Avoiding Polarization

Adrian Liu, adrian.liu@rutgers.edu
July 16th, 2024, ANU Philosophy, Formal Epistemology Workshop
This handout also at constitutive.net/few

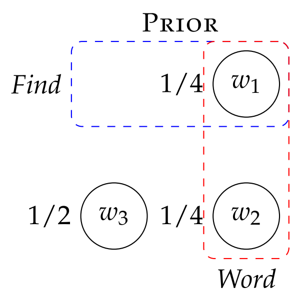
Dorst (2023): *Epistemic polarization can be rational, since it can be rational to expect your future rational credence to diverge from what it is now.* **Question:** *When* can it be rational to expect your future rational credence to diverge? **Kevin's story:** When you get *ambiguous evidence* that leaves it rational to be *higher-order uncertain*: uncertain about what the rational opinions are. **This talk:** I question the story.

Plan:
Part I: Give you Kevin's story.
Part II: Question Kevin's story.
Part III: Discuss an alternate story.
Part IV: Subvert expectations (shh).

I. Introducing Uncertainty

N E _ _ _ _ O N N _
_ _ V E U _ P _ E R

I flip a fair coin and show you a string. If Heads, the string can be completed into a word by filling in the blanks. If Tails, it cannot be completed. I ask for your credence that the coin came up heads.[1] *You know you are 50% accurate: you find a word half the time there is one. How should you update?*



← Before you see the string, you should think it 1/4 likely there is a word and you find it (w_1), 1/4 likely there is a word and you don't find it (w_2), and 1/2 likely there is no word (w_3).

[1] Since you know there is a word iff the coin came up heads, this is the same as the chances that there was a word in the string: P(*Heads*) = P(*Word*) everywhere.

Conditionalize recommends:

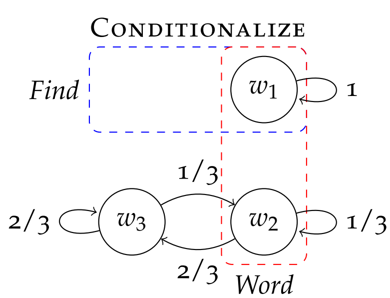
1. (w_1) If you find a word, you know there is a word. $P_1^+(Word) = 1$.
2. (w_2, w_3) If you don't find a word, calculate the chance that there is a word *given that you didn't find a word*: $P_{2,3}^+(Word) = 1/3$.[2]

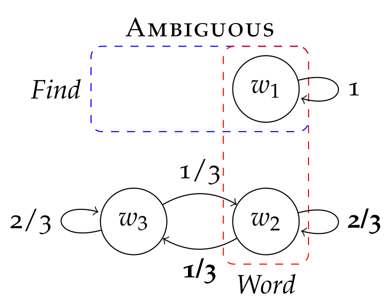
Ambiguous recommends: If you didn't find a word, you could be *uncertain* whether there is a word: perhaps you get *ambiguous evidence* in the form of a subtle feeling that there is a word.[3] Thus:

1. In w_1 and w_3, do the same thing as in Conditionalize.
2. But if you get ambiguous evidence (in w_2), raise your credence somewhat: $P_2^+(Word) = 2/3$.

[2] Where P is the prior credence and P_w^+ is the posterior in world w, $P_{2,3}^+(Word) = P(Word \mid \neg Find) = \frac{P(Word \& \neg Find)}{P(\neg Find)} = \frac{1/4}{3/4} = \frac{1}{3}$.

[3] Here we make the idealizing assumption that in fact you get this ambiguous evidence only if there is in fact a word.





← In Prior the rational credence was the same everywhere. After you see the string you have different evidence, and thus different rational credences, at different worlds. A labeled arrow from w_i to w_j represents $P_i^+(w_j)$, the rational credence at world w_i that one is at w_j. (I omit arrows with zero probability).

• Ambiguous is always at least as accurate as Conditionalize: it is exactly as accurate in w_1 and w_3; in w_2 it is more accurate. But:

• Prior *experts to diverge* on *Heads* if following Ambiguous: the average posterior credence in *Heads*, is 7/12 > 0.5.[4] So it seems the uncertainty makes you expect to think that the fair coin is biased!

[4] it's 1/4 likely that you end up with $P_1^+(Word) = 1$, 1/4 likely you end up with $P_2^+(Word) = 2/3$, and 1/2 likely you end up with $P_3^+(Word) = 1/3$.

