Rational Polarization

Kevin Dorst
kevindorst@pitt.edu

CPS Lunchtime Talk
March 5, 2021

I. A Standard Story

Societal polarization is *profound* and *persistent*.

But it's also *predictable*: your choices and circumstances have a predictable effect on which direction your opinions will shift.

Where to live? What to read? Who to follow? How to engage?

Not new. But increasing? **Ask me!**

Standard Story: This is driven by epistemically *ir*rational processes. Motivated reasoning, confirmation bias, conformism, etc.

Kunda (1990); Klein (2020); Nickerson (1998); Taber and Lodge (2006); Axelrod (1997); Sunstein (2009), etc.

New Story: Some kinds of evidence are *more ambiguous*—harder to know how to react to—than others. Our choices lead to ambiguity asymmetries, which in turn lead *rational* people to polarize.

Familiar fact: ambiguity leads to biased processing (Petty and Wegener 1998). New idea: the "bias" can be *in the evidence*, rather than the person.

Claim: This is both theoretically possible and empirically plausible.

Focus more on latter, today.

II. A Theoretical Possibility

Started in an unlikely place...

Idea: your evidence is **ambiguous** iff it's rational to be unsure how confident to be in response to it (warrants higher-order uncertainty).

Let P be the current rational probability function; iff $\exists q \forall t$: $P(P(q) = t) < 1$.

Evidence is (rationally) **predictably polarizing** about q iff you should expect it to move the rational opinion in a particular direction.
⇒ Starting with same beliefs, you and I can expect to diverge.

$\vec{P}$ = future rational probabilities.
$\mathbb{E}(\vec{P}(q))$ = current rational estimate of future rational probability in q.
Pred. polarizing on q: $P(q) \neq \mathbb{E}(\vec{P}(q))$.

Evidence is **valuable** iff, no matter what choice you face, you should prefer to use the evidence to help guide your decision.

You should *want* evidence. Good 1967; Geanakoplos 1989; Dorst 2020; Dorst et al. 2021.

Fact 1. Suppose evidence is valuable. Then if it's *un*ambiguous, it's *never* (rationally) predictably polarizing.

Best arg. for Standard Story. **Ask me!** van Fraassen 1995; Kadane et al. 1996; Briggs 2009; Huttegger 2014

Fact 2. If evidence is ambiguous, then—even if it's valuable—it's *always* (rationally) predictably polarizing (about some q).

Generalization of Salow 2018

Intriguing... But so what?
Example: *Word-completion task* generates asymmetric ambiguity.

Three Qs: Why polarizing? Why valuable? Would it work?

Why Predictably Polarizing?

It's easier to recognize that *there is* a completion than to recognize that there's *no* completion ($\exists$ vs. $\forall$). So:

- If there is a word, you should (on avg.) be confident there is.
- If there's not, you *shouldn't* be very confident there's not.

You won't find one; but you should be unsure whether you *should* find one.

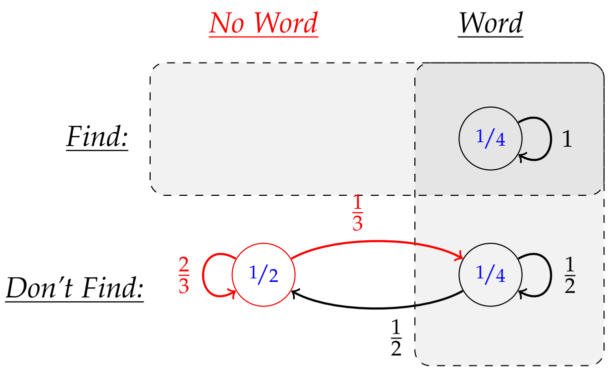


Diagram Key:

Blue numbers = prior probabilities.

Labeled arrows from world w to x = posterior probabilities at w of x.

If *Don't Find*, evidence is ambiguous: unsure whether to be $\frac{1}{3}$ or $\frac{1}{2}$ confident there's a word.

Prior confidence there's a word: $1/2$.

Prior estimate of *future* rational confidence?

