Avoiding Polarization — *draft: please do not cite without permission*

Adrian Liu, December 15 2024, 6,000 words (✉: adrian.liu@rutgers.edu)

Abstract: Epistemic agents have higher-order uncertainty when they are unsure what the rational credences are. If higher-order uncertainty is rationally permissible, then agents could obey many standard constraints on epistemic rationality yet fail to match their current credences and their expected future credences. Two agents who fail to match and do so in different directions will polarize their opinions. Exploiting this possibility, Dorst (2023) exhibits a mechanism in which ambiguous evidence – evidence that warrants higher-order uncertainty – generates expectable rational polarization. He hypothesizes that the mechanism can explain real-world cases of political belief polarization. I argue that ambiguous evidence does not generate rational pressure to polarize nor eliminate rational pressure to avoid it. I further propose a constraint on rational responses to ambiguous evidence that does not result in expectable rational polarization.

Avoiding Polarization

A fair coin lands heads 50% of the time. Suppose you now know that a certain coin is fair and will be tossed n times and that you will get reliable evidence each time about how the coin landed (e.g. a signal you know to be positively correlated with the outcome of a flip). Then you expect your evidence to tell you that around 50% of the tosses land heads, and thus you expect that if you respond rationally to your evidence, you will think that around 50% of the tosses landed heads.

Now suppose that over a large number of coin flips in the same conditions, you would think that around 60% of the coins landed heads. There seems to be some problem. If you would rationally expect around 60% of the flips to land heads, but you now think the coin is fair, then you think that your future self will wrongly think the coin is biased, while making no rational mistakes!

Dorst (2023) argues that such a series of events can happen, and that it can explain how rational agents can polarize in their beliefs. Suppose you are given a string with letters and blanks (Elga and Rayo 2022):

P _ A _ E T,

and asked whether or not there is some way to fill in the blanks to complete the string into a word. Such a task, says Dorst, leads to an asymmetric ambiguity: if you find a word, you are sure there is a completion. But if you don't find a word, it still might be that you've missed a word, so you will be uncertain what the rational credence to have in there being a word is. And this uncertainty can push your rational credences in asymmetric ways. If we then hook up a word-search task to the outcome of a coin flip such that the word is completable if and only if the coin came up heads, then it turns out we can make you polarize on the coin flip outcomes, rationally expecting around 60% of the flips to land heads!

Dorst argues that this coin-flip word-search game roughly models 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 (Dorst argues that this follows from confirmation bias, and that confirmation bias is not necessarily irrational because it can increase accuracy). When we don't find a problem

in an article, that evidence is ambiguous (like when we don't find a word in a string): perhaps there is a problem and we haven't found it, but perhaps 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.

So the coin-flip word-search game serves as a toy model of a mechanism for how we could think our future rational selves could diverge from what we now think, even while we make no rational mistakes: how we could *expectably rationally polarize*. If the story works, we have a plausible mechanism for rational polarization of opinion, of how people could look at the same information and diverge in their beliefs, even while making no rational mistakes.[1] It's a deep and insightful proposal. But I worry it does not work as it stands. This paper puts forth my worry.

1 | The Toy Model

In this section I will explain Dorst's example mechanism for generating polarization. In the remainder of the paper I'll show that although the *asymmetric ambiguity* in Dorst's proposed mechanism allows polarization, it does not necessitate or generate it (§2), and that there may remain rational pressure to avoid polarizing (§3).

Dorst argues that *when evidence is ambiguous, rational updates can lead to expectable polarization*[2] Here is the toy model:

> **Coin-Flip Word Search:** Haley (*H* for Heads) 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 for her credence that the coin landed Heads.

Let WORD be the proposition that there is a word (I'll use small caps for propositions), FIND be the proposition that Haley finds a word, and HEADS be the proposition that the coin came up heads. Given the setup, WORD is true if and only if HEADS is. Stipulate three features of Haley:

1. Haley is certain about all the facts of the setup above.
2. Haley is certain that: she is 50% accurate at finding words in completable strings. If there is a word in the string, she will find it half the time.
3. Haley is certain that: *if and only if* the string has a word but she doesn't find it, she will get evidence that there is a word in the form of a subtle hint that there is a word. Let HINT be the proposition that Haley got a subtle hint that there

1. Actually, this story is just the first part of Dorst's story for how agents can *predictably* rationally polarize. See Dorst (2023), §5 presents the second part. Though I focus on the argument of the first part in my paper, I discuss the second part in my §3
2. I'll use the term "ambiguous" to follow Dorst, since there isn't much of a standard in the literature. The phenomenon is connected to epistemic externalism and failures of failures of access principles and is sometimes called 'opaque' (Williamson 2002; Williamson 2014, Srinivasan 2015, Gallow 2021, Salow 2018, Salow 2019, Schoenfield 2017, Isaacs and Russell 2023, Schultheis forthcoming, Zendejas Medina 2024, Dorst 2019, Dorst 2020, Elga 2013.

was a word, and stipulate that HINT is *ambiguous* – when Haley gets the hint, she shouldn't be sure that she has.

How should Haley respond to her evidence in Coin-Flip Word Search? If Haley finds a word, she should be certain there is a word. That much is not at issue. But how about when she does not find a word? Well, the probability of there being a word that she doesn't find is $1/4$, and the probability of there being no word is $1/2$. Since these are the two possibilities in which Haley does not find a word, Haley can calculate that, conditional on her not finding a word, there is a $\frac{1/4}{1/4+1/2} = \frac{1/4}{3/4} = 1/3$ chance that there is a word. A simple answer, then, is the *conditionalizing* answer (what Dorst calls the Standard Bayesian answer): Haley should adopt credence 1 in WORD if she finds a word, and credence $1/3$ in WORD otherwise.

Recall that HINT (by stipulation) is ambiguous: when HINT is true, Haley should still be unsure whether it is true. The Standard Bayesian ignores HINT because its calculus assumes that evidence is never ambiguous and thus cannot work with ambiguous evidence. Dorst argues that the Standard Bayesian is incorrect in suggesting Haley ignores HINT: it is rational for Haley to take HINT into account. Since HINT is evidence in favor of WORD, Haley should raise her credence in WORD if HINT is true.

