

# Confirmation Bias and Motivated Reasoning

Kevin Dorst

24.805, Fall 2025

## Kelly 2008

### I. Biased Assimilation

What happens when people with different opinions are both presented with mixed evidence?

**Biased Assimilation:** People are inclined to interpret mixed evidence in a way that favors their prior beliefs.

Example: we disagree over *Deterrent*; both presented with two studies, one favoring and one disfavoring.

$S_1$ : States A and B are next to each other; A has capital punishment, B does not; A has lower murder rate.

$S_2$ : States C and D are next to each other; C has capital punishment, D does not; they have the *same* murder rate.

Result? I increase my confidence in *Deterrent*; you decrease yours.

"Belief polarization", in Kelly's terminology.

Avg; only under certain conditions (non-obvious evidence, strong priors); can be hard to measure effects...

*Deterrent* = capital punishment has a deterrent effect

### II. Psychological story

People don't ignore/dismiss incongruent evidence. Instead, they engage in **selective scrutiny**.

E.g. I take  $S_1$  at face value, while you scrutinize it and realized *B* has more poverty than *A*. Meanwhile, you take  $S_2$  at face-value, while I scrutinize and realize *C* has more poverty than *D*.

Searching for *alternative explanations*.

### III. Normative story

Kelly argues that most people are *unaware* of this general tendency for selective scrutiny. Three questions:

**Q1:** Is selective scrutiny reasonable?

**Q2:** *If* we do so, is the resulting polarization reasonable?

**Q3:** Does it remain reasonable once we become *aware* of this process?

Kelly: Yes, Yes, No.

Q1: Is selective scrutiny reasonable?

Kelly says this is a question about *practical* rationality.

Analogy: science is *anomaly-driven*.

Time- and resource-constraints.

Likewise, says Kelly, with investigation generally. Unreasonable to demand equal scrutiny for surprising vs. unsurprising bits of evidence.

"I need to let my dog out" vs. "I need to let my elephant out".

## Q2: Is resulting polarization rational?

*Key Epistemological Fact:* How confident you should be of a hypothesis depends on the available alternatives.

- These alternative explanations are part of your “broad evidence”.

So given that you have an alternative for  $S_1$  and not  $S_2$ , you are rational to lower confidence in Deterrent. Vice versa for me.

*Worry: commutativity.* Intuitively, the order in which you receive evidence shouldn't which beliefs you form.

*Does Kelly's proposal violate commutativity?* If you first get evidence  $e_1$  that convinces you of  $q$ , and then you are presented with  $e_2$  you'll explain it away and maintain belief.

Notice: Commutativity is easy to state, but near-impossible to find real-world violations of.

*Case 1:* Me: “I'm hungry. Also, I'm thirsty.” What's your credence in ‘Kevin told me he was hungry before he told me he was thirsty’.

*Case 2:* Me: “I'm thirsty. Also, I'm hungry.” What's your credence in ‘Kevin told me he was hungry before he told me he was thirsty’?

Generically: given that we're *aware* of the order evidence came in, (1) we can never<sup>1</sup> get the exact same evidence in different orders, and (2) these differences can affect what we should believe.

Kelly: Order effects what *other* pieces of evidence you gather, and so affects what total evidence you end up with.

→ If get  $S_1$  first, end up with  $S_1 \wedge S_2 \wedge \text{explanation-of-}S_2$ .

→ If get  $S_2$  first, end up with  $S_2 \wedge S_1 \wedge \text{explanation-of-}S_1$

So if you're unaware of the selective scrutiny effect (it's not part of your broad evidence), then you end up with different total (broad) evidence, which point in different directions, in the two cases.

## Q3: How should learning about this process affect our beliefs?

But if you're *aware* of this process of selective scrutiny, you should realize it's no accident that you ended up with the alternative-explanations you did. Now in the two cases you end up with:

- $S_1 \wedge S_2 \wedge \text{explanation-of-}S_2 \wedge \text{selectively-scrutinized-}S_2$
- $S_1 \wedge S_2 \wedge \text{explanation-of-}S_1 \wedge \text{selectively-scrutinized-}S_1$

And, he seems to say, *this* should lead to the same credence.

That's wrong. *Compare:* two envelopes; the first *may or may not* contain further info about  $S_1$ ; the second may contain further info about  $S_2$ . I open envelope 2 it says ‘there was a selection effect’. I now have an explanation of  $S_2$ . Even though I know I selectively scrutinized (I didn't

E.g. design vs. natural selection.

Vs. “narrow evidence”  $\approx$  “data”

Standardly claimed true for Bayesians who update by conditioning:  
 $P(q|e_1 \& e_2) = P(q|e_2 \& e_1)$ .

But if you first get  $e_2$ , vice versa!

What's your credence in ‘Kevin's more hungry than he is thirsty’?

What's your credence in ‘Kevin's more thirsty than he is hungry’?

<sup>1</sup> without memory loss

open envelope 1), I should still increase my credence in *Deterrent*.

Generally: so long as selective scrutiny doesn't entail finding an explanation,<sup>2</sup> *explanation-of-S<sub>2</sub> ∧ selectively-scrutinized-S<sub>2</sub>* will still weaken *S<sub>2</sub>*.

**Better:** If you know you're selectively scrutinizing, you can't expect it to make your opinion about *Deterrent* more extreme.

