Chapter 11

The Science of Irrationality

Abstract

Mechanistic psychology views human cognition as a collection of flawed heuristics. Rational psychology views it as a finely-tuned system. This book has given detailed empirical arguments for rational psychology. But there is more to theory choice than that. This chapter argues that a broader view provides clear evidence for rational psychology, rebuts standard arguments against it, and shows why the results of this book help debunk the popularity of mechanistic psychology.

In the previous chapters, I've argued that we are

Reasonably Polarized: People reasonably approximate Bayesian solutions to the problems they face. Under clarity, this leads to convergence—but under ambiguity, it leads to bias and polarization.

Suppose we were to provisionally accept this hypothesis. What would follow?

In the remaining chapters, I'll trace the consequences. This chapter will argue that it should shift both our understanding of ourselves and our approach to psychology. Chapter 12 will argue that it should shift both our understanding of our political opponents and our approach to politics. Chapter 13 will argue that it promises to make philosophy more relevant, the behavioral sciences more rigorous, and both more fruitful.

11.1 Two Faces of Psychology

Chapter 1 introduced two competing views of the human mind within cognitive science.

Mechanistic psychology says that human cognition is best understood as a collection of heuristics and mental shortcuts that worked well-enough in our evolutionary past, but that misfire systematically in our contemporary environment—leading to pervasive bias, inaccuracy, and polarization (e.g. Tversky and Kahneman 1974; Ariely 2008; Thaler 2015). It centers high-level cognition—'judgment and decision-making'. It is often supplemented with the hypothesis that there is a deliberative 'System 2' to think carefully when heuristics give out—but that this system is slow, effortful, and itself quite limited in both what it can accomplish and how diligently it can be deployed (Kahneman 2011). The emphasis is on the *mis*firings of the system.

Mechanistic psychologists view human cognition the way we view a clumsy person. Why does he fall over so often? Well, walking is hard and the terrain is difficult. But still, when he slips that is a *slip*—a failure of the system to work properly; something to be explained, perhaps excused, but not justified. It would be a mistake to posit that slipping *makes sense* from his perspective, or to in other ways rationalize his clumsiness. He doesn't want to slip, nor does he want something that justifies slipping. Slips—whether of the feet, or of the mind—are not in themselves sensible steps.

In contrast, **rational psychology** says that human cognition is best understood as a finely-tuned system that flexibly deploys optimal strategies for solving intractable problems. It centers low-level cognition—the everyday feats of seeing, walking, and talking—but with increasing forays into higher levels (e.g. Gershman 2021; Page 2022; Griffiths et al. 2024). The standard way it explains biases is by appealing to resource limitations (e.g. Lieder and Griffiths 2019; Icard 2023), arguing that once we realize *how intractable* many problems are, even optimal strategies will often result in biases, inaccuracies, and polarization. This book has used a different but complementary strategy: even computationally *un*limited agents would fall into bias, inaccuracy, and polarization so long as they have ineliminable higher-order uncertainty about their own opinions. Regardless, the emphasis is on the sensible (or even optimal) nature of the system.

Rational psychologists view human cognition the way we view a skilled athlete. Why does an elite marathoner look like she's 'shuffling'—with short strides and heel strikes—toward the end of a race? Because it lets her apply force gently for longer, smoothing out inefficient peak-outputs from tired muscles, and uses the spring forces of tendons to do more of the work. It would be a mistake to posit that shuffling is a *misfire*, excusable only because marathons are hard and she's tired. Shuffling is a sensible—perhaps even bio-mechanically optimal—strategy.

Mechanistic and rational psychologists disagree over how human cognition is 'best understood'. This phrase is intentionally ambiguous between (1) the adequacy of a set of explanations, and (2) the promise of a research program. I'll take it to mean both, since they tend to go hand in hand.

In popular discourse, mechanistic psychology has won the day. There are literally hundreds of popular books dedicated to describing the foibles of human cognition, but a mere handful expounding its feats. A more rigorous way to assess their comparative impact on public discourse is to use data from Google Ngram to plot that change in relative frequency of their key terms after their inception. The locus classicus for mechanistic psychology is Tversky and Kahneman's 1974, introducing the phrase 'heuristics and biases' for their research program, later taken on in economics under the label 'behavioral economics'. Meanwhile, the locus classicus for rational psychology is Anderson's 1990, *The Adaptive Character of Thought*, introducing the term 'rational analysis' for his program. The left side of Figure 11.1 shows that the phrase 'heuristics and biases' is currently over 150 times more common than it was in 1974, 'behavioral economics' is about 70 times more common than it was in 1974, and 'rational analysis' is about 1.2 times more common than it was in 1990.[1]