- If there's a word, on average $\frac{3}{4}$ confident.
- If not, $\frac{1}{3}$ confident. $\Rightarrow$ Average $> \frac{1}{2}$.

Might go way up; won't go way down.

Half the time 1, half the time $\frac{1}{2}$

Likewise for all models like this.

Why Valuable?

Notice that posterior probability is more accurate in every world.

More centered on the actual world, whatever it is.

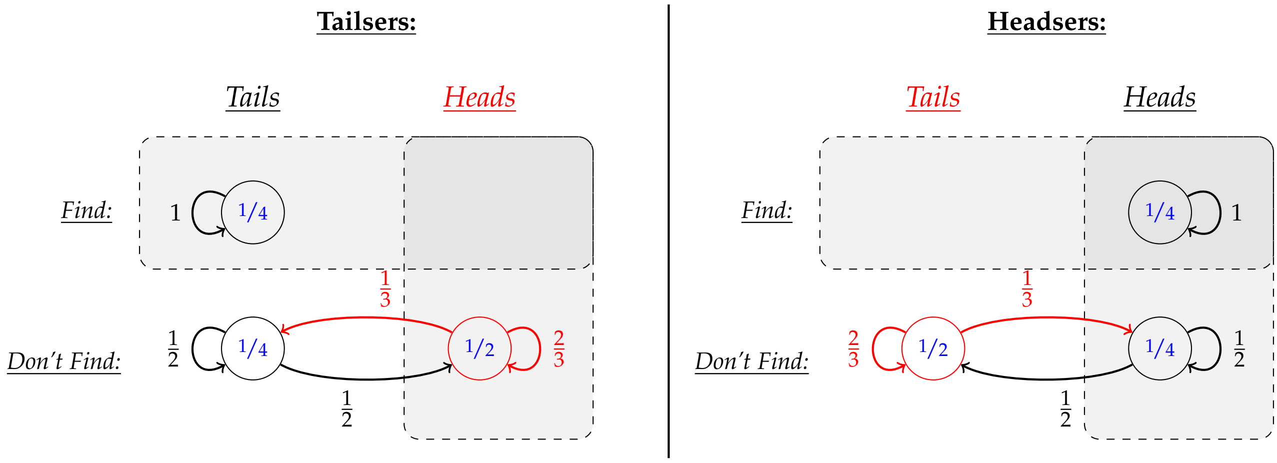
So accuracy always increases. It just increases *asymmetrically*: increase is greater if the string is completable than if not.

Ask me! If we repeat with many WC tasks, *at each stage* the evidence is valuable, yet you can predict with confidence that, overall, you'll end up polarized. (Diachronic tragedy.)

Would it work?

Did this. Divide into two groups: Headsers vs. Tailsers. Show different WC task; Headser's would be completable iff coin landed H, Tailsers would be iff coin landed T.

Induces two mirror-image models:



Prediction: Headsers better at recognizing H; Tailsers better at recognizing T.

So Hsers end up on avg. more confident of H than Tsers.

It worked.

In 6 of 7 presentations.

III. An Ambiguous Effect

Pre-registration:
https://aspredicted.org/8jg3e.pdf

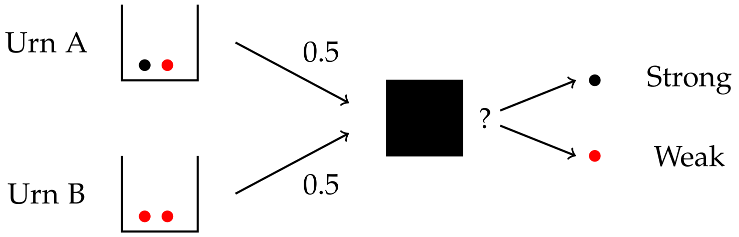
Ambiguous evidence $\neq$ weak evidence.

Ambiguous evidence is evidence for which *it's hard to know how weak it is*.

Urn A: 1 black, 1 red. Urn B: 2 red. Chosen randomly.
Drawing a red marble is weak *but unambiguous* evidence for B.

If see *black*: $P(B|black) = 0$, and $P(P(B|black) = 0) = 1$.