But if HINT is ambiguous then, Dorst shows, being sensitive to HINT can generate *rational expectable polarization.* For if Haley adopts any credence in WORD above $1/3$ when HINT is true, while keeping her credences the same otherwise, then over a series of trials she will have a credence in WORD, and thus HEADS, that is more than $1/2$ (this math will be explained in §3). This despite her knowing that the coin is fair, and thus that the average credence in HEADS should be $1/2$. So Haley polarizes away from the truth *by her own lights*. Furthermore, if we give Thomas (T for Tails) the same game, but flip it so he gets completable words if and only if the coin comes up tails, then Haley and Thomas can expect to polarize from each other on the ground truth of how many coins came up heads, even if they know that their strings are being chosen from the same coin flips. In this way, Dorst argues, Coin-Flip Word Search provides a model of rational expectable polarization generated by ambiguous evidence.

1.1 | *Formal Setup*

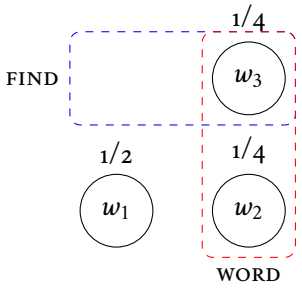
For our purposes, a credence function is a probability function whose domain is a set of propositions. Propositions themselves can be modeled as sets of possible worlds. For the Coin-Flip Word-Search example, we can let the set of possible worlds be $W = \{w_1, w_2, w_3\}$, where w_1 is the world where there is no word, w_2 is the world where there is a word but Haley doesn't find it, and w_3 is the world where there is a word and Haley finds it. Then the set of propositions is the powerset $\mathcal{P}(W)$ of W, and the credence functions are the probability functions $\{\pi : \mathcal{P}(W) \to [0, 1]\}$.

Let H be a description of Haley's prior credences: her credences before she sees the string. We can model H as a function from W to credence functions: $H[w_1]$ (which we can write as H_{w_1}) is Haley's prior credence function in w_1, $H[w_2]$ (or H_{w_2}) is Haley's prior credence function in w_2, and $H[w_3]$ (or H_{w_3}) is Haley's prior credence function in w_3.[3]

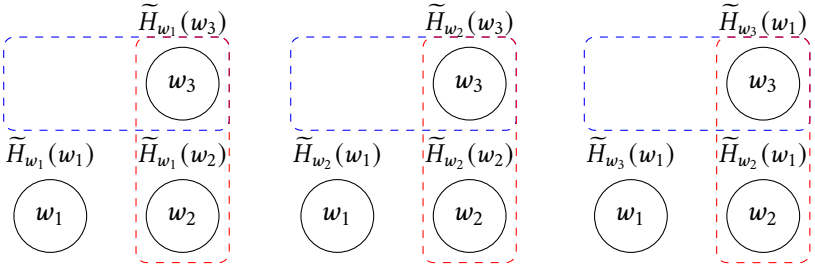
3. Note that even though Haley has not seen the string yet, we can still say she is in one of w_1, w_2, w_3: she is in a world where either there will be no word, or there will be a word she doesn't find, or there will

Before she sees the string, Haley has the same total evidence in every world: regardless of what world she is in, all she knows is what is specified in the setup. So $H_{w_1} = H_{w_2} = H_{w_3}$. I'll call this H_c for simplicity (c for constant). The subscript is meant to keep clear the distinction between credence functions (which are individuated by their inputs and outputs) and *descriptions* of credence functions, which can pick out different credence functions at different worlds. In my notation, descriptions will always be unsubscripted capital Roman letters (and it will almost always be H), and functions will always be either *subscripted* capital Roman letters or π. The figure below provides a diagram of H_c.

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

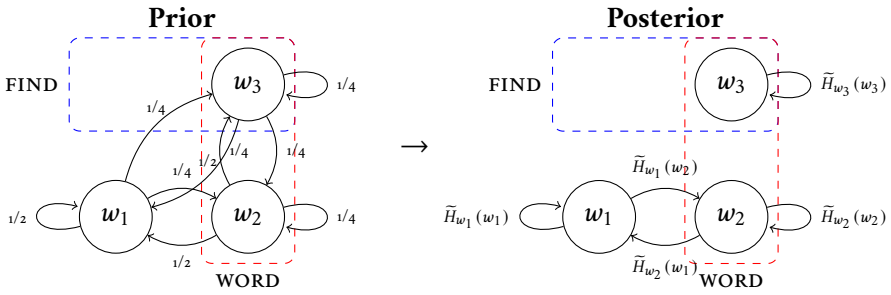


How should Haley respond to the evidence she has after she tries to complete a word? We seek an *update plan* for Haley that prescribes a set of opinions at t_1, when she sees the string, for each possibility. A complete rational update plan for Haley would involve a rational posterior credence function $\widetilde{H}_{w_i}$ at t_1 for each world w_i she might be in. In other words, it would involve a *description* $\widetilde{H}$ of Haley's posterior credences. From left to right below we have $\widetilde{H}_{w_1}$, $\widetilde{H}_{w_2}$, and $\widetilde{H}_{w_3}$, the values of $\widetilde{H}$ at every world. (We could also have drawn out three diagrams for H_{w_1}, H_{w_2}, H_{w_3} above: but since they are the same I just drew it once.)



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 $x \in [0, 1]$ if $\widetilde{H}_{w_i}(w_j) = x$. I will adopt this way of drawing posteriors, and omit arrows whenever $\widetilde{H}_{w_i}(w_j) = 0$. Using this diagramming, we can describe an *update* as a transition from a prior description H at t_0 (left diagram) to a posterior description $\widetilde{H}$ (right diagram) from worlds to credence functions, as below. (Don't worry too much about the left diagram: it's the constant prior copied three times and then condensed: so every arrow pointing into w_1 has value 1/2, every arrow pointing into w_2 or w_3 has value 1/4.)

be a word she finds.

