Avoiding Polarization

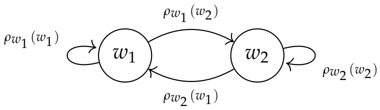
When evidence is ambiguous, a norm against being more wrong and a norm against being more biased come apart. Dorst (2023) argues that when evidence is asymmetrically ambiguous, agents can rationally get *less* wrong while getting *more* biased: ***ambiguous evidence generates expectable rational polarization***. I argue: even when evidence is ambiguous, it is possible to avoid polarization. Ambiguity does not generate expectable polarization. I suggest independent motivation for having a norm against being biased in addition to the norm against being more wrong.

Adrian Liu, adrian.liu@rutgers.edu
January 10, 2025, APA Eastern

Handout: adrianliu.net/avoid
Paper: adrianliu.net/avoidpaper

1. Ambiguity, Modesty, and Rational Norms (Background Setup)

Let R be the rational opinions, whatever they are. Say that R is modest if it is uncertain that its actual opinions are the rational ones to have. We can model R as a map $w \mapsto \rho_w$ which maps each world w to a credence function ρ_w describing the rational opinions at w. Say that R is **modest** if at some world w, ρ_w is not sure that the rational opinions are ρ_w, and **immodest** if at every world w, ρ_w is certain that the rational credences are ρ_w. Say some evidence e is **ambiguous**[1] if when one gets e at some world, it is rational for R to be modest.



$$R : w_1 \mapsto \rho_{w_1}, w_2 \mapsto \rho_{w_2}$$

Def: R is **modest** if $\exists w : \rho_w(R = \rho_w) < 1$ and **immodest** if $\forall w : \rho_w(R = \rho_w) = 1$.

Why is ambiguous evidence significant? Because two constraints are *equivalent* when evidence is unambiguous and thus R is immodest, but *inequivalent* when evidence is ambiguous and thus R is modest:

1. **Total Trust**: If π thinks R is rational, then π expects R to be more accurate than it in every possibility.
2. **Expectation Reflection**: If π thinks R is rational, then π expects R to have the same opinion, on average, as π.

Total Trust is an *anti-being wrong* norm, and Expectation Reflection is an *anti-being biased* norm. Both seem important, but if evidence can be ambiguous, the two can come apart, and we should answer the question: are they *both* rational norms?

Def: Where π is a credence function, X is a random variable, and

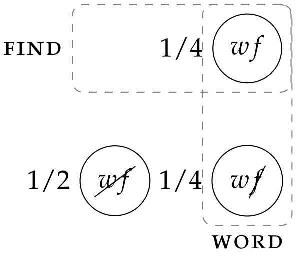
$$\mathbb{E}_\pi(X) := \sum_x x \cdot \pi(X = x)$$

is π's **expectation** of X, π **totally trusts** R if $\forall X, t : \mathbb{E}_\pi(X \mid \mathbb{E}_R(X) \geq t) \geq t$. Here $\mathbb{E}_R(X)$ is the random variable $w \mapsto \mathbb{E}_{\rho_w}(X)$.

Def: With π, X as above, π **expectation reflects** R if $\forall X : \mathbb{E}_\pi(X) = \mathbb{E}_\pi(\mathbb{E}_R(X))$.

2. Does Ambiguity Generate Polarization? (Kevin's Story)

Coin-Flip Word Search: *A fair coin is flipped. Haley is shown a string (of letters) If* heads*, the string can be completed into a word by filling in the blanks. If* tails*, it cannot be completed. Haley is asked her credence in* heads.[2] *She knows she is 50% accurate at word-search: she finds a word half the time there is one. When there is a word Haley doesn't find: she gets ambiguous evidence that there is a word: a subtle hint.*



P_A_ET ____MT
P_G_ER _R__R_

Before Haley sees the string, she should think it 1/4 likely there is a word and she finds it (wf), 1/4 likely there is a word and she doesn't find it ($w\cancel{f}$), and 1/2 likely there is no word ($\cancel{wf}$).