If see *red*: in 2 of 3 of possibilities, it's B. So $P(B|red) = \frac{2}{3}$, and moreover $P(P(B|red) = \frac{2}{3}) = 1$.



Gallow's Challenge: What if it's not ambiguity, but just that people under-react to weak evidence?

Conservatism (Edwards 1982). **Ask me!**

Prediction: ambiguity *exacerbates* polarization.

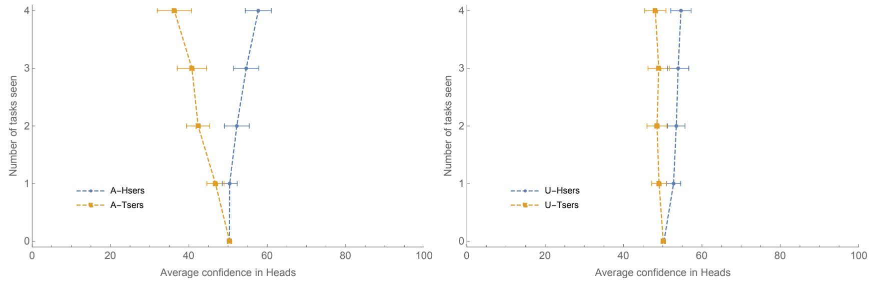
Setup: Divide between Hsers and Tsers: Hsers get strong evidence when H, weak evidence when T; Tsers vice versa. Two conditions:

- *Ambiguous condition:* evidence comes as WC task.
- *Unambiguous condition:* evidence comes as draw from an urn of unknown composition.
  → Hsers: if H, 1 black and 1 red; if T, 2 red. (Tsers: vice versa.)

Prediction 1: Mean posterior credence in H polarizes in Ambiguous condition.

Prediction 2: It polarizes *more* in Ambiguous than Unambiguous condition.

Both confirmed:
1: $t(101) = 7.98$, $p < 0.001$, $d = 1.58$.
2: 2x2 ANOVA interaction effect $p < 0.001$; empirically bootstrapped 95% CI for diff of diffs, (A-Hser – A-Tser) – (U-Hser – U-Tser), is $[7.19, 22.59]$.



Upshot: Asymmetric ambiguity *could* drive real-world polarization. Does it?

Other details: **Ask me!**

IV. A Confirmed Bias

Confirmation bias: tendency to seek and interpret evidence in way that favors your prior beliefs (Nickerson 1998; Whittlestone 2017).

Focus on "interpret" side, aka *biased assimilation*. **Ask me!**

D = capital punishment has Deterrent effect. Present two bits of evidence. Those who believed D took them to support it; those who didn't, didn't.

Lord et al. 1979; Taber and Lodge 2006

Mechanism: *selective scrutiny*. Spend more time scrutinizing incongruent study; often find flaws in it.

Claim: this is rational.

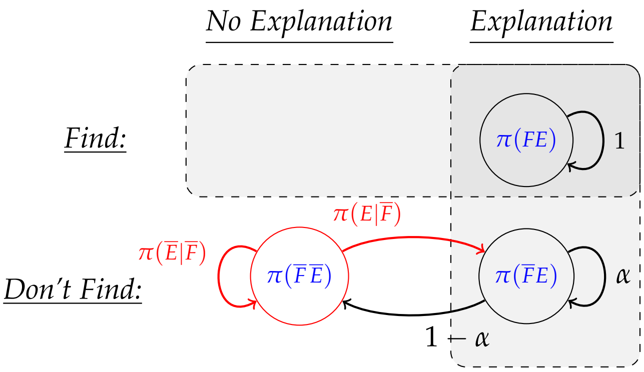
Inspired by / reacting to Kelly 2008. **Ask me** about differences.

Scrutinizing a study is a form of *cognitive search.*

Like a word-completion task.

- If there's an alternative explanation, can get unambiguous evidence there is.
- If there's not, can only get ambiguous evidence that there's not.

Prior = π. Is there an Explanation (E)? Will you Find one (F)?