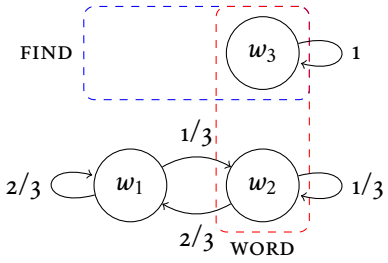


Definition 1. An **update** is a pair $\langle H, \widetilde{H} \rangle$, where H and $\widetilde{H}$ are both descriptions of credence functions, defined over the same domain.

1.2 | The Bayesian Response

What update plan is most rational for Haley to adopt? Bayesians say that Haley should conditionalize on the total body of information she becomes certain in. If HINT is not something Haley becomes certain in, then all Haley is certain in 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, since $\widetilde{H}_{\{w_1,w_2\}}(\text{WORD}) = H_c(\text{WORD} \mid \neg\text{FIND}) = \frac{H_c(\text{WORD}\&\neg\text{FIND})}{H_c(\neg\text{FIND})} = \frac{1/4}{3/4} = \frac{1}{3}$. Call it **Standard Bayesian**:

Update 1: *Standard Bayesian*

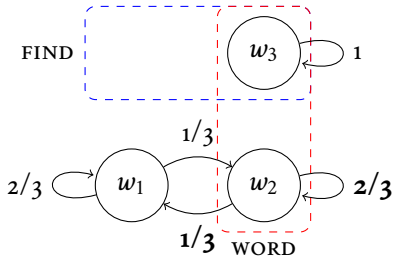


1.3 | Dorst's Ambiguous Update

Dorst argues that HINT warrants Haley increasing her credence 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 HINT is true, and leaves it unchanged from Standard Bayesian otherwise. Call it **Heads-Polarizing**:

In Heads-Polarizing, Haley *expectably polarizes*. Haley can expect 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! To write this formally, we'll use the notion of the *expectation* of a function – in this case, the function from worlds to posterior credence functions.

Update 2: *Heads-Polarizing*



Definition 2. If H_c is a credence function over $\mathcal{P}(W)$ and $\widetilde{H}$ is a description of a credence function, i.e. a function from W to credence functions $\widetilde{H}_w$, then H_c's **expectation** of $\widetilde{H}$, written $\mathbb{E}_{H_c}\left[\widetilde{H}(q)\right]$, is given by the following equation, where brackets within the arguments of probability functions mean propositions interpreted as sets of possible worlds – e.g. $[\widetilde{H}(q) = x] =_{def} \{w \mid \widetilde{H}_w(q) = x\}$:

$$\mathbb{E}_{H_c}\left[\widetilde{H}(q)\right] =_{def} \sum_{x\in[0,1]} H_c([\widetilde{H}(q) = x]) \cdot x = \sum_{x\in[0,1]} H_c(\{w \mid \widetilde{H}_w(q) = x\}) \cdot x.$$

Note that

$$\begin{aligned}\sum_{x\in[0,1]} H_c(\{w \mid \widetilde{H}_w(q) = x\}) \cdot x &= \sum_{x\in[0,1]} \sum_{w:\widetilde{H}_w(q)=x} H_c(w) \cdot \widetilde{H}_w(q) \\ &= \sum_{w\in W} H_c(w) \cdot \widetilde{H}_w(q).\end{aligned}$$

The first form weights the value of each possible posterior credence by how likely it is to occur. The second calculates expectation equivalently by weighting the posterior credence at a given possible world by how likely that world is to occur.

Then we can say that Haley's rational prior H_c *expectation reflects* her rational posterior on a proposition q if her prior rational credence H_c in q equals her prior rational credence's expectation of her posterior rational credence $\widetilde{H}$ in q:

Constraint 1 (Expectation Reflection)**.** *Say that H_c* ***expectation reflects*** *$\widetilde{H}$ on q if*

$$H_c(q) = \mathbb{E}_{H_c}\left[\widetilde{H}(q)\right]$$

and say that H_c expectation reflects $\widetilde{H}$ if $H_c(q) = \mathbb{E}_{H_c}\left[\widetilde{H}(q)\right]$ for all propositions q.[4] *Note that Dorst sometimes speaks of a* description *H expectation reflecting another description. We can say that a description H expectation reflects another description $\widetilde{H}$ if $\forall w \in W$, H_w expectation reflects $\widetilde{H}$.*

In Standard Bayesian, 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

4. Expectation reflection is a weaker form (implied by but not implying) of the principle Van Fraassen (1984) calls "reflection": $\forall q, x : H_c(q \mid \widetilde{H}(q) = x) = x$. See Elga (2007), 2013, and Briggs (2009).

credence 1/3 in heads (3/4). Then we add up the posterior credences weighted by how likely Haley's prior thinks they are: $\mathbb{E}_{H_c}\left[\widetilde{H}(\textsc{word})\right] = 1 \cdot H_c([\widetilde{H}(\textsc{word}) = 1]) + \frac{1}{3} \cdot H_c([\widetilde{H}(\textsc{word}) = \frac{1}{3}]) = (1 \cdot \frac{1}{4}) + (\frac{1}{3} \cdot \frac{3}{4}) = \frac{1}{2}$. But in Heads-Polarizing, Haley's prior does not expectation reflect its posterior – her expected posterior is higher than her prior: $H_c(\textsc{heads}) = 0.5$ but $\mathbb{E}_{H_c}\left[\widetilde{H}(\textsc{heads})\right] = (1 \cdot \frac{1}{4}) + (\frac{1}{3} \cdot \frac{1}{2}) + (\frac{2}{3} \cdot \frac{1}{4}) = \frac{7}{12} > \frac{1}{2}$.

2 | Does Ambiguity Generate Polarization?