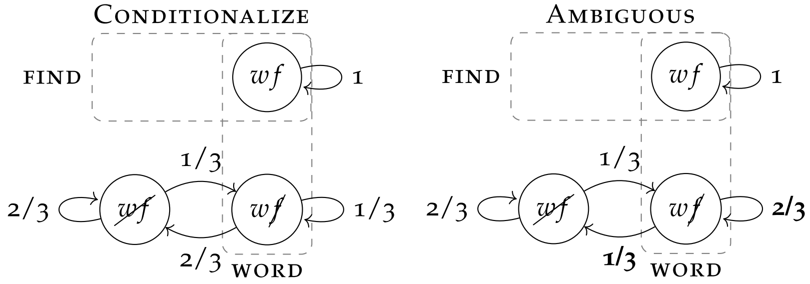
[1] *Possible examples:* faraway signs, vague feelings, hunches, looking at unmarked clock, being unsure if bored.

[2] Since she knows 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: $\pi(\textsc{heads}) = \pi(\textsc{word})$everywhere.

What should Haley do? Conditionalize says:

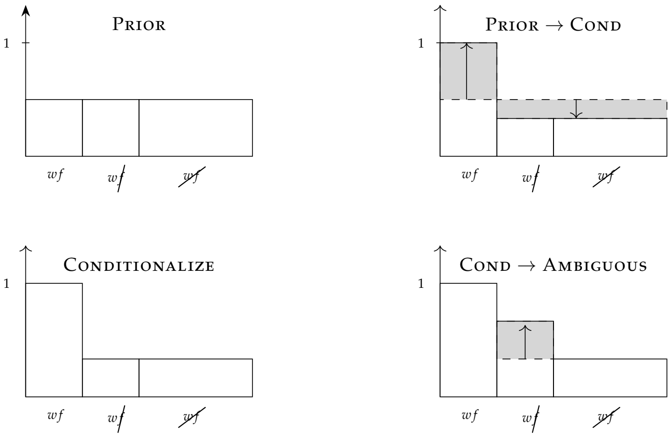
1. If Haley finds a word, she knows there is one: $\pi^+_{wf}(\textsc{word}) = 1$.
2. If Haley doesn't find a word, she calculates the chance that there is a word *given that she didn't find one*: $\pi^+_{w\cancel{f}}(\textsc{word}) = \pi^+_{\cancel{wf}}(\textsc{word}) = 1/3$.[3]

Ambiguous says: Haley shouldn't ignore the subtle hint. The hint is evidence! When Haley doesn't find the word but there is one, she should raise her credence somewhat: $\pi^+_{w\cancel{f}}(\textsc{word}) = 2/3$.



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

Haley's prior trusts both Conditionalize and Ambiguous. And Ambiguous is always at least as accurate as Conditionalize. But Haley *expects bias* on heads if she follows Ambiguous: her average credence in word, and thus heads, in the future is 7/12 > 0.5.[4] So ambiguous evidence makes her expect to think a fair coin is biased!



← A "credence-mud" representation for the proposition heads/word. The y-axis is credence in heads/word; x axis is different worlds. So the height of a given rectangle is Haley's credence in word at some world (labeled at its bottom), and the width of the rectangle is the prior's credence in being at that world. The area of each rectangle thus models the value $\pi(w)\pi^+_w(\textsc{word})$, where w is the world labeled below the rectangle. A posterior expectation reflects the prior if it has the same total area as prior.

Kevin's story: If two people use Ambiguous and we give them word searches in opposite directions (word iff heads / word iff tails), they expect their posteriors to diverge in opposite directions. *So if* Ambiguous *can be rational, then expectable polarization can be rational.* But Kevin makes a stronger claim:[5] ambiguity not only allows but *generates* polarization. Does it?

[3] Where π is the prior credence and π^+_w is the posterior in world w, we have $\pi^+_{w\cancel{f}}(\textsc{word}) = \pi^+_{\cancel{wf}}(\textsc{word}) = \pi(\textsc{word} \mid \neg\textsc{find}) := \frac{\pi(\textsc{word}\&\neg\textsc{find})}{\pi(\neg\textsc{find})} = \frac{1/4}{3/4} = \frac{1}{3}$.