Cognitive Search Model:

$\pi(D|FE) = \pi(D|\overline{F}E)$. (Higher than $\pi(D)$ if E explains a disconfirming study; lower if confirming one.)

$\alpha \in [\pi(E|\overline{F}), 1]$, to potentially engender ambiguity.

Fact 3: Any model of this structure is both valuable and predictably polarizing. → Same reason as in word-completion task.

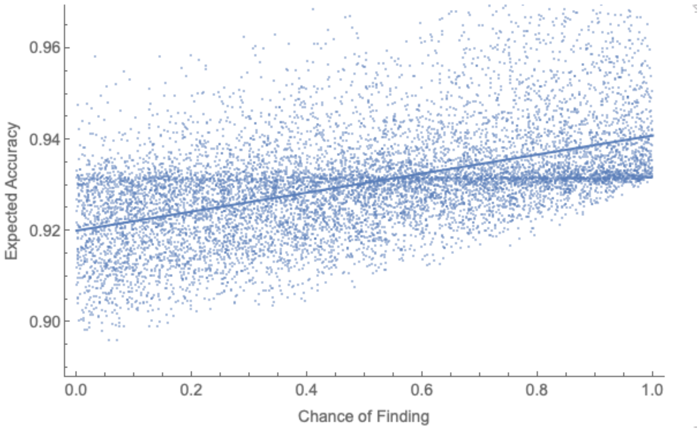
If E would explain away disconfirming study, then expected *rise*; if would explain confirming study, expected *drop*.

What drives *choice* of which to scrutinize? Get accurate beliefs! So avoid ambiguity. So scrutinize the one where you're more likely to *find* an explanation if there is one—the incongruent study.

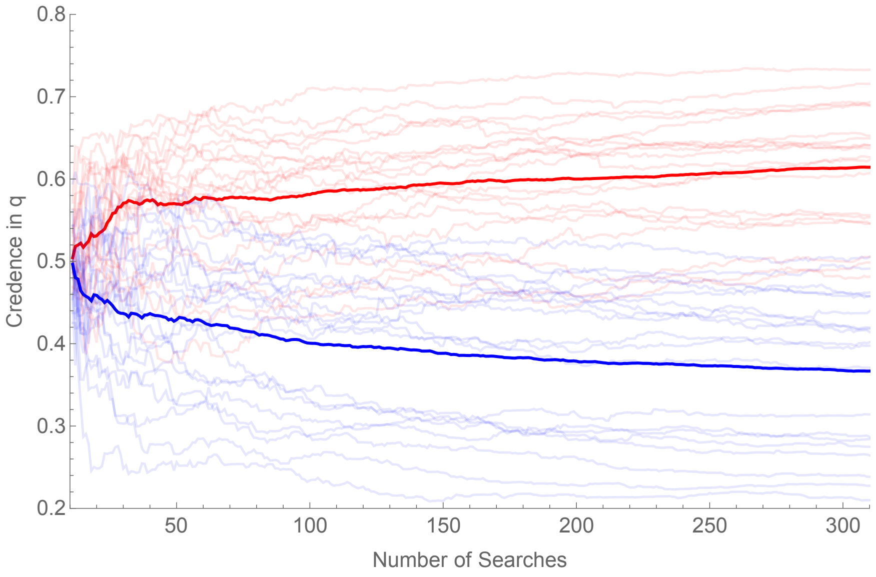
Part of being convinced of D is learning how to rebut arguments against it.

Correlation between chance of finding an explanation if there is one, and expected accuracy of doing the search:

Randomly generated models; correlation between $\pi(F|E)$ and expected accuracy of posterior (Brier score).



Two groups of agents, one (red) is better at explaining disconfirming studies; the other (blue) is better at explaining confirming studies. Choose which to scrutinize based on expected accuracy:



A single (but representative) run.

Thin red = individual pro agents (20)
Thick red = average of pro agents
Thin blue = individual con agents (20)
Thick blue = average of con agents

Upshot: Confirmation bias can be driven by a rational attempt to get accurate beliefs in the face of the ambiguity.

V. A Clarified Argument

Group Polarization Effect: Discussion amongst like-minded people tends to lead them to become *more extreme* in their opinions.