2.1 | *Ambiguity Allows Polarization*

We have an example of expectable polarization. What's Dorst's story? That there is an asymmetric ambiguity: in some cases evidence is unambiguous in one direction; in other cases it is ambiguous in the other direction. So on average credence moves in the direction of the unambiguous evidence. When Haley finds a word she has unambiguous evidence that there is a word, but when she doesn't find a word, she has only ambiguous evidence that there is no word. Accordingly, Dorst uses higher-order uncertainty to generate the failure of expectation reflection:

Definition 3. Let H be a description of a credence function and let H_c be a credence function in the range of H. Say that H_c is **higher-order uncertain** if $\exists q, x : H_c(q) = x$ and $H_c([H_w(q) = x]) < 1$. In general, say that H is higher-order uncertain if some H_c in the range of H is higher-order uncertain.

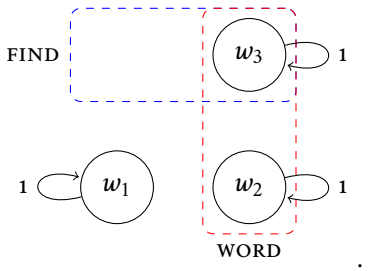
Definition 4. Say that some evidence e is **ambiguous** if it is rational to have higher-order uncertainty in some proposition in response to e.

Why is ambiguous evidence and the resultant higher-order uncertainty important? Because if being higher-order *certain* (i.e. satisfying $\forall q, x$: if $H_c(q) = x$ then $H_c(\{w \mid H_w(q) = x\}) = 1$) is a contraint on epistemic rationality, then failures of expectation reflection lead to paradigmatic failures of rationality, like susceptibility to Dutch Books and decreases in accuracy (Dorst et al. 2021, 2023). So if having no higher-order uncertainty is a contraint on epistemic rationality, Expectation Reflection is as well. Conversely, if it can be rational to have higher-order uncertainty, then rational agents can satisfy plausible rational constraints like being immune to Dutch Books and valuing one's future rational credences as more accurate, without necessarily expectation matching future rational credences. In this case, expectation Reflection would have to be *independently* motivated as a principle of rationality.

In fact, Dorst proves (2023, Theorem 3.1) that if a description $\widetilde{H}$ has no higher-order uncertainty, then a credence function H_c expectation reflects $\widetilde{H}$ if and only if it epistemically *values* it in the sense that H_c expects $\widetilde{H}$ to be strictly more accurate than itself. Dorst argues that *value* is a plausible necessary condition on epistemic rationality. If he is right, then whenever evidence is unambiguous, expectation reflection is also a necessary condition on epistemic rationality. The sense of epistemic *value* at issue above is formalized in Dorst (2019) and Dorst (2023), but I will not discuss it more here. All of the updates in this paper satisfy *value*, so it is not at issue: the question is rather whether there are *further* constraints on rationality, and in particular ones that prevent polarization.

Let us see more concretely why HINT must be ambiguous in order for Heads-Polarizing to be rational. Suppose HINT were *unambiguous*. Then it would not be rational for Haley to have higher-order uncertainty in any proposition in response to HINT. In particular, suppose $\widetilde{H}_{w_2}(\text{HINT}) = 2/3$, as in Heads-Polarizing. Then Haley would have to satisfy $\widetilde{H}_{w_2}([\widetilde{H}(\text{HINT}) = 2/3]) = 1$. Based on the diagram for Heads-Polarizing, $[\widetilde{H}(\text{HINT}) = 2/3] = [\widetilde{H}(\text{WORD}) = 2/3] = w_2$. So we must have $\widetilde{H}_{w_2}(w_2) = 1$. In other words, Haley in w_2 could reverse-engineer the fact that she is in w_2 and there is a word if the hint were unambiguous. By similar reasoning, Haley in w_1 could reverse engineer the fact that there is no word. Then the rational update would be not Heads-Polarizing, but instead what we can call **Certain Update**, which does not polarize:[5]

Update 3: *Certain Update*



Certain Update makes Haley perfectly accurate on the question of whether there is a word, even though she is only 50% accurate at actually finding the word. It also satisfies expectation reflection: H_c expectation reflects $\widetilde{H}$. But obviously Certain Update is not in the spirit of the *subtle hint* as Dorst originally conceives it. Certain Update illustrates both that ambiguity as formally defined is a necessary condition for rational polarization and that it is a good model of the intuitive sense in which the subtle hint is ambiguous. So ambiguity is necessary for polarization. but does it *generate* polarization, and does it *rationalize* polarization?

2.2 | Ambiguity Does Not Asymmetrically Generate Rational Polarization

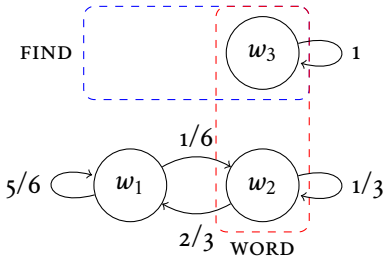
Heads-Polarizing shows that if ambiguous evidence like HINT is possible, then there exist updates that polarize. But for the general structure of Coin-Flip Word Search to be a good story of *rational* expectable polarization, we would need to establish that the ambiguity of Haley's evidence *generates* Haley's polarization – that the rationality of Haley's uncertainty around HINT is connected to Haley's polarization in a way that rationalizes the polarization.