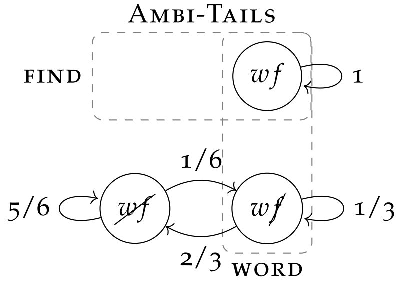
[4] it's 1/4 likely that she ends up with $\pi^+_{wf}(\textsc{word}) = 1$, 1/4 likely she ends up with $\pi^+_{w\cancel{f}}(\textsc{word}) = 2/3$, and 1/2 likely she ends up with $\pi^+_{\cancel{wf}}(\textsc{word}) = 1/3$.

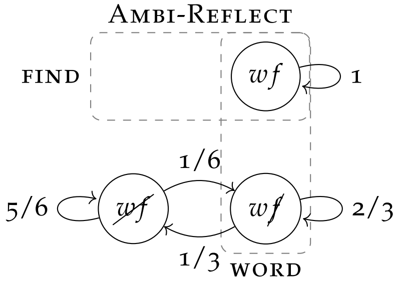
[5] This is just in §4 of "Rational Polarization". In §5, Kevin makes an even stronger claim: that ambiguity can generate *predictable* rational polarization. We can discuss this in Q&A, but I'll be focusing on the claim about *expectable* polarization, a necessary component of the stronger claim.

3. Avoiding Polarization (My Counterexamples)

Not in this example, because even given ambiguous evidence, Haley can avoid polarization: The ambiguity does not favor polarization in any *particular* direction, and it doesn't even *necessitate* polarization. So ambiguous evidence does not *generate* polarization.

For Ambi-Tails, we just biased the weights in the opposite direction as Kevin did. For Ambi-Reflect, we calibrate to expectation-reflect. This is always possible: for this setup, we just have to solve the equation $\pi^+_{\cancel{wf}}(w\cancel{f}) = \frac{1}{2}(1-\pi^+_{w\cancel{f}}(w\cancel{f}))$.

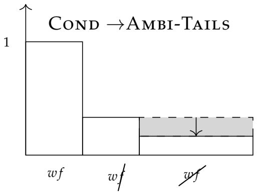


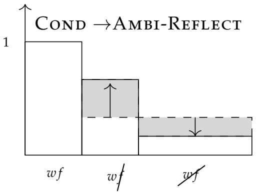


← For Ambi-Tails, $\mathbb{E}_\pi(\pi^+(\textsc{word})) = \frac{1}{4}1 + \frac{1}{4}\frac{1}{3} + \frac{1}{2}\frac{1}{6} = \frac{5}{12}$.

For Ambi-Reflect, $\mathbb{E}_\pi(\pi^+(\textsc{word})) = \frac{1}{4}1 + \frac{1}{4}\frac{2}{3} + \frac{1}{2}\frac{1}{6} = \frac{1}{2}$.

The credence mud diagrams:





← The grey rectangle in Ambi-Tails has the same area as the one in Ambiguous. So in Ambi-Reflect, the areas of the two grey rectangles cancel each other out, and so the total area under the rectangles is the same as in Conditionalize and Prior.

So even if expectation reflection and total trust are inequivalent, this setup is not a case where total trust *precludes* expectation reflection: being less wrong doesn't seem to require being more biased. This clears the way for 'don't be biased' as a possible rational requirement.

4. Don't Be Biased? (Further, More Tentative Thoughts)

Independent motivation for expectation reflection as a rational requirement: If HINT is (ambiguous) evidence in favor of WORD, then a lack of HINT is, to a proportional extent, evidence *against* WORD.

> **Proportional Sensitivity**: Choose two worlds $v, w \in W$ and ignore the others, and suppose that the prior has credence x in w *rather than* v.[6] Let $\pi^+_w(w) = x + \delta$. Then P^+ is **proportionally sensitive** to evidence if $\pi^+_v(w) = x - \frac{x}{1-x}\delta$ (in which case π trusts P^+ iff $\delta \geq 0$).[7]

If proportional sensitivity is satisfied at *every* pair of worlds, this is equivalent to expectation-reflection for every subset $W' \in W$:

$$\forall W', \forall X: \quad \mathbb{E}_\pi(X \mid W') = \mathbb{E}_\pi\left(\mathbb{E}_R(X \mid W') \mid W'\right). \qquad \text{(Subset ER)}$$

Conjecture: ***evidence balances for rational agents.*** What is a "proportional extent"? Whatever extent satisfies expectation reflection, as illustrated by the equivalence between Proportional Sensitivity and Subset ER, [8] and thus allows us to avoid polarization.[9]

[6] That is, $\pi(w \mid v, w) = x$. We conditionalize all the credence functions on $\{v, w\}$ to ignore the other worlds.