Kevin's story: In Ambiguous, the prior *expects the posterior to diverge* on some proposition if it does not **expectation-reflect** it:[5] if its credences do not equal its calculations of the average credence it expects Ambiguous to have. If two people use Ambiguous and we give them word searches in opposite directions (*Word* iff *Heads* / *Word* iff *Tails*), they will expect their posteriors to diverge in opposite directions. *So if* Ambiguous *can be rational, then polarization can be rational.*

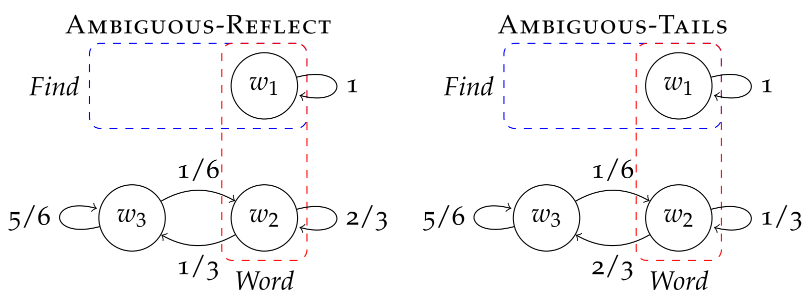
Ambiguous can be rational only if *higher-order uncertainty* can be rational. It must be possible to rationally update on the subtle feeling while being uncertain if the feeling is good evidence (and thus whether I should update on it). Say I am **higher-order uncertain** in a proposition q if I have credence t in q but I am uncertain that t is the rational credence to have.[6] If rationality disallowed higher-order uncertainty, expected divergence would be impossible.[7]

But Kevin makes a stronger claim: in Ambiguous, higher-order uncertainty not only allows but also *generates* polarization. Does it?

II. Uncertainty Underdetermines

For all that Kevin's story builds in, we can still avoid polarizing.

1. The higher-order uncertainty does not *necessitate* polarization: Prior does not expect Ambiguous-Reflect to diverge.[8]
2. Nor does it favor polarization in any *particular* direction: Prior expects Ambiguous-Tails to diverge in favor of *tails*.[9]



What did we do here? For Ambiguous-Reflect, we calibrated to expectation-reflect.[10] For Ambiguous-Tails, we just biased the weights in the opposite direction as Kevin did.[11] *So the uncertainty resulting from ambiguous evidence does not* ***generate*** *polarization.*

The Dialectic Now: A prior can *value* (fn7) a higher-order uncertain posterior while failing to *expectation-reflect* it. But as we've seen, the prior doesn't *have to* fail to expectation-reflect the posterior! If we give up on expectation-reflection for higher-order uncertain posteriors, rationality neither *rules out* nor *generates* polarization.

[5] A credence function π expectation-reflects a family of credence functions $P^1 : W \to \{P^1_w\}$ on a proposition q if $\pi(q) = \mathbb{E}_\pi(P^1(q))$, where $\mathbb{E}_\pi(P^1(q)) := \Sigma_{w\in W}\left(\pi(w)\cdot P^1_w(q)\right)$. A prior expects a posterior to diverge if it does not expectation-reflect it.

[6] Letting $R : W \to \{R_w\}$ be a definite description for the rational credence function family, whatever it is, a credence function π is higher-order uncertain in a proposition q if $\pi(q) = t$ but $\pi([R(q) = t]) < 1$. Here $[R(q) = t] = \{w \in W \mid R_w(q) = t\}$.

[7] Dorst (2023) If a prior *values* a posterior and the posterior is not *higher-order uncertain*, then the prior cannot *expect the posterior to diverge.* A prior *values* a posterior when it defers decisions to the posterior, in a way that can be formalized (Dorst et al 2021). All the examples in this talk satisfy value, so it's not directly at issue.

[8] $\mathbb{E}_P(P^+(Word)) = \sum_{w\in W} P(w)\cdot P^+_w(Heads) = \frac{1}{4}1 + \frac{1}{4}\frac{2}{3} + \frac{1}{2}\frac{1}{6} = \frac{1}{2}$.

[9] $\mathbb{E}_P(P^+(Word)) = \frac{1}{4}1 + \frac{1}{4}\frac{1}{3} + \frac{1}{2}\frac{1}{6} = \frac{5}{12}$.

[10] This is always possible if the prior is higher-order certain (Dorst et al 2021). For this setup, the prior expectation-reflects any posterior satisfying the equation $P^+_3(w_2) = \frac{1}{2}(1 - P^+_2(w_2))$.