Suppose I have credence 0.5 the envelope contains an explanation of *S<sub>2</sub>*, and 0.5 that it contains a Haiku. Suppose seeing an explanation *E* would raise my credence in *D* by 0.2. If I'm a Bayesian, it follows that getting a Haiku would lower my credence in *D* by 0.2

So what's my best estimate for my future credence in *D*? I know I'll either learn *E* or  $\neg E$ , so 0.5 likely it'll go up by 0.2 and 0.5 it'll go down by 0.2; this averages out to 0 net change—I don't expect, on average, for scrutinizing to make me more confident.

### Williams 2023

Does partisan cognition—differing reasoning and conclusions of different partisans—come from *directional* or *epistemic* goals.

→ Practically- vs. epistemically-rational explanations.

Existing evidence is equivocal:

- Correlation between beliefs and party's conclusions?
  - Confounded by social networks and info environments
- Seeking out congenial info selectively?
  - Confounded by assessments of trustworthiness
- 'Matched-info' experimental designs?
  - Party-cue designs?
    - Confounded by trust, and our epistemic inter-dependence.
  - Biased-assimilation designs?
    - Confounded by priors
  - Motivated-numeracy designs?
    - Confounded by priors and therefore 'belief bias'—people find it harder to process weird claims.

Williams doubts that experiments will settle it—underdetermination.

Proposal: turns on broader theoretical virtues.

### **Coalitional Press Secretaries**

Groups tend to form big coalitions. Individuals then start to identify their interests with the success of their group.

This obviously induces motivations to advocate for party, like a press

Knowledge of selective scrutiny shouldn't nullify its results, so long as they weren't totally predictable.

<sup>2</sup> And it had better not, the explanation isn't news and you should've priced it in already

Total probability:

$$P(D) = P(E) \cdot P(D|E) + P(\neg E) \cdot P(D|\neg E)$$

If  $P(D|E) = P(D) + 0.2$ , for this to average out to  $P(D)$ , we need  $P(D|\neg E) = P(D) - 0.2$

Motivated cognition vs. not

'Arbitrary influences'

'Selective exposure'

E.g. Cohen 2003.

E.g. Lord et al. 1979

E.g. Kahan et al. 2017

'Mice are insects' vs. 'Roses are plants'

secretary—search for flattering info, good talking-points, etc.

This shifts people's beliefs via the self-persuasion effect.

### Virtues:

- Fits with what we know about non-political cases.
- Explains directions of bias, including believing both (1) good things about party, but also (2) bad news that's rhetorically useful.
- Explains why people *are* sensitive to the facts (need to be good press secretaries), but they still end up with polarized beliefs.
- Explains why people go to obviously-biased sources: they want to arm themselves with arguments.

E.g. Schwarzmann et al. 2022.

E.g. sports fans

'We're persecuted!', or other threats

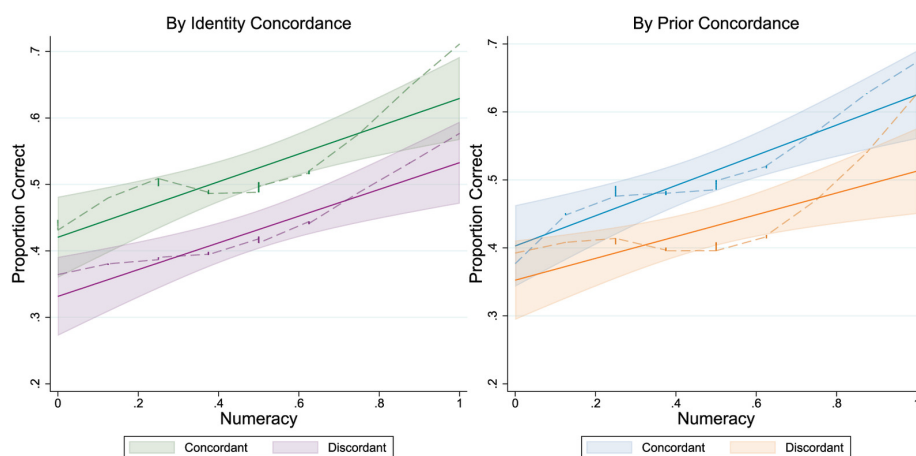
### Worries:

- How explain why people failed to converge *before* partisan sorting?
- Why does self-persuasion happen? Wouldn't you be a *better* press secretary if you didn't believe your own spin?
- This works best for highly politically-engaged partisans. What explains why the (vast majority who are) unengaged are polarized?
- This seems to predict that highly numerate people will be *more* biased, since better able to be secretaries. But initial findings to that effect (Kahan et al. 2017) have failed to replicate:

In 1950s–70s, parties were unsorted

Better grip on the facts, etc.

→ To be more convincing? Then why can't you adopt beliefs at will?



Stagnaro et al. 2023.

Incidentally, in this study, when controlling for numeracy and priors, identity-concordance has precisely zero effect on success rate.

## References

- Cohen, Geoffrey L, 2003. 'Party over policy: The dominating impact of group influence on political beliefs.' *Journal of personality and social psychology*, 85(5):808.
- Kahan, Dan M., Peters, Ellen, Dawson, Erica Cantrell, and Slovic, Paul, 2017. 'Motivated numeracy and enlightened self-government'. *Behavioural Public Policy*, 1:54–86.
- 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.
- Schwarzmann, Peter, Tripodi, Egon, and van der Weele, Joël J., 2022. 'Self-Persuasion: Evidence from Field Experiments at International Debating Competitions'. *American Economic Review*, 112(4):1118–1146.
- Stagnaro, Michael N, Tappin, Ben M, and Rand, David G, 2023. 'Unmotivated Numeracy and Self Governance: No Evidence of Motivated System Two Reasoning Across 5 Issues Using a Gold-Standard Representative Sample'.