

Polarization

Kevin Dorst

24.805, Fall 2025

I. Empirical Features

Polarization \approx persistent disagreement despite plentiful evidence.

Polarization vs. convergence

Polarization is Normal: Even for factual matters where there's plentiful evidence and shared standards, given competing narratives or incentives, people *normally* polarize.

- Religion, history, art, science, politics, and who does more chores.
- Strong and common over factual matters; little evidence that it's driven by *purely* normative disagreements.
- The fact that Arguments Work suggests that people have (relatively) shared standards.

Often factual disagreements are larger, and usually normative disagreements mediated by factual ones.

Polarization is Persistent: Even after extensive discussion amongst open-minded and reasonable parties—and study of and reflection on its causes—polarization often remains.

- Often persists after extensive communication and reflection on ins and outs of disagreement.

Including amongst researchers who study polarization.

Polarization Isn't Inevitable: People often *do* converge to the truth: on small questions, and sometimes—when they agree on a shared paradigm—on big ones.

- On middle-sized dry goods—the everyday feats.
- On esoteric topics, in well-developed science.

II. Against Standard Bayes

Save 'against irrationalism' for Ch. 11

Standard-Bayes has trouble explaining why Polarization is Normal.

Bayesian convergence results often have to do with 'long run'.¹ Savage (1972)-style results assume (1) you leave open the true chance hypothesis and (2) draws are iid.

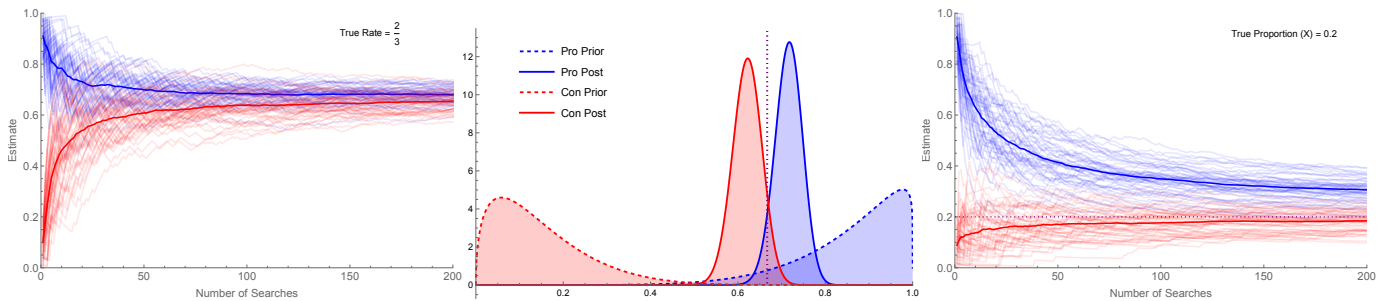
¹ Infinite evidence. Coin toss example.

But, in practice, even when conditions don't quite hold and data is limited, Standard Bayesians with reasonable priors often converge quite quickly, given a decent amount of (even quite noisy) evidence.

Eg uncertain what proportion of immigrants are net taxpayers.

1) Known random samples from true proportion: beta priors with clear outcomes. → Left and middle

2) Selective search with calibrated priors: fractional beta approximation with different directions of (clear) search. → Right