[7] This equation is basically conservation of credence mud across v, w.

[8] *Approximately:* I think expectation reflection needs at least to be *strengthened* to apply across any subsets of worlds, not just the set, and also *weakened* to be able to handle modest priors.

[9] At least in this setup: Kevin will give you some new ones in which this may be insufficient!

Avoiding Avoiding Avoiding Polarization (Replies to Kevin)

A1. The Strength of Proportional Sensitivity

How strong is proportional sensitivity? Pretty strong.

> ***Fact:*** Consider any good-case bad-case setup with two worlds b, g where it is rational to be certain of the good case in the good case ($\pi_g^+(g) = 1$). Then a prior π expectation reflects a posterior P^+ if and only if $\pi(b) = 0$ or $\pi_b^+(b) = 1$.[1]

That is: either the prior should be certain about being in the good case, or the posterior in the bad case should be certain about being in the bad case. Both cases are incompatible with asymmetry of good case/bad case. So *if proportional sensitivity is a rational constraint, then*:

1. *Certainty* good case/bad case scenarios are forbidden: it's not rational to become certain in the good case if there is some prior chance of the bad case.[2]
2. *Asymmetric* good case/bad case scenarios are forbidden: it's not rational to think that you might learn something and you might learn nothing: if you might $pr \in (0, 1)$ gain some evidence, then there is no possibility in which you gain no evidence.

So proportional sensitivity is quite a strong constraint, and I do have to take on the externalists more broadly if I want to defend it.

A2. Modest Priors, Unambiguous Evidence

This is a case Kevin first showed me back in January 2024 – I think it's really cool and have nothing concrete to say about it. However:

1. I wonder if when the prior is modest (as P is), conditionalization is no longer most rational response even if new unambiguous evidence comes in (the partition *tails&odd*, $\neg$[*tails&odd*]).
2. What would be the most rational response? Suppose there is some update where each $\pi_w \in P$ trusts P^+, and each $\pi_w \in P$ also expectation reflects P^+ conditional on being **informed** – on being certain that π_w is indeed the rational credences to have:[3]
$$\mathbb{E}_{\pi_w}(X) = \mathbb{E}_{\pi_w}\left(\mathbb{E}_R(X) \mid [P = \pi_w]\right). \qquad \text{(Informed ER)}$$
3. Conditionalizing on a modest prior does not (in general) satisfy Informed ER.[4]
4. And informed ER seems to better capture what we care about for polarization: we care about whether rational agents *actually* polarize, and Informed ER measures what agents' *actual* credences will be on average.
5. If P informedly expectation reflects P^+ and the prior π_0 over the full space expectation reflects P, then π_0 expectation reflects P^+.

[1] *Proof*: Expectation reflection is satisfied if $\pi(g) = \pi(g)\pi_g^+(g) + \pi(b)\pi_b^+(g)$. If $\pi_g^+(g) = 1$, then the equation reduces to $\pi(b)\pi_b^+(g) = 1$. And this is only satisfied if $\pi(b) = 0$ or $\pi_b^+(b) = 1$.

[2] So by ***Fact***, proportional sensitivity says something is wrong in **memory loss**. I don't have a good *error theory* for **memory-loss**: something seems fishy about memory loss (not irrational but *arational*?) but I don't have a good story of why Kevin's proposed response for this case would not be rational.

[3] This modification is needed because there is no prior P^+ that P expectation reflects, without the modification. For everything in the first three pages, the prior was already immodest, so the modification makes no difference.

[4] For example: $\pi_{he}(\textsc{heads}) = 2/3$ but $\mathbb{E}_{\pi_{he}}(P^+(\textsc{heads})) = \frac{1}{2}\frac{4}{5} + \frac{1}{2}\frac{4}{5} = 4/5 > 2/3$. Note that the polarization goes in the "correct" direction: this may be good for Kevin's case!