However, the ambiguity as characterized leaves a wide range of responses on the

5. Certain Update aligns with a rule from Schoenfield (2017) that I call metaconditionalization, following Schultheis (forthcoming). This rule is discussed as "Auto-Epistemic Conditionaliation" in Hild (1998) and Zendejas Medina (2024), Conditionalization* in Schoenfield (2017) and Isaacs and Russell (2023), and "Schoenfield Conditionalization" in Gallow (2021). Instead of conditionalizing on the propositional content of one's total evidence E, metaconditionalization says to conditionalize on the proposition [my total evidence is E]. Following metaconditionalization, Haley conditionalizes on [my total evidence is *oe*] when she has HINT and [my total evidence is ¬FIND] when she doesn't have HINT.

table besides the ones Dorst suggests. And when we consider other plausible options, it becomes less clear that the ambiguity of HINT pushes Haley's credences toward a bias in HEADS rather than TAILS. Dorst reasons 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 why should Haley keep her confidence steady when she does not have HINT? If Haley gets a hint if and only if WORD is true, then getting a hint is evidence for WORD, and not getting a hint is evidence for ¬WORD. If the HINT/¬HINT evidence for and against WORD has the same absolute degree (say, 1/6), then we get **Tails-Polarizing**, wherein Haley polarizes in the *opposite direction* from Heads-Polarizing:

Update 4: *Tails-Polarizing*

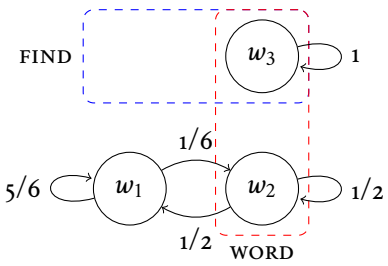


Tails-Polarizing polarizes in the opposite direction and to the same degree as Heads-Polarizing: $H_c(\text{HEADS}) = \mathbb{E}_{H_c}\left[\widetilde{H}(\text{HEADS})\right] = 5/12$ ($\mathbb{E}_{H_c}(\widetilde{H}(\text{WORD})) = (1 \cdot \frac{1}{4}) + (\frac{1}{6} \cdot \frac{1}{2}) + (\frac{1}{3} \cdot \frac{1}{4}) = \frac{5}{12} < \frac{1}{2}$). So even if ambiguity generates polarization, it doesn't generate polarization in any particular direction.

2.3 | Ambiguity Does Not Necessitate Polarization

Even if asymmetric ambiguity did not asymmetrically 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 perhaps we can find an update that does not polarize at all. A first pass: instead of shifting only when there is a subtle hint or only when there isn't, suppose Haley shifted both times, and to the same absolute degree of, say, 1/6: from 1/3 to 1/2 in w_2 and from 1/3 to 1/6 in w_1. Then we have **Tails-Polarizing 2**:

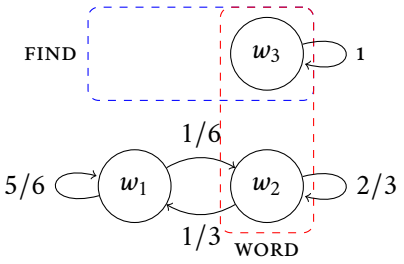
Update 5: *Tails-Polarizing 2*



This still polarizes (barely): $\mathbb{E}_{H_c}(\widetilde{H}(\text{WORD})) = (1 \cdot \frac{1}{4}) + (\frac{1}{6} \cdot \frac{1}{2}) + (\frac{1}{2} \cdot \frac{1}{4}) = \frac{11}{24}$. It turns out that what is for Haley to respond to the absence of the hint *half as much*

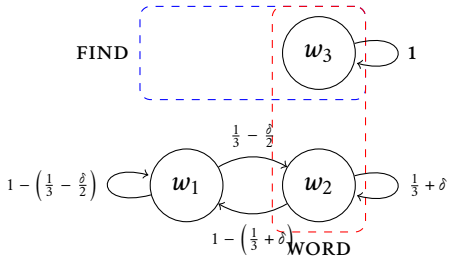
as she responds to its presence (I'll explain below why this makes sense). If she does this, we get **Nonpolarizing**, wherein Haley does not have to polarize:

Update 6: *Nonpolarizing*



We can generalize. If Haley adopts credence $1/3+\delta$ in HINT for some $\delta \in (0, 2/3)$ when there is a hint, and credence $1/3 - \delta/2$ when there is no hint, she will not polarize, and as long as $\delta \geq 0$, it is at least as accurate as Standard Bayesian.[6] In fact, we have a general formula of predicting when polarization will occur: if $\widetilde{H}_{w_2}(\text{WORD}) = 1/3 + \delta$, then the update will polarize towards HEADS if $\widetilde{H}_{w_1}(\text{WORD}) > 1/3 - \delta/2$, towards TAILS if $\widetilde{H}_{w_1}(\text{WORD}) < 1/3 - \delta/2$, and will expectation reflect otherwise:

Update 7: *Nonpolarizing Schema*



3 | Avoiding Polarization

Dorst's update and his generalization demonstrates that a range of choices lead to polarization in favor of heads. My updates demonstrate that it is also easy to either generate polarization in favor of tails or to avoid polarizing altogether. Of all these potentially rational updates, which should be favored, and why? The constraints that Dorst commits to underdetermine the answer. I'm partial to expectation Reflection, and despite sustained criticism of it from Dorst and others, I think a restricted version of it may be applicable here. In this section I will give a tentative defense. My argument will depend on considerations of how to respond to both the *presence* and *absence* of ambiguous evidence. I will argue in two steps:

1. **Be Sensitive Both Ways**: If it is rational to be sensitive to evidence, then it is rational to be sensitive to the absence of evidence.

6. Specific case: $\mathbb{E}_{H_c}(\widetilde{H}(\text{WORD})) = (1 \cdot \frac{1}{4}) + (\frac{1}{6} \cdot \frac{1}{2}) + (\frac{2}{3} \cdot \frac{1}{4}) = \frac{1}{2}$.
General case: $\mathbb{E}_{H_c}(\widetilde{H}(\text{WORD})) = (1 \cdot \frac{1}{4}) + ((1/3 - \frac{\delta}{2}) \cdot \frac{1}{2}) + ((1/3 + \delta) \cdot \frac{1}{4}) = \frac{1}{4} + \frac{1}{6} - \frac{\delta}{4} + \frac{1}{12} + \frac{\delta}{4} = \frac{1}{2}$.

2. **Be Proportionally Sensitive**: If it is rational to be sensitive to ambiguous evidence, then it is rational to be *proportionally* sensitive to its absence.

Because I like a picture in which anything an epistemic agent should be sensitive to counts as evidence, defending the positions above will amount to defending a version of the claim that absence of evidence is evidence even when evidence is ambiguous.