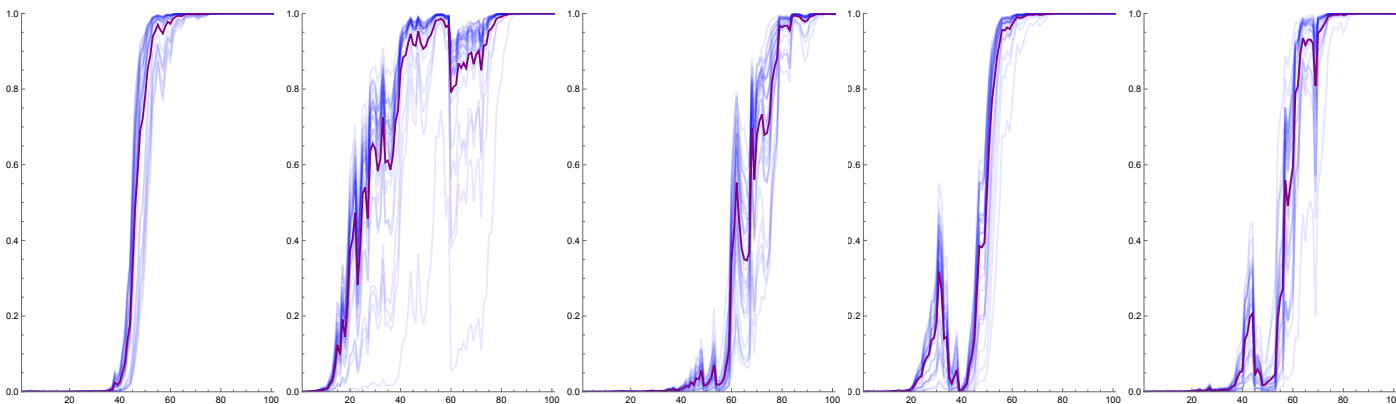
This generalizes. Suppose:

- 10,000 different chance hypotheses for generating different signals. Think of as combinations of (1) what proportion of are net taxpayers and (2) how might Fox report on immigrants and taxes?
- Each signal can be one of 100 different outcomes.
- Suppose 30 Bayesian agents who (1) defer to the chances² but (2) start with random priors over the chance hypotheses.
- Collectively observe 100 signals (rolls of the die).

100 different possible proportions (1%, 2%,...) and 100 different possible reporting strategies $\rightsquigarrow 100 \times 100 = 10,000$ different chance hypotheses.

So like rolling a 100-sided die, with 10,000 different possible weightings ² i.e. obey Principal Principle

Plot individual and average credences in *true* chance hypothesis:

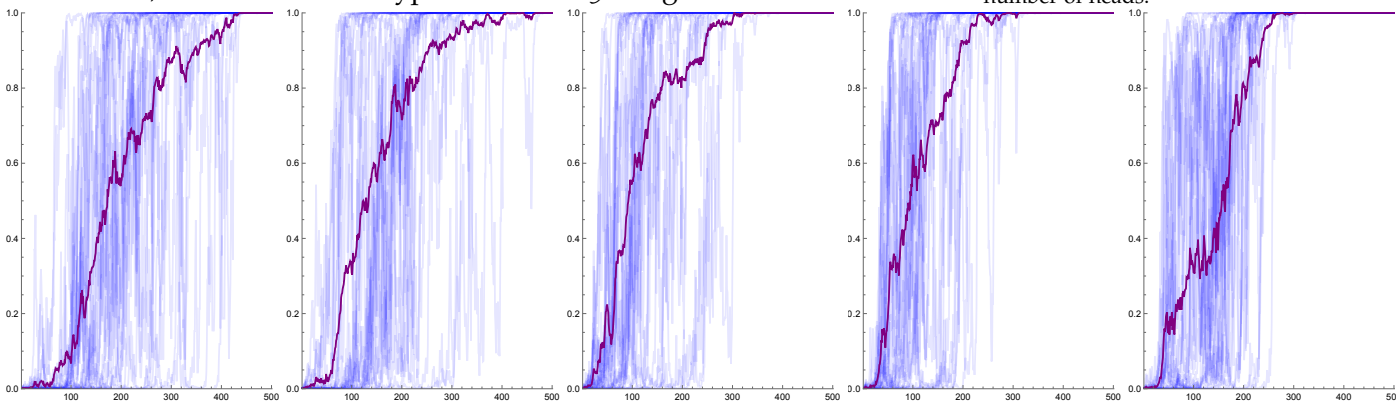


What if likelihoods are noisy?

When get e_i , update as if $P_a(e_i|ch = \pi)$ is a noisy indicator of $\pi(e_i)$.

