

# Mechanistic vs. Rational Psychology

Kevin Dorst

24.805, Fall 2025

## I. The Debate

Mechanism:

- Heuristics and biases; behavioral economics.

Conjunction fallacy via representativeness

Rationalism:

- Bayesian cognitive scientists; folk psychology.

Conjunction fallacy via informativity

## II. Against Mechanism

First-Person Stability: Accepting mechanism while remaining polarized is incoherent from the first-person perspective.

Take a claim you're confident in—say,  $t = \text{Trump has violated the Constitution in his second term}$ .

*Intuitively*: Suppose you buy the mechanistic narrative (eg from Williams 2023). What should you think about the rationality of your belief in  $t$ ? It seems that *either* you must do special pleading,<sup>1</sup> *or* you will be akratic.<sup>2</sup>

<sup>1</sup> 'I didn't do motivated reasoning'

<sup>2</sup> 'If I were ideally rational, I'd be less confident than I am in  $t$ .'

More precisely:

**(P1) Anti-Akrasia**: If your estimate of the ideally-rational probability to have in  $q$  should be  $x$ , then your probability in  $q$  should be  $x$ .

If  $\mathbb{E}_{P_a}(\mathcal{P}(q)) = x$ , then  $P_a(q) = x$ .

**(P2) Political Confidence**: You're reasonable to be (say, 95%-)confident in your predictably-polarized belief.

$P_a(t) \geq 0.95$

**(P3) No Special Pleading**: If you should doubt rational psychology, then you should lend significant (say, 50%-)probability to the hypothesis that you should be a bit (say, 5%-)less confident in your predictably-polarized beliefs.

If you reasonably doubt rational psychology and  $P_a(t) = x$ , then  $P_a(\mathcal{P}(t) < x - 0.05) \geq 0.5$ .

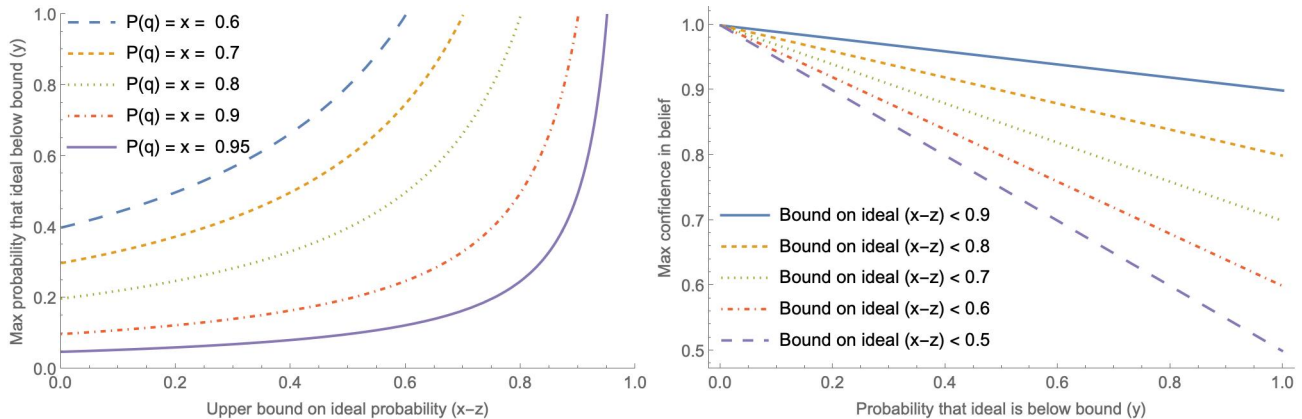
Suppose, for reductio, you should. Derive a contradiction.

(P1)–(P3) imply that you shouldn't doubt rational psychology.

Reject (P1)? Follows from Informed Reflection. If deny it, then No Foregone Questions is false, and rationality is even further separated from truth than on *my* theory.

Reject (P2)? Risk of being disingenuous.

Reject (P3)? Dynamics:



Assuming (P1) Anti-Akrasia, the options:

- 1) *Reduce confidence in predictably-polarized beliefs.*  
→ Do you think that's the right response for your beliefs about politics, religion, and who does more chores?
- 2) *Think you are an exception to mechanistic psychology's explanations.*  
→ Strong evidence that people of all political stripes, intelligence levels, etc. are influenced by these biases.
- 3) *Be believe rationalism, even for predictably-polarized beliefs.*

This, I think, is what Griffiths et al. would say. They would appeal to approximations or resource constraints to explain why we don't converge on topics like this.

I think Ambiguous Bayes gives us a better version of the argument.

- Rational psychologists are right that we approximate Bayesianism.
- For topics where we can achieve sufficient clarity<sup>3</sup> we approximate *Standard*-Bayesians and so converge. That explains our feats.
- For topics where ambiguity is endemic<sup>4</sup>, even though we approximate optimal, Bayesians, we still polarize.

<sup>3</sup> Well-constrained domains like vision and language; some parts of science

<sup>4</sup> Politics, religion, departmental dramas, (sometimes) household chores

### Better explanations

Explanatory closure—microprocessors and Marr.

Robustness.

Griffiths et al. make a version of this argument

## III. Defending Rationalism

Other salient ones, besides (1)–(5) below?

- 1) Bayesian complexity
- 2) Structural biases
- 3) (Moral) dumbfounding

Dumbfounding experiments: chicken carcass, dead dog, incest.

Objection: getting people confused about their intuitive judgment is what philosophy professors *do*. Doesn't mean their intuitive judgments were wrong.

Eg skepticism. Eg 'flarp'.

4) Bad reasoning

5) Failure to converge

A priori, it is way more likely that we would polarize if mechanism were true than if rationalism were:

$$P(\text{polarize} \mid \text{rationalism}) \ll P(\text{polarize} \mid \text{mechanism})$$

So polarization is evidence for mechanism.

But we all know that *if* we have anything like Bayesian opinions, then they're ambiguous. And we now know that *if* they're ambiguous, then we would expect polarization either way:

$$P(\text{polarize} \mid \text{ambiguity \& rationalism}) \approx P(\text{polarize} \mid \text{ambiguity \& mechanism})$$

Analogy: red-looking wall is evidence that it's red. But if we find out that it's bathed in red light, that undermines our confidence that it's red.