[1] This ratio scale is an imperfect measure. 'Rational analysis' is a much more common phrase in standard English than 'heuristics and biases', so its instances are less diagnostic of Anderson's research program. Still, in absolute terms the changes are stark. In 1974, 'rational analysis' was 152 times more common than 'heuristics and biases'; by 2020, it was only 1.8 times more common. Meanwhile, in 1974 'rational analysis' was 27 times more common than

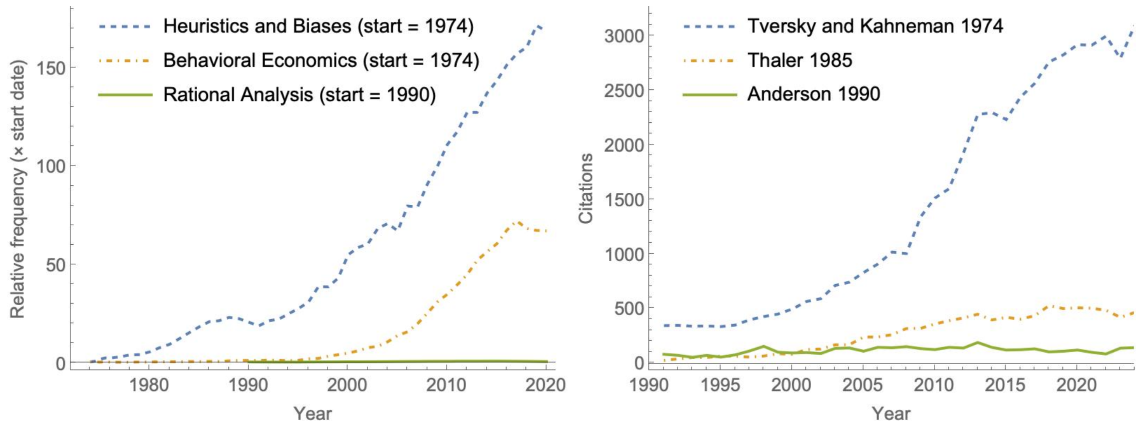

Figure 11.1: Popularity of mechanistic vs. rational psychology. ***Left:*** Data from Google Ngram (5-year rolling averages) tracking change in relative frequency for the key terms of mechanistic (dashed and dot-dashed lines) vs. rational psychology (solid line) in English-language books. ***Right:*** Citations per year of representative seminal works in mechanistic (dashed and dot-dashed lines) vs. rational psychology (solid line).

Another way to compare their influence is to measure their impact on academia. The right side of Figure 11.1 plots the citations per year (according to Google Scholar) from 1991 to 2024 of three representative works: Tversky and Kahneman 1974 (for heuristics and biases), Thaler 1985 (for behavioral economics), and Anderson 1990 (for rational analysis). Tversky and Kahneman's paper has received 51,312 citations; Thaler's paper has received 9,792; Anderson's book has received 4,171. Despite the successes of rational psychology in recent years, it is losing. Badly.

This book has been an extended defense of rational psychology. I have argued that once we take proper account of the *ambiguity* in many of our judgments, rational models offer a better explanation of not only the feats of human cognition, but also the biases that send us off the rails.

I have done so using Bayesian models. But the distinction between mechanistic and rational psychology doesn't align neatly with the distinction between non-Bayesian and Bayesian models. Although Bayesian inference and decision-making is in a sense optimal *given* the starting beliefs and values, nothing guarantees that they themselves are sensible. There are Bayesian agents who, upon seeing 100 green marbles drawn from a bag in sequence, will—because of a silly prior—bet their lives that the next draw will be blue (cf. Goodman 1955).

In practice, many Bayesian models of biases work by positing an unreasonable and unexplained mistake, and then see what's optimal given that mistake. Call such models **error-optimal models**. For example, Rabin and Schrag (1999) give a model of confirmation bias on which Bayesian agents have a positive chance of mistaking a disconfirming signal for a confirming one, zero chance of mistaking a confirming signal for a disconfirming one, and assign probability 0 to such asymmetries. Unsurprisingly, such agents easily go off the rails. Although in a narrow sense 'optimal' given their limitations (asymmetric misclassification rates) and beliefs (probability 0 that they have such asymmetric classification rates), if a real person were found to exhibit such behavior, no one would conclude that they are being reasonable. They have a dogmatic, unexplained, and incorrect

'behavioral economics'; by 2020, 'behavioral economics' was 2.5 times more common than 'rational analysis'.

belief that they couldn't be making such asymmetric mistakes; hence they ignore the accumulating evidence that they have been doing so; and hence, after drawing (with replacement) 501 green marbles and 499 blue marbles from an urn, they'll bet their lives that the majority is blue.

