.12258.

Dorst, Kevin (May 2020). "Evidence: A Guide for the Uncertain". In: 100.3, pp. 586–632. ISSN: 0031-8205, 1933-1592. DOI: 10.1111/phpr.12561.

— (July 2023). "Rational Polarization". In: *Philosophical Review* 132.3, pp. 355–458. ISSN: 0031-8108, 1558-1470. DOI: 10.1215/00318108-10469499.

Dorst, Kevin et al. (Dec. 2021). "Deference Done Better". In: *Philosophical Perspectives* 35.1, pp. 99–150. ISSN: 1520-8583, 1520-8583. DOI: 10.1111/phpe.12156.

Elga, Adam (Sept. 2007). "Reflection and Disagreement". In: *Noûs* 41.3, pp. 478–502. ISSN: 0029-4624, 1468-0068. DOI: 10.1111/j.1468-0068.2007.00656.x.

— (May 2013). "The Puzzle of the Unmarked Clock and the New Rational Reflection Principle". In: *Philosophical Studies* 164.1, pp. 127–139. ISSN: 0031-8116, 1573-0883. DOI: 10.1007/s11098-013-0091-0.

Elga, Adam and Agustín Rayo (Sept. 2022). "Fragmentation and Logical Omniscience". In: *Noûs* 56.3, pp. 716–741. ISSN: 0029-4624, 1468-0068. DOI: 10.1111/nous.12381.

Gallow, J. Dmitri (Sept. 2021). "Updating for Externalists". In: *Noûs* 55.3, pp. 487–516. ISSN: 0029-4624, 1468-0068. DOI: 10.1111/nous.12307.

Greco, Daniel (Oct. 2019). "Fragmentation and Higher-Order Evidence". In: *Higher-Order Evidence*. Ed. by Mattias Skipper and Asbjørn Steglich-Petersen. 1st ed. Oxford University PressOxford, pp. 84–104. ISBN: 978-0-19-882977-5 978-0-19-186827-6. DOI: 10.1093/oso/9780198829775.003.0004.

— (July 2021). "Fragmentation and Coarse-Grained Content". In: *The Fragmented Mind*. Ed. by Cristina Borgoni, Dirk Kindermann, and Andrea Onofri. Oxford University Press, pp. 54–77. ISBN: 978-0-19-885067-0 978-0-19-188562-4. DOI: 10.1093/oso/9780198850670.003.0003.

Isaacs, Yoaav and Jeffrey Sanford Russell (Sept. 2023). "Updating without Evidence". In: *Noûs* 57.3, pp. 576–599. ISSN: 0029-4624, 1468-0068. DOI: 10.1111/nous.12426.

Schoenfield, Miriam (Dec. 2015a). "Bridging Rationality and Accuracy:" in: *Journal of Philosophy* 112.12, pp. 633–657. ISSN: 0022-362X. DOI: 10.5840/jphil20151121242.

— (Oct. 2015b). "Internalism without Luminosity". In: *Philosophical Issues* 25.1, pp. 252–272. ISSN: 1533-6077, 1758-2237. DOI: 10.1111/phis.12049.
— (Oct. 2017). "Conditionalization Does Not (in General) Maximize Expected Accuracy". In: *Mind* 126.504, pp. 1155–1187. ISSN: 0026-4423, 1460-2113. DOI: 10.1093/mind/fzw027.
— (May 2018). "An Accuracy Based Approach to Higher Order Evidence". In: *Philosophy and Phenomenological Research* 96.3, pp. 690–715. ISSN: 0031-8205, 1933-1592. DOI: 10.1111/phpr.12329.
Williamson, Timothy (Dec. 2002). *Knowledge and Its Limits*. 1st edition. Oxford: Oxford University Press. ISBN: 978-0-19-925656-3.
Zendejas Medina, Pablo (Feb. 2024). "Just As Planned: Bayesianism, Externalism, and Plan Coherence". In: *Philosophers' Imprint* 23.0. ISSN: 1533-628X. DOI: 10.3998/phimp.1300.