AKA **enclave deliberation**
Myers and Lamm 1976; Isenberg 1986

Mechanism: you receive more arguments favoring your position—which tend, on average, to persuade!

At least if engaged with openly.
Ask me about scrutinizing arguments.

Why? Arguments can't *guarantee* a rise in credence.

If the argument is worse than you expected, should lower your credence.

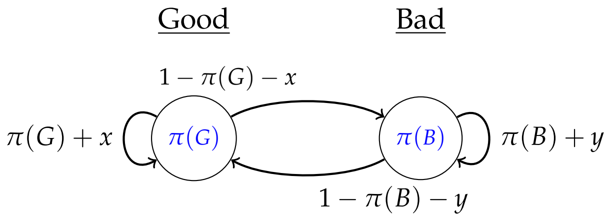
But what they *can* do is make is easy to recognize favorable reasons and hard to recognize unfavorable ones.

They can manipulate ambiguity.

Example: "All the victims friends came to the party. As we know, my defendant was at the party—so he was a friend."
vs. "All those who came to the party were the victim's friends. As we know, my defendant was at the party—so he was a friend."

People are worse at recognizing (tempting) fallacies than analogous validities (Cariani and Rips 2017, Fig. 1).

Simple model:

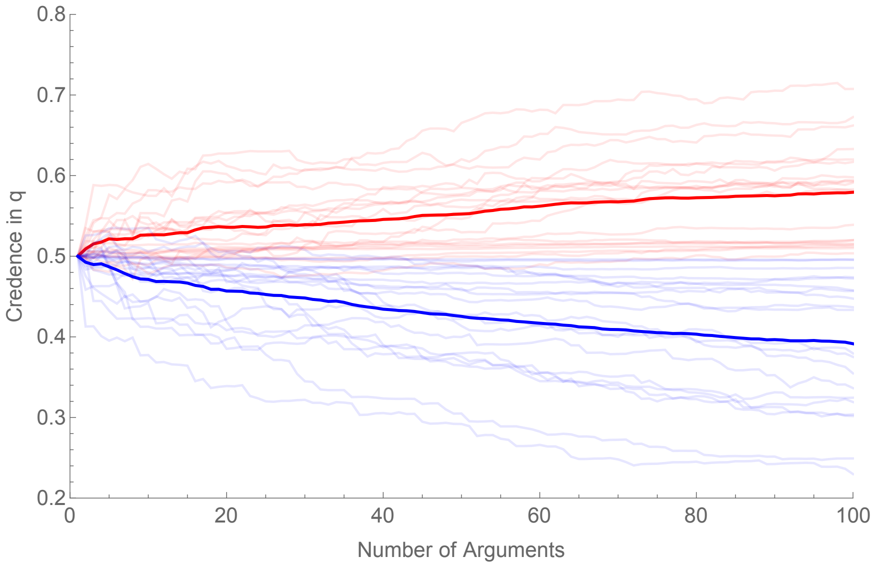


If arg in favor of q, $\pi(q|G) > \pi(q)$;
if arg against q, $\pi(q|G) < \pi(q)$.

Since Bad more ambiguous, $y \leq x$.

Valuable because probabilities shifting towards truth.
Polarizing because shift if *Good* is greater than shift if *Bad*.

Split into two groups. One (red) receives arguments favoring q; the other (blue) receives arguments against q. Polarized:



A single (but representative) run.

Thin red = individual pro agents (20)
Thick red = average of pro agents
Thin blue = individual con agents (20)
Thick blue = average of con agents

Upshot: The group polarization effect can be driven by rational sensitivity to ambiguity-asymmetries in arguments.

VI. A Needed Story

New Story: A rational sensitivity to ambiguous evidence plays a significant role in driving predictable polarization.

This story has a firm theoretical foundation, fits with old and predicts new empirical findings, and plausibly plays a role in helping explain some of the core mechanisms of polarization.

That is how I became predictably (rationally!) polarized about the possibility of rational polarization.

Notice: a story for the opposite conclusion would be incoherent.