If $\pi(e_i) = x$, flip an x -biased coin 100 times and update as if likelihood is the number of heads.

This time 1,000 total chance hypotheses and 500 signals:



Standard-Bayes has trouble explaining why Polarization is Persistent.

Standard Bayesians with commonly-known shared priors cannot 'agree

to disagree', i.e. have common knowledge of their posteriors about q and still disagree.

Agreement Theorem (Aumann 1976) Let W be finite, and each of (W, P, P^+) and (W, C, C^+) be Standard Bayesian. Suppose there's a π such that for all w , $P_w = \pi = C_w$. Then if $\langle P^+(q) = x_1 \rangle \& \langle C^+(q) = x_2 \rangle$ is common knowledge at any world v , then $P_v^+(q) = x_1 = x_2 = C_v^+(q)$.

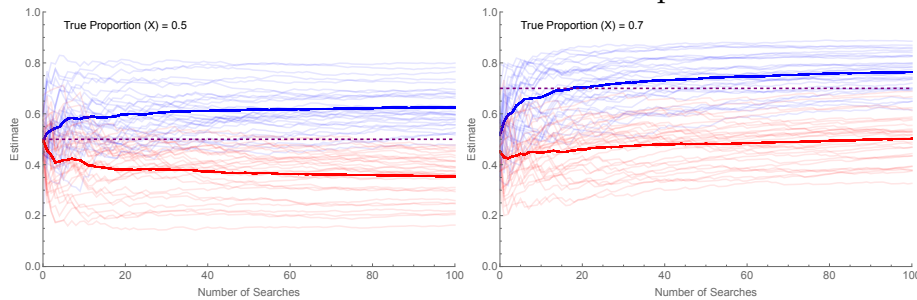
Even when strict conditions aren't met (we might have had somewhat-different priors; we might not have been perfect Bayesians), still that the other person is confident of q is good evidence for q , and vice versa.

→ Hard to explain disagreement in scientific contexts, where standards are broadly shared and data, reasoning, etc. can all be shared.

III. Ambiguous Bayes

Polarization is Normal

Fractional betas with selective search + calibrated priors:



Polarization is Persistent

Ambiguous-Bayesians can agree to disagree.

→ We can know we share a prior, but be unsure what it is, and have common knowledge that learning will lead to different posteriors.

Inspired by Lederman 2015. Same idea, but his models aren't factorable.

Order worlds as $\{w_1, w_2, w_3, w_4\}$.

Let $q = \{w_2, w_4\}$. $Q_1 = \{\{w_1, w_2\}, \{w_3, w_4\}\}$, and $Q_2 = \{\{w_1, w_4\}, \{w_2, w_3\}\}$.

$$P = \begin{pmatrix} .05 & .45 & .45 & .05 \\ .01 & .09 & .09 & .81 \\ .01 & .09 & .09 & .81 \\ .45 & .05 & .05 & .45 \end{pmatrix} \quad P^{Q_1} = \begin{pmatrix} .1 & .9 & 0 & 0 \\ .1 & .9 & 0 & 0 \\ 0 & 0 & .1 & .9 \\ 0 & 0 & .1 & .9 \end{pmatrix} \quad P^{Q_2} = \begin{pmatrix} .5 & 0 & .5 \\ 0 & .5 & .5 & 0 \\ 0 & .5 & .5 & 0 \\ .5 & 0 & .5 \end{pmatrix}$$

Start out unsure whether your credence in q is 0.5 or 0.9.

If you learn answer to Q_1 , it'll definitely go to 0.9.

If you learn answer to Q_2 , it'll definitely go to 0.5.

So if I learn Q_1 and you learn Q_2 , we'll agree to disagree!

Shit's getting crazy...

What about in case iterating Martingale failures with independent Q_i ?

Take simple search model, where $P_a(c_i) = 0.5$ and $\mathbb{E}_{P_a}(P^+(c_i)) = 0.55$.