[11] A bunch of posteriors will be valued by the prior (Dorst 2023, Thm3.2), and at least one will be expectation-reflected by it (the prior is a Markov chain with a stationary distribution).

Should rationality require expectation-reflecting? **Arguments Against**:

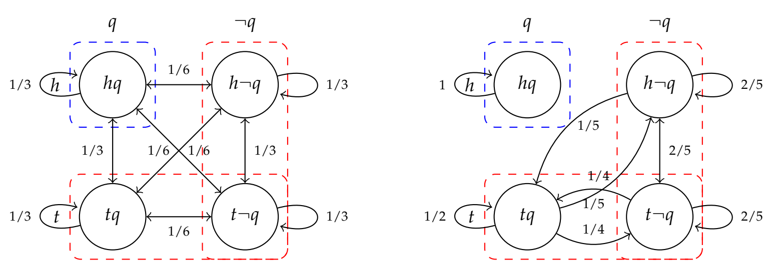
1. *It is too onerous to calculate an expectation-reflecting posterior.*
   Response: maybe rationality is hard when evidence is ambiguous!
2. *If we required expectation-reflecting in general, it would forbid higher-order uncertainty. So we need a* positive *argument for it in specific cases.*
   Response: In our cases the posterior is higher-order uncertain. But the prior is higher-order certain. So it *is certain* how likely it thinks the posterior is to be in any given one of the situations and can calibrate accordingly,[12] even though it knows that the prior will rationally be unsure about how likely it is to be in those scenarios.

So consider this **Constraint**, which guarantees that agents avoid polarizing when they begin with certainty: *If the prior is not higher-order uncertain, then it should expectation-reflect the posterior, even if the posterior is higher-order uncertain.*[13]

III. Uncertain Beginnings

But what if an agent *begins* with higher order uncertainty? Then even conditionalizing on propositions can lead to new bias.[14] For instance, suppose that you are uncertain how good you are at finding a word. You think you are 50% accurate, but you leave open that you could be more accurate (say, 75%) or less accurate (say, 25%).[15] Then conditionalizing results in small amounts of polarization.[16]

A Simpler Demonstration:[17] Suppose you start out with higher-order uncertainty about a proposition q (left).[18] And suppose a coin is tossed and you're told whether you're in the world where q is true and the coin came up heads (upper left, $\{hq\}$) or not ($\{tq, h\neg q, t\neg q\}$). In this case the prior P, conditionalized on the evidence, returns a posterior P^+ (right) that is not expectation-reflected by P.[19,20]



But *starting out with* higher-order uncertainty raises new questions:

i. In higher-order uncertain cases, what notion of *expectation* do we end up with, and is this notion relevant for polarization? **(§IV)**

ii. Higher-order uncertain priors are, on Kevin's picture, *already polarized*, because they do not expectation-reflect themselves.[21] But then how seriously do we take their expectations of other credences?

[12] A conditionalizing update will always be calibrated correctly. So if the posterior responds to uncertainty by diverging from conditionalization in ways that are *symmetric* around conditionalization from the persepective the prior, it will continue to be expectation-reflected by the prior.

[13] If your prior obeys $\forall w, q, t[P_w(q) = t \rightarrow P_w(R(q) = t) = 1)$ then your prior should also obey $P_w(q) = \mathbb{E}_{P_w}(P^1(q))$.

[14] And then **Constraint** does not apply: its antecedent is not satisfied.

[15] Prior is uncertain what the posterior conditional credences should be.

[16] Dorst, in conversation. The model assumes that each trial is independent, and that you don't update on your revised estimates of how good you are at finding a word between trials.

[17] Dorst, in conversation / unpublished.

[18] E.g.: at q worlds (left) you are 2/3 confident in q but leave open that you are in $\neg q$ worlds and thus should be 1/3 confident in q. In this case every node has a 2/3 arrow to itself, a 1/3 horizontal arrow, a 1/3 vertical arrow and a 1/6 diagonal arrow.

[19] Nor is P^+ expectation-reflected by the constant prior $\pi_c := (\frac{1}{4}\,\frac{1}{4}\,\frac{1}{4}\,\frac{1}{4})$.

[20] It's stronger than this: the only update that does not expectably diverge is one that is certain in $\{tq\}$ whenever $\{tq\}$ is true! And this certainty seems unwarranted, if the uncertainty in q were formerly warranted.

[21] They are *synchronic* expectation-reflection failures, where $P_v(q) \neq \sum_{w \in W} P_v(w) \cdot P_w(q)$. Any frame with higher-order uncertainty fails to expectation-reflect itself (Dorst 2019).