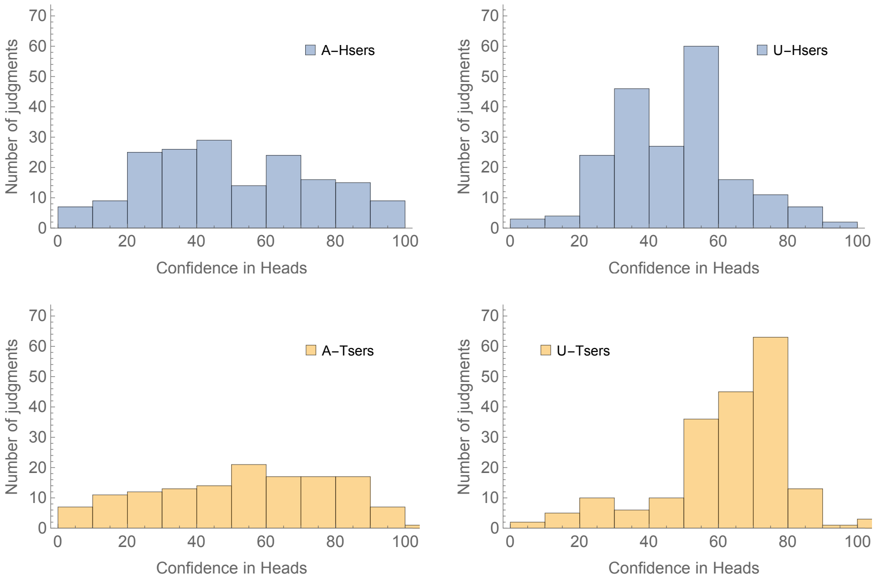
- 'Polarization is rational, and I came to believe that through rational polarizing mechanisms' ✓ (Coherent)
- 'Polarization is *ir*rational, and I came to believe that through *ir*rational polarizing mechanisms' ✗ (Akratic)

⇒ If you want to hold on to your predictably-polarized beliefs, you'd better buy into rational polarization!

VII. The Bonus Material

Experiment

Can see difference in ambiguity in the weak-evidence cases:



Left: Ambiguous; cases where didn't find a word (non-extreme credence).

Right: Unambiguous; cases where didn't see black marble (non-extreme credence).

Differences in variances significant at $p < 0.001$ (Conover).

Average confidence that it landed heads across cases:

	A-Hsers	A-Tsers	U-Hsers	U-Tsers
Overall:	57.7	36.29	54.64	48.10
Heads cases:	67.42	47.73*	66.89	59.95
Tails cases:	48.00*	24.84	42.39	36.25

* = not significantly different from 50 (= the prior confidence).

Note: average posterior is more accurate than prior!

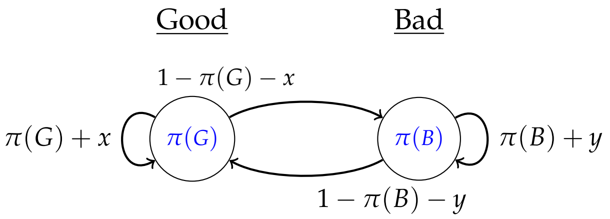
Further prediction: Ambiguous condition, when people don't find a word (= non-extreme credence), their confidence that there's a word is on average higher *if there is one* than if not.

Confirmed: 44.6 vs. 52.3; $t(309) = 2.77$, $p = 0.0030$, $d = 0.32$.

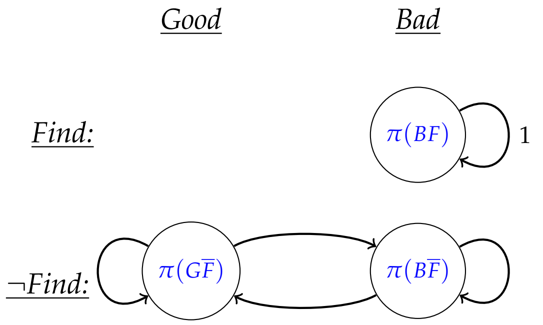
Scrutinizing Arguments (Combined Model)

Do you engage with an argument uncritically, or scrutinize it for flaws? Turns argument model into a cognitive search!