While you (Con) search opposite direction, so $\mathbb{E}_{P_a}(C^+(c_i)) = 0.45$. Suppose we know we're clones.

My posterior now thinks around 55% of them are true.
 And it now thinks *you'll* think that around 47.5% of them are true—
 more of them are true than I thought, so I expect your estimate is higher
 than I initially expected it. But I still expect it to have gone down, so
 I'm unsurprised that you're confident that less than half landed heads.

What if we start sharing our evidence?

If in fact 50% are true, then *in fact* we'll de-polarize.

But neither of us expect that. I expect doing so to reveal that around
 55% landed heads.

Polarization Isn't Inevitable

Predictions:

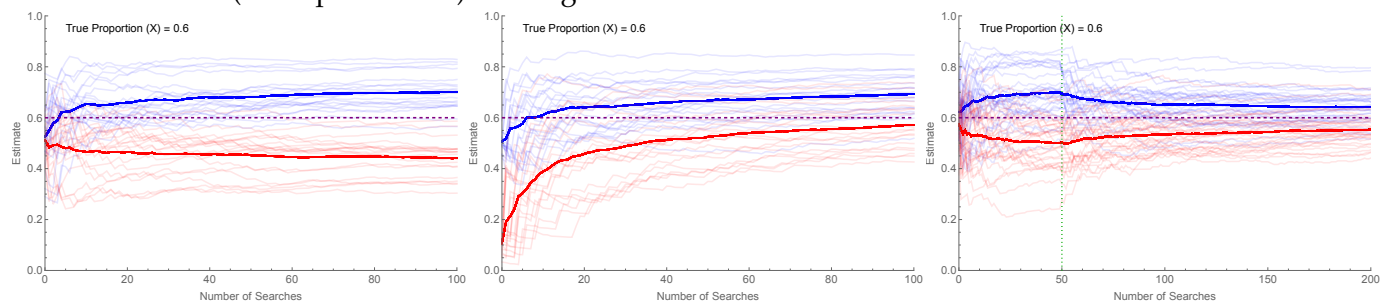
- 1) Under ambiguity and differential search, polarization.
- 2) Under ambiguity and univocal search, agreement but not truth.
- 3) Under clarity, should see convergence *regardless* of search direction.

Kuhn on paradigm shifts. Mysticism about anti-realism and no need
 for 'truth' in his explanation of science.

Need to distinguish *unifying* vs. *clarifying* paradigm shifts.

→ Former leads to (often temporary) agreement.

→ Latter lead to (often permanent) convergence to truth.



Sometimes less evidence is better, if it's clear and sufficiently probative.

→ Significance testing and meta-analyses: clarifying what we expect.

Is this reasonable?

When ambiguity is Foxing can improve convergence. But:

- Diachronic tragedy.
- Must balance it perfectly, depending on actual degree of ambiguity-
 asymmetry—but you don't know that!

Learning *exactly* what you think (say, that 45% are true) would be surprising; but of course even you don't know exactly what you think. Our disagreement itself isn't surprising.

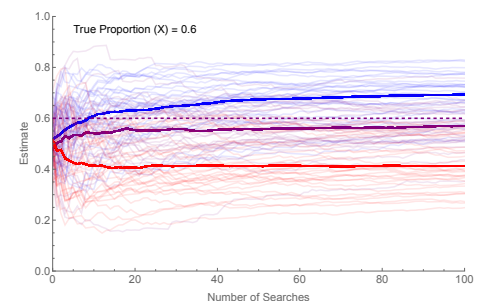
Sharing would bring clarity, i.e. \hat{P} , and since I obey Informed Reflection, my credence equals my expectation of \hat{P} .

Politics, philosophy

Fads

Established science.

'Less is More'



References

Aumann, R, 1976. 'Agreeing to Disagree'. *The Annals of Statistics*, 4:1236–1239.

Lederman, Harey, 2015. 'People with Common Priors Can Agree to Disagree'. *The Review of Symbolic Logic*, 8(1):11–45.