3.1 | *Be Sensitive Both Ways*

In Dorst's update, Haley responds in w_2 to the presence of ambiguous evidence by increasing her credence in WORD by 1/3 compared to the conditionalizing update. But in w_1, where she has not found a word and doesn't get the ambiguous evidence, she does not change her credences compared to the conditionalizing update. In Nonpolarizing, Haley lowers her credences in w_1 compared to the conditionalizing update. I think this makes sense: if Haley is sensitive to the presence of some evidence, then she should be sensitive to the absence of that evidence.

To see what I mean by "sensitive to the absence of evidence", suppose we separate out the two pieces of evidence Haley gets after being shown the string. First, there is the *unambiguous* evidence of whether or not Haley has found a word. Second, there is the *ambiguous* evidence of whether or not Haley there are subtle hints that there is a word. Now suppose that, through some Frankfurt-style neural implant, we can delay the ambiguous evidence by two seconds: if Haley gets the subtle hint, it will be with a two second delay after she learns whether or not she found a word. (We don't need this two-second delay, but it makes imagining the example easiser.) Then we can split Haley's update into two parts:

1. First, there is the update she does when she learns whether or not she found a word. Because the evidence is unambiguous and her prior has no higher-order uncertainty, let's assume she should conditionalize. So she ends up with the conditionalizing update.
2. Second, there is the further update she does when she learns, two seconds later, whether or not there was a subtle hint. Because, by hypothesis, the subtle-hint evidence is ambiguous, conditionalization is not appropriate (since conditionalizing on evidence leaves one certain that one has the evidence).

Let's limit our attention to w_1 and w_2, the cases in which Haley does *not* find a word. Then the conditionalizing update gives a constant credence function description that has credence 1/3 in w_2 and credence 2/3 in w_1. Since it is constant across w_1 and w_2, we can treat it as a new prior.

If Haley should be *sensitive* to the subtle-hint evidence, it should involve some movement in her credences in having the subtle-hint evidence. If she should be *uncertain* about the subtle-hint evidence, i.e. if it is ambiguous, then she should not become certain either that there is a hint (when there is one) or that there isn't a hint (when there isn't one). But in Heads-Polarizing, Haley is *insensitive* to the absence of the hint: when she does not get the hint, she does not change her credences at all.[7] And in Tails-Polarizing, Haley is insensitive to the presence of the hint: when she does get the hint, she does not change her credences at all.

7. Note that she is uncertain in whether she has the hint, since she has credence 1/3 that there is a word, and she knows those are correlated: however, she doesn't *change* her credence in there being a hint.

I think these two insensitivities should both strike us as an odd response to evidence. Consider the period of time after Haley has not found a word but before the hint (or lack thereof) comes through. At this point Haley knows she may or may not get a hint. In fact: she predicts that she will get a hint with probability $1/3$, since her new-prior credence in WORD is $1/3$. If getting a hint should cause her to raise her credence in having the hint (and thus WORD), then in the case where there is no word and she doesn't have the hint, the fact that she *could* have gotten the hint but did not should be evidence against WORD. In other words, there are two circumstances in which Haley doesn't get a hint: first, there is the circumstance in which a hint was never on the table. Second, there is the circumstance where there could have been a hint. My claim is that Haley should not treat these circumstances the same.

This connection between the presence and absence of evidence is clear when evidence is unambiguous. FIND is unambiguous evidence: when Haley finds a word, she is certain she has found one; when does hasn't, she is certain she hasn't. FIND and ¬FIND partition the possibilities, and Haley is always certain which of the partitions she is in. What I'm suggesting here is that the connection between presence and absence should be preserved even when evidence is ambiguous: if Haley should be sensitive to the presence of the hint, she should be sensitive to the absence of HINT, even if in both cases she is uncertain whether HINT is true or not.

Objection: there *does* seem to be a difference between HINT and ¬HINT: in HINT there is new perceptual or cognitive input, in ¬HINT there isn't. So HINT should count as evidence while ¬HINT should not. *Response:* in w_1 Haley gets new perceptual and cognitive input (new photons, the time passing, etc) – it's just that none of this new information is *hinty*. But by parity of reasoning we can say that w_2 Haley gets new perceptual and cognitive input, none of which is *non-hinty*. So it's not clear that a presence/absence distinction can create an asymmetry in what counts as evidence.

Indeed, *modeling-wise*, modeling the absence of a hint as getting no additional evidence at all (or getting the trivial proposition $\top$) is an unforced choice: it is not determined by the informal description of the coin-flip word search. Restricting attention to the second update and w_1 and w_2, we could instead say that Haley faces two possibilities, in both of which she will get new evidence: in w_2 she gets evidence HINT and in w_1 she gets evidence ¬HINT, and both pieces of evidence are ambiguous. Then if the presence of HINT warrants an increase in credence in HINT, so too does the presence of ¬HINT (which we had before been modeling as the absence of HINT) warrant a decrease in credence in HINT.

I'm suggesting a picture here wherein the absence of ambiguous evidence is itself ambiguous evidence. Another way to achieve commensurate formal results is to say that the absence of possible ambiguous evidence is not itself evidence, but is still something that an epistemic agent should update in response to (Isaacs and Russell (2023) pursue an approach along these lines). I prefer a picture in which anything an agent should be sensitive to counts as evidence, but the arguments of this paper do not depend on this framing.

3.2 | Be Proportionally Sensitive

Nonpolarizing and Tails-Polarizing 2 are sensitive to both the presence and absence of ambiguous evidence. Yet only Nonpolarizing avoids polarizing. This is because

only Nonpolarizing shifts in response to presence in absence in ways that balance out. In this subsection I defend the balancing of Nonpolarizing.