In contrast, **bounded-optimal models** hard-code in limitations that are sensible, justified, or excusable, and then derive what inferences and decisions are optimal given those limitations. When done well, such models allow the agent to learn about their possible limitations—but show that despite such self-awareness, biases emerge. For example Wilson 2014 shows that under a natural model of limited memory, a tendency to have asymmetries in updating (having a positive chance of ignoring disconfirming signals, but zero chance of ignoring confirming signals) is optimal in the long run because you're likely to have forgotten confirming signals along the way.

Error-optimal models can be unifying and theoretically useful, but they are not *vindicating* models: like mechanistic models, they explain behavior without justifying it. And—as I'll argue below—like other mechanistic models, they are reflectively unstable: no person can sensibly and stably think, of themselves, that they exhibit a given bias in the way an error-optimal model does—for awareness of that error should provide normative pressure for correction.

This is why it was crucial to use models that were not only Bayesian, but that started with sensible beliefs, had sensible limitations, and did sensible updates. Provably, the agents modeled in this book are aware of the causes of their biases, respond optimally to that awareness, and yet are still biased. This is why it was crucial to derive Martingale failures (Ch. 8)—so that even agents who are aware of their tendency to polarize will nonetheless do so.

Thus far, the arguments have been localized, focusing on particular findings and sticking close to the methods of the behavioral sciences. Much of the focus has been on empirical coverage: I've argued that when we look at the empirical trends surrounding probability-weighting, hindsight bias, confirmation bias, polarization, and overconfidence, we find that rational-psychological explanations rival or exceed mechanistic accounts in explanatory power.

But philosophy aims to explain *all* of the phenomena (Sellars 1963). And the special sciences would often benefit from broadening their view of what calls out for explanation. In this chapter, I'll leverage this fact to make a direct, high-level arguments in favor of rational psychology. I'll first argue against mechanistic psychology (§11.2), then rebut arguments in favor of it (§11.3), and conclude with a debunking story for its popularity (§11.4).

11.2 Against Mechanistic Psychology

I'll offer three arguments favoring rational over mechanistic psychology. Rational psychology (1) can better reconcile the two faces of human cognition and (2) gives more-complete explanations, while mechanistic-psychology (3) is unstable from the first-person perspective.

11.2.1 Reconciling the Two Faces

A standard argument for rational psychology is the successes of Bayesian models of cognition. Even mechanistic psychologists admit that there are few if any rivals to Bayesianism when it comes to re-engineering the mind (e.g. Mandelbaum 2019). We saw a version of this argument in Chapter

1: there is no theory of vision, intuitive physics, sensorimotor control, theory of mind, or language that does not require us to be adeptly deploying vast amounts of information every waking minute (see Griffiths et al. 2024).

A standard rebuttal is that there is no plausible way that the same sensible (perhaps Bayesian) processes could be governing high-level cognition—for if they were, people wouldn't persistently exhibit the polarizing biases that predictably send them off the rails (e.g. Tetlock 2005).

The standard counter-rebuttal is that resource-limitations and the open-ended nature of high-level cognition is what explains our biases and limits our performance—but we are still approximating (Standard) Bayesian processes (e.g. Lieder and Griffiths 2019). Mechanists find this unconvincing, pointing out that it's a stretch of the term 'approximation' to say that we are approximating Bayesian responses—for those responses would converge to the truth, rather than predictably polarizing (e.g. Mandelbaum 2019).

The result is a stalemate, with rationalists and mechanists each emphasizing their preferred face of this duality. But any theory of the mind needs to explain *both* faces of human cognition—both our astounding feats and our embarrassing foibles. I promised you a theory that could square this circle. And I claim to have delivered.

My theory says that the rational psychologists are right: in most of our cognitive processes, we are approximating sensible Bayesian inferences given our beliefs and evidence. When our prior beliefs about the relevant question and evidence are sufficiently *clear*, we perform our feats and converge to the truth; but when they are irreducibly *ambiguous*, we exhibit our foibles and polarize away from it (Ch. 9).