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, and audiences at the Formal Epistemology Workshop and the MIT Epistemology Reading Group in 2024.

Comments on Adrian Liu's, "Avoiding Polarization"

Kevin Dorst
kmdorst@mit.edu

Eastern APA
January 10, 2025

Fantastic paper.

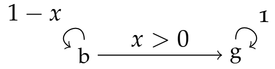
- Adrian might be right that 'ambiguity asymmetries' don't in themselves *generate* polarization
- Interesting case for 'only generate ambiguity *symmetrically*' constraint.
- Has pushed me to rethink which types of models/updates I think are the most compelling for the rationality of polarization.

At least in the models from the paper; TBD if this is ever right.

So: I've learned a lot from Adrian.

1) How strong is Proportional Sensitivity (Constraint 2) meant to be?

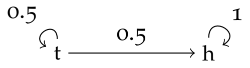
Good/bad cases (externalism):
The rational prior to have in *hands* is 0.9. But if you do have hands (good case, g), your evidence entails as much so the rational posterior is 1. And if you don't have hands (bad case, b), you can't be sure you're in the bad case, so the rational posterior is greater than 0.



$\mathbb{E}_P(P^+(g)) = 0.9(1) + 0.1(x) > 0.9 = P(g)$

Memory loss:
Every morning I toss a fair coin. (You remember this.) If today's landed heads, I'll send you all an email saying so on Jan 10, 2026. If not, I won't say anything—and you'll forget about this.

Future credences:



$\mathbb{E}_P(P+(h)) = 0.75 > 0.5 = P(h)$.

2) Does symmetric ambiguity generation prevent polarization?

Once we allow (non-polarizing) symmetric ambiguity, overlaying a bit of *clear* (partitional, conditioning) evidence can turn it polarizing.

Suppose (for simplicity) Haley knows she won't find a word, but will just get (symmetric!) hints. She starts with prior $\boldsymbol{\pi = (0.5, 0.5)}$ over (h, t), and her initial posterior is [»]:

$$P = \left(\begin{array}{c|cc} & h & t \\ \hline h & 2/3 & 1/3 \\ t & 1/3 & 2/3 \end{array}\right)$$

But I'll also roll a fair die, and she's 50-50 between even (e) and odd (o). So her initial posterior P over the *full space* is on the left.

And her prior over the full space is

$$\pi = \left(\begin{array}{cccc} he & ho & te & to \\ \hline 0.25 & 0.25 & 0.25 & 0.25 \end{array}\right)$$

$$P = \left(\begin{array}{c|cccc} & he & ho & te & to \\ \hline he & 2/6 & 2/6 & 1/6 & 1/6 \\ ho & 2/6 & 2/6 & 1/6 & 1/6 \\ te & 1/6 & 1/6 & 2/6 & 2/6 \\ to & 1/6 & 1/6 & 2/6 & 2/6 \end{array}\right) \qquad P^+ = \left(\begin{array}{c|cccc} & he & ho & te & to \\ \hline he & 2/5 & 2/5 & 1/5 & 0 \\ ho & 2/5 & 2/5 & 1/5 & 0 \\ te & 1/4 & 1/4 & 1/2 & 0 \\ to & 0 & 0 & 0 & 1 \end{array}\right)$$

(π, P) isn't polarizing. But suppose she knows that after, I'll tell her whether or not *tails&odd* is true.[1] Then her final posterior is P^+ (right).

And $\boldsymbol{(\pi, P^+)}$ ***is* polarizing**: $\mathbb{E}_\pi(P^+(h)) = 0.525 > 0.5 = \pi(h)$.

$\mathbb{E}_\pi(P^+(h)) = \frac{1}{4}(\frac{4}{5}) + \frac{1}{4}(\frac{4}{5}) + \frac{1}{4}(\frac{1}{2})$

Once we allow ambiguity, avoiding polarization is hard.

Interestingly, in this case the polarization goes in the *opposite* direction from the ambiguity-asymmetry—supporting Adrian's point.

[1] So she'll condition on the true cell of the partition, $\{\{he, ho, te\}, \{to\}\}$.