Still restricting attention to w_1 and w_2 and the second update, let's suppose now that instead of having $1/3$ credence in w_2 and $2/3$ in w_1, Haley is split 50/50 between them. Then her credence in getting the hint is $1/2$, since she gets the hint if and only if there is a word. Now suppose that the hint is ambiguous evidence that warrants a rise in credence of $\delta \in (0, \frac{1}{2})$. How much does the absence of the hint warrant a drop in credence? I claim it should just be δ. Why? If my claims in the subsection above are correct, then Haley faces a situation in which the only evidential difference between w_1 and w_2 is that in w_2 she gets a subtle hint and in w_1 she doesn't. These possibilities are equally likely according to the prior. So Haley faces two equally-likely possibilities, one of which is indicated by evidence HINT and the other of which is indicated by evidence ¬HINT. Therefore, to the extent that the presence of HINT warrants an increase in credence in HINT, then the presence of ¬HINT warrants a decrease in credence in HINT *to that same extent*: nothing about HINT and ¬HINT warrants a difference in how much of a shift in credence they warrant.

Objection: there *is* a difference between HINT and ¬HINT. HINT can be *stronger* evidence than ¬HINT. The subtle hint is a *positive* new feeling that provides evidence that there is a word, while a lack of a subtle hint is.. just the same phenomenology as before. *Response*: Perhaps HINT is *more vivid* than ¬HINT. But vividness doesn't have a straightforward connection to strength. Suppose we have a perfect fire alarm that sounds iff there is a fire in the vicinity. Then a lack of ringing is just as strong evidence in favor of there being no fire as a ringing is in favor of there being a fire, despite it being less vivid.

We should be careful talking about strength of evidence. Typically evidence is *stronger* evidence in favor of some proposition q if it warrants higher credence in q (compared to other evidence). Unambiguous evidence is always maximally strong evidence in favor of itself, since it warrants certainty in itself. But ambiguous evidence can be ambiguous to different degrees and thus have strength in favor of itself to different degrees: e.g. a subtle hint that warrants credence $5/6$ in HINT is stronger than a subtle hint that only warrants credence $2/3$ in HINT.

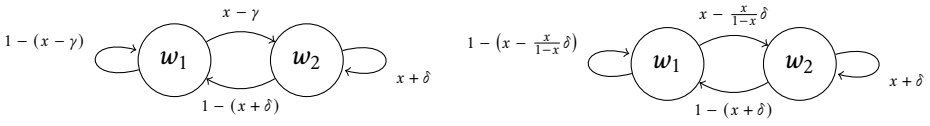
What *does* determine strength of ambiguous evidence in favor of itself? Intuitively it depends on how easy it is to tell that you have the evidence, when you in fact have it. But because increasing your credence that you have some evidence is the same as decreasing your credence that you *don't* have it, the question of how easy it is to tell when you have some evidence is tied up with the question of how easy it is to tell when you *don't* have it. My conjecture is that these two questions are tied up such that when the presence or absence of evidence is equally likely, it is exactly as easy to tell that you have ambiguous evidence (when you do have it) as to tell that you don't have it (when you don't have it). I presently have no successful positive arguments for thinking the conjecture is true: so here my argumentation is limited to insisting that there is nothing to break the symmetry.

3.3 | Proportional Sensitivity and Expectation Reflection

Instead of defending Be Proportionally Sensitive further, I will explore how it might be generalized. Return to the case where Haley has $1/3$ credence in w_2 and $2/3$ in w_1,

and suppose that the hint is ambiguous evidence that warrants a rise in credence of $\delta \in (0, \frac{2}{3})$. The generalization is that ¬HINT should take Haley the same proportion of the way to certainty in ¬HINT as HINT takes her in the direction of certainty in HINT. If at w_2 she should increase her confidence 50% of the way to certainty in w_2, then at w_1 she should decrease it 50% of the way to certainty in w_1. Mathematically, this is the most obvious way to deliver parity of strength in the presence and absence cases. Intuitively, it also reflects the idea that evidence is more informative the more unlikely it is.[8]

Generalizing further: Suppose that Haley's prior credence function has credence x in the world where she gets e and thus credence $1 - x$ in the world in which she does not get e. Now suppose that the presence of e warrants an increase of δ for $\widetilde{H}_{w_2}(w_2)$, and the absence of e warrants a decrease of γ for $\widetilde{H}_{w_1}(w_2)$. Then evidence is proportional in strength in the way specified above just when $\gamma = \frac{x}{1-x}\delta$:[9]



Constraint 2 (Proportional Sensitivity)**.** *Let w_1 and w_2 be two worlds that differ evidentially only in that a subject at w_2 will get evidence e and a subject at w_1 does not. Let H_c be a prior credence function such that $H_c(w_2) = H_c([e \in E]) = x$ (and thus $H_c(w_1) = 1 - x$), where E is the subject's total posterior evidence. Suppose that e is ambiguous and warrants posterior credence $\widetilde{H}_{w_2}([e \in E]) = x + \delta$, for $\delta \in (0, 1 - x)$. Then $\widetilde{H}$ satisfies **proportional sensitivity** if $\widetilde{H}_{w_1}([e \in E]) = x - \frac{x}{1-x}\delta$.*[10]

As it turns out, updates following Proportional Sensitivity satisfy expectation reflection.[11] In fact, updates that are proportionally sensitive to the presence and absence of ambiguous evidence in the way specified above will satisfy this constraint:

Constraint 3 (Restricted Expectation Reflection)**.** *If a description H is higher-order certain, then an update $\langle H, \widetilde{H}\rangle$ satisfies **restricted expectation reflection** if $H_w(q) = \mathbb{E}_{H_w}\left[\widetilde{H}(q)\right]$ for all propositions q and credence functions H_w in the range of H.*