So the mechanists were right to balk at the claim that we are approximating Standard-Bayesian updating—we can't be, since we are violating Reasonable Convergence. But the problem is not the claim that we are approximating Bayesian updating, but the assumption that the updates are *Standard*-Bayesian, and hence clear. They are not; hence Reasonable Convergence is false—the (ambiguous) Bayesian updates we are approximating should *not* be expected to converge to the truth in the contexts where we polarize.

In short: once we recognize the significance of ambiguity and the falsity of Reasonable Convergence, we can give a full-throated version of the standard argument. Not only can Bayesian models of cognition explain our feats—but they also can explain and predict our foibles.

11.2.2 Better Explanations

A second standard argument for rational psychology is that it gives *better* explanations of the phenomena (e.g. Chater et al. 2024a). Suppose we have competing mechanistic and rational explanations of a given empirical finding—say, hindsight bias—each with an equal ability to explain the empirical data. Which explanation should we prefer? Rationalists will point out that there are at least two reasons to prefer their explanation.

First, rational explanations are more explanatorily closed. As Marr (1982) pointed out, we cannot understand a complex system unless we know what problem it is solving and why; and as Jonas and Kording (2017) show, even a full functional map of a microprocessor leaves us in the dark about how it works until we know what it's doing—namely, playing Donkey Kong. Likewise: having a mechanistic explanation of hindsight bias leaves open the question of *why* an otherwise well-designed cognitive system would do so, and why selective pressures (from evolution or markets) haven't selected it out. A (good) rational explanation automatically answers these questions, for it says both what the system is doing and why it makes sense.

Second, rational explanations are often more robust. A familiar point from philosophy of science is that overly-specific explanations are bad because they don't capture robust patterns in what would've happened otherwise (e.g. Woodward 1997; Batterman 2001, 2002). Suppose I drop a marble and it ends up at the bottom of a bowl. Contrast two explanations:

a) 'The marble was at height h and so had momentum m by the time it hit point p of the bowl, which was at angle θ and therefore the force vector coming off the bounce was $\vec{v}$, sending it...'
b) 'The marble was above the bowl, which is curved so that the bottom is the only stable point.'

Explanation (a) makes much more specific predictions, but it is *over*-specific: it doesn't say what would've happened if I'd dropped the marble a centimeter to the left. Explanation (b) is robust: it immediately implies that I could move my hand anywhere over the bowl and get the same result, while anywhere not over the bowl would lead to a very different result.

Rational explanations are generally more robust than mechanistic ones. They specify the global conditions (including the state of the agent and the environment) from which it makes sense that they would exhibit the behavior they do—in other words, they reveal points of stability or attractors in the space of possible behaviors. As a result, they immediately explain why similar systems in similar circumstances would (be under pressure to) behave in similar ways. The mechanistic theory of hindsight bias makes it a fluke of how we're built—for example, due to how our memory works or our motivations to look smart. The rational theory, in contrast, explains why exhibiting hindsight bias is the result sensibly updating uncertain beliefs about yourself, and therefore to be expected from similar cognitive systems.

Of course, robustness is not *always* a good thing. If in fact it is not a bowl beneath the marble but is a wire rack, and the marble just *happened* to hit the wires rather than falling through them, then explanation (a) is better. But in the absence of further information, we should expect that what we're trying to explain is not a fluke. (Imagine we are only told where the marble started and ended up, and need to posit what led it there.)

In short, when the empirical data is equivocal we should prefer explanations that are more explanatorily-closed and robust. Rational psychology delivers those virtues better than mechanistic psychology.

11.2.3 First-Person Stability

In Chapter 1, I emphasized that 'they' are us: any plausible theory of how the mind works needs to be one that we can reasonably believe about how *our own* mind works. It needs to be stable from the first-person perspective, in the sense that we can believe it applies to us as we go about our lives, forming our own beliefs and making our own choices.

Social scientists may fear that such insistence requires adopting a woolly, philosophical ('non-scientific') perspective. But it requires no such thing. Rather, it requires paying close attention to *all* our evidence about the topic, not merely a narrowly-circumscribed subset of it that is easy to run randomized experiments on. Of course, as we've seen (Ch. 9), when a type of evidence is sufficiently ambiguous that we don't know what to make of it, it can make sense to ignore it and focus on our clear evidence—provided that the latter is sufficiently abundant. But this book has argued that the experimental evidence we have bears far less clearly on the rational-vs.-mechanistic debate than has been assumed. And I intend to show that first-person perspective can actually provide quite strong and clear evidence on this front.

Here's the four-sentence version. If you doubt rational psychology, then you should doubt the rationality of *your own* predictably-polarized beliefs—after all, the mechanist's theories apply just as much to