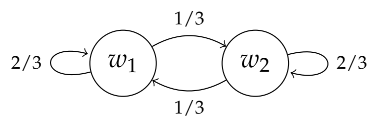
IV. Uncertain Expectations (Question (i.))

Kevin: *if a prior does not expectation-reflect a posterior, then over repeated trials the prior expects the posterior to polarize.* **Question:** In cases with higher-order uncertainty, what formalization of "expectation" validates this? The standard definition (call it S-Expectation) [22] delivers a weird result in cases with higher-order uncertainty: higher-order uncertain credences always fail to expectation-reflect themselves on some proposition.[23]

[22] $\mathbb{E}_\pi(P(q)) := \Sigma_{w \in W}\,(\pi(w) \cdot P_w(q))$

[23] Dorst 2019.

Example: This diagram represents a rational credence family P that knows that the rational credence family is described by the diagram and at each world knows its own credences, but is unsure whether it is rational.[24] What does the standard expected-value calculation do? Let's walk through it.[25]



← Also note it is just the bottom two worlds of Ambiguous.

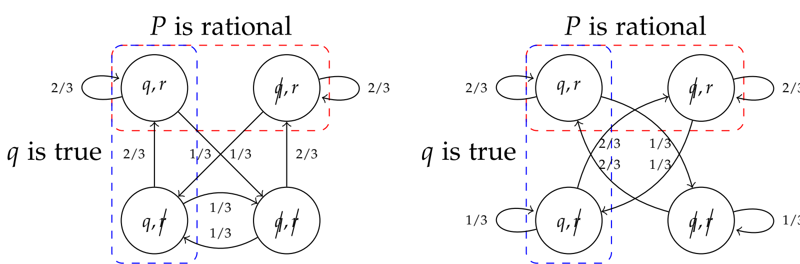
What we think we're asking: "How does P_w expect itself to do?"
What we're actually asking: "How does P_w expect to do *if it is rational at every world*?" If some P_w is uncertain that it is rational, the two diverge:[26] In fact, P will think that at some possible world, it is *irrational*.[27] How do we capture this possiblity of irrationality?

[24] Since it is unsure what world it is at, and thus what credence is rational.

[25] E.g.: $\mathbb{E}_{P_v} P(w_1) := \sum_w P_v(w) P_w(w_1)$. For each w, we ask how likely P_v thinks we are to be at w, we ask what the rational credence in w_1 to have at w is, and we multiply. We sum the results. So we have $\frac{2}{3}\frac{2}{3} + \frac{1}{3}\frac{1}{3} = \frac{5}{9} < \frac{2}{3} = P_{w_1}$.

[26] If P were higher-order certain, these would be equivalent.

[27] If P_w is higher-order uncertain in q and $P_w(q) = t$, then P_w leaves open some world where it has credence t in q irrationally (and if P is certain of its own credences it is certain it will be irrational somewhere). *Proof: exercise.*



Left: S-Expectation. *Right:* U-Expectation (what we thought we were asking). $\mathbb{UE}_\pi P(q) := \sum_{w,\rho} \pi(w) P_w(\mathbb{M}_@ = \rho)\rho(q)$.[28]

1. S-Expectation takes the expectation of *the rational credence, whatever it is.* It characterizes cases in which *I will actually be correct, but I am uncertain whether I will be (I underestimate my rationality).*
2. U-Expectation takes the expectation of *the rational credence's modest predictions of its own performance.* It characterizes cases in which *I correctly suspect that I have some chance of being incorrect.*[29]

Claim: U-Expectation is more relevant for actual polarization.

[28] where $\mathbb{M}_@$ is an indexical term for "my actual credence", π is a credence function, and ρ is a variable ranging over credence functions.

[29] The constant prior π_c $\mathbb{UE}$-reflects $P+$ in the example on the previous page (fn19). But the uncertain prior P does not. This relates to question (ii) above.

I am grateful to Juan Comesaña, Andy Egan, Adam Elga, Isabel Uriagereka Herburger, Yong Xin Hui, and especially Dmitri Gallow and Kevin Dorst, for extensive discussion and comments. Kevin in particular has essentially taught me formal epistemology via email correspondence.

References:
Dorst (2019) "Higher-Order Uncertainty"
Dorst et.al (2021) "Deference Done Better"
Gallow (2021) "Updating for Externalists"
Dorst (2023) "Rational Polarization"