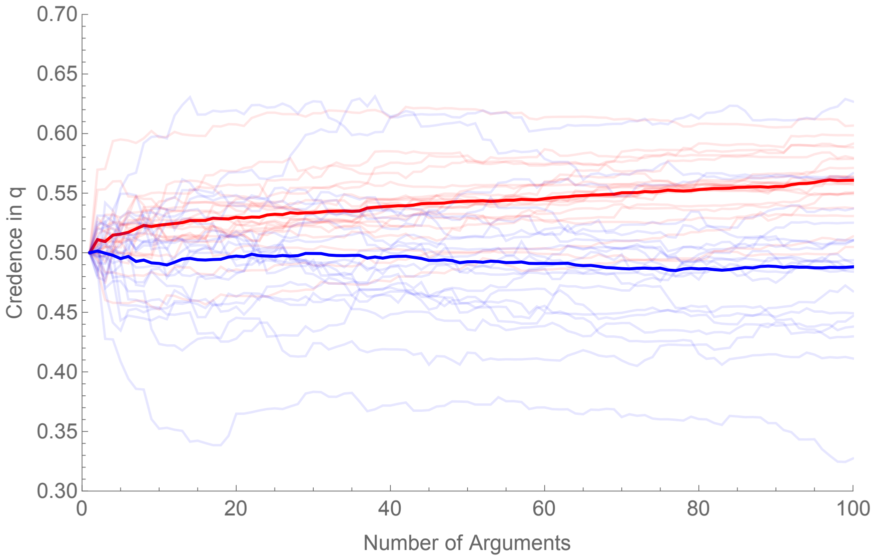
Don't Scrutinize:



Scrutinize:

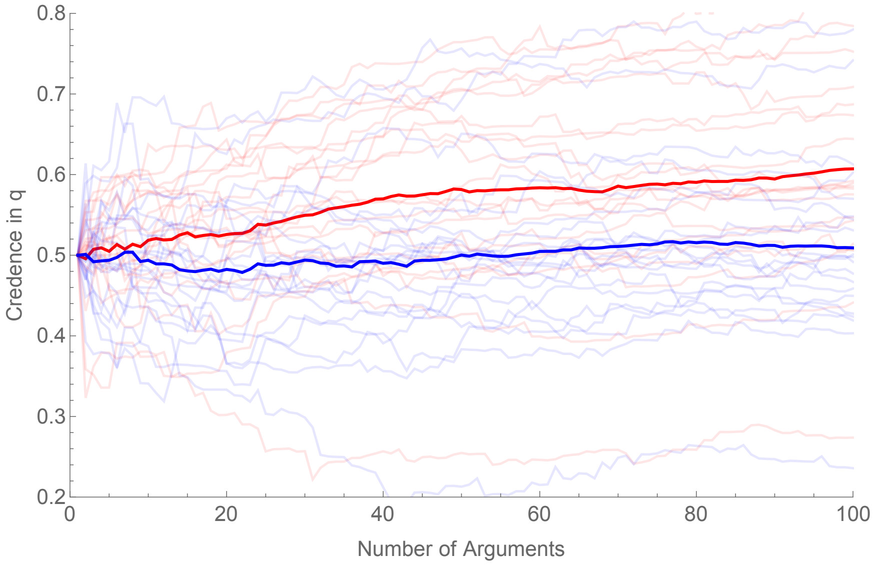


Presented with q-favoring arguments. Red (pro) never scrutinize, while blue (con) always do:



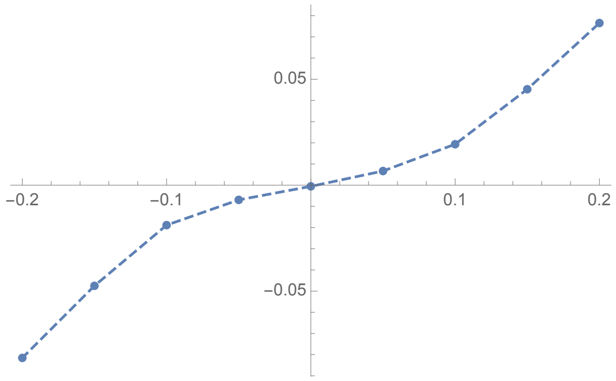
Chance of scrutiny proportional to how likely the argument is to be Bad? Red (pro) think less likely than blue (con):

Lot's of noise, but this direction of divergence is the trend.