8. When talking about the strength of evidence in favor of other propositions, this is known as Eddington's observation, and can be modeled by Standard Bayesianism's commitments to probabilism and conditionalization (Lin 2024). Here, where we're talking about the strength of ambiguous evidence in favor of itself, I'm suggesting that the general norm below captures this version of the observation.
9. If the presence of e warrants an increase of credence from x to $x + \delta$ in w_2, this is an absolute increase of δ toward 1 (certainty), and a *proportional* increase of $\epsilon = \delta/(1 - x)$ toward 1. In the other direction, a proportional increase of $\epsilon = \delta/(1 - x)$ in w_1 from $1 - x$ toward 1 (certainty). This is an *absolute* increase of $\epsilon(1 - (1 - x)) = \frac{x}{1-x}\delta$.
10. If we model the absence of evidence e with another piece of evidence f that the subject gets only in w_1, then we can equivalently state proportional sensitivity as the constraint that if e is ambiguous and warrants posterior credence $\widetilde{H}_{w_2}([e \in E]) = x + \delta$, for $\delta \in (0, 1 - x)$, then f warrants posterior credence $\widetilde{H}_{w_2}([f \in E]) = x - \frac{x}{1-x}\delta$.
11. We want $\mathbb{E}_{H_c}\left[\widetilde{H}(w_2)\right] = H_c(w_2)$. Now, we have $H_c(w_2) = x$ and $\mathbb{E}_{H_c}\left[\widetilde{H}(w_2)\right] = x(x + \delta) + (1 - x)\gamma$, so we want to solve the equation $x(x + \delta) + (1 - x)\gamma = x$ for γ. Rearranging terms gives $\gamma = \frac{x}{1-x}\delta$.

Restricted Expectation Reflection says that we should preserve expectation matching when the prior is higher-order certain. It is satisfied by Proportional Sensitivity and thus by the structural relationship I've defended between a prior's estimation of getting evidence and the absence- and presence-ambiguity of evidence. Sensitivity to presence and absence, I've suggested, are not allowed to float free of each other: if you think you might get ambiguous evidence, then its absence is also ambiguous evidence. Just as Dorst objects that the Bayesian leaves evidence on the table if they ignore HINT, I am suggesting that Dorst's update leaves evidence on the table if it is not similarly sensitive to the absence of HINT.

Proportional sensitivity to the presence and absence of opaque evidence is compatible with updates that don't use conditionalization, as we've already seen. It follows naturally from an accuracy-first approach to updating, and is implied, for instance, by the updating rules proposed by Schoenfield (2017), Gallow (2021), and Isaacs and Russell (2023) (though I do not derive it from accuracy-first considerations).[12] Proportional Sensitivity also imposes no constraints on any higher-order uncertain description. The functions in a higher-order uncertain description do not even expectation reflect themselves (**dorstHigherOrderUncertainty2019aa**, 2020), so if they are rational then it is dubious that rationality could require them to expectation reflect a further posterior. So restricted expectation reflection does not give any guidance to agents that are *already* higher-order uncertain, but it can still restrict allowable ways to *become* uncertain.

This result, even if limited in scope, is not trivial. Dorst's overarching story is about the possibility of rational polarization that is not only *expectable* but *predictable*: he argues that not only can rational agents expect themselves to polarize *on average*, but they can also predict with high confidence that they *will* polarize. His mechanism of generating predictable polarizing over a series of coin-flip word-search trials is for Haley to forget her higher-order uncertainty after each trial. When faced with the next trial, Haley can be modeled as having a higher-order certain prior once more – in which case each of her updates can satisfy Proportional Sensitivity and Restricted Expectation Reflection. Following Proportional Sensitivity is one way of following restricted expectation reflection, and is thus one way of avoiding polarization.[13]

References

Briggs, R.A. (Jan. 2009). "Distorted Reflection". In: *The Philosophical Review* 118.1, pp. 59–85. ISSN: 0031-8108, 1558-1470. DOI: 10.1215/00318108-2008-029.

Dorst, Kevin (Oct. 2019). "Higher-Order Uncertainty". In: *Higher-Order Evidence*. Ed. by Mattias Skipper and Asbjørn Steglich-Petersen. 1st ed. Oxford University PressOxford, pp. 35–61. ISBN: 978-0-19-882977-5 978-0-19-186827-6. DOI: 10.1093/oso/9780198829775.003.0002.

12. Proportional Sensitivity is derivable from but weaker than the update rules proposed by these authors. I derive it from and defend it with much weaker assumptions about the nature of ambiguous evidence.

13. I am grateful to Juan Comesana, Kevin Dorst, Andy Egan, Adam Elga, Dmitri Gallow, Isabel Uriagereka Herburger, Simon Hutteger, Jill North, Miriam Schoenfield, Yong Xin Hui, Pablo Zendejas Medina, participants of the MIT Epistemology Reading Group in November 2024, and members of seminars at Rutgers in 2023 and 2024.

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.

Hild, Matthias (1998). "Auto-Epistemology and Updating". In: *Philosophical Studies* 92.3, pp. 321–361. ISSN: 00318116. DOI: 10.1023/A:1004229808144.

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.

Lin, Hanti (2024). "Bayesian Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University.

Salow, Bernhard (July 2018). "The Externalist's Guide to Fishing for Compliments". In: *Mind* 127.507, pp. 691–728. ISSN: 0026-4423, 1460-2113. DOI: 10.1093/mind/fzw029.

— (Apr. 2019). "Elusive Externalism". In: *Mind* 128.510, pp. 397–427. ISSN: 0026-4423, 1460-2113. DOI: 10.1093/mind/fzx015.

Schoenfield, Miriam (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.

Schultheis, Ginger (forthcoming). "Accurate Updating". In: *Philosophy of Science*.

Srinivasan, Amia (Oct. 2015). "Normativity without Cartesian Privilege". In: *Philosophical Issues* 25.1, pp. 273–299. ISSN: 15336077. DOI: 10.1111/phis.12059.

Van Fraassen, C. (May 1984). "Belief and the Will". In: *The Journal of Philosophy* 81.5, p. 235. ISSN: 0022362X. DOI: 10.2307/2026388.

Williamson, Timothy (Dec. 2002). *Knowledge and Its Limits*. 1st edition. Oxford: Oxford University Press. ISBN: 978-0-19-925656-3.

— (Oct. 2014). "Very Improbable Knowing". In: *Erkenntnis* 79.5, pp. 971–999. ISSN: 0165-0106, 1572-8420. DOI: 10.1007/s10670-013-9590-9.

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.