Increasing Polarization?

Randomly generate valuable, ambiguous-evidence models. Measure degree of degree of ambiguity-asymmetry. Plot against expected shift in opinion on q:



***x*-axis:** degree to which q-favoring evidence less ambiguous than q-disfavoring evidence.
***y*-axis:** average expected shift on q, $\mathbb{E}(\vec{P}(q)) - P(q)$.

Increased social and informational sorting (Mason 2018; Klein 2020) may increase ambiguity-asymmetries, and so increase polarization.

References

Axelrod, Robert, 1997. 'The Dissemination of Culture'.

Briggs, R., 2009. 'Distorted Reflection'. *Philosophical Review*, 118(1):59–85.

Cariani, Fabrizio and Rips, Lance J., 2017. 'Conditionals, Context, and the Suppression Effect'. *Cognitive Science*, 41(3):540–589.

Dorst, Kevin, 2020. 'Evidence: A Guide for the Uncertain'. *Philosophy and Phenomenological Research*, 100(3):586–632.

Dorst, Kevin, Levinstein, Benjamin, Salow, Bernhard, Husic, Brooke E., and Fitelson, Branden, 2021. 'Deference Done Better'. *Philosophical Perspectives*, To appear.

Edwards, Ward, 1982. 'Conservatism in Human Information Processing'. *Judgment under Uncertainty: Heuristics and Biases*, 359–369.

Geanakoplos, John, 1989. 'Game Theory Without Partitions, and Applications to Speculation and Consensus'. *Research in Economics*, Cowles Fou(914).

Good, I J, 1967. 'On the Principle of Total Evidence'. *The British Journal for the Philosophy of Science*, 17(4):319–321.

Huttegger, Simon M, 2014. 'Learning experiences and the value of knowledge'. *Philosophical Studies*, 171(2):279–288.

Isenberg, Daniel J., 1986. 'Group Polarization. A Critical Review and Meta-Analysis'. *Journal of Personality and Social Psychology*, 50(6):1141–1151.

Kadane, Joseph B., Schervish, Mark J., and Seidenfeld, Teddy, 1996. 'Reasoning to a foregone conclusion'. *Journal of the American Statistical Association*, 91(435):1228–1235.

Kelly, Thomas, 2008. 'Disagreement, Dogmatism, and Belief Polarization'. *The Journal of Philosophy*, 105(10):611–633.

Klein, Ezra, 2020. *Why We're Polarized*. Profile Books.

Kunda, Ziva, 1990. 'The case for motivated reasoning'. *Psychological Bulletin*, 108(3):480–498.

Lord, Charles G., Ross, Lee, and Lepper, Mark R., 1979. 'Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence'. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

Mason, Lilliana, 2018. *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Myers, David G. and Lamm, Helmut, 1976. 'The group polarization phenomenon'. *Psychological Bulletin*, 83(4):602–627.

Nickerson, Raymond S., 1998. 'Confirmation bias: A ubiquitous phenomenon in many guises.' *Review of General Psychology*, 2(2):175–220.

Petty, Richard E. and Wegener, Duane T., 1998. 'Attitude change: Multiple roles for persuasion variables'. *The handbook of social psychology*, 323–390.

Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.

Sunstein, C, 2009. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press.

Taber, Charles S and Lodge, Milton, 2006. 'Motivated Skepticism in the Evaluation of Political Beliefs'. *American Journal of Political Science*, 50(3):755–769.

van Fraassen, Bas, 1995. 'Belief and the problem of Ulysses and the sirens'. *Philosophical Studies*, 77(1):7–37.

Whittlestone, Jess, 2017. 'The importance of making assumptions : why confirmation is not necessarily a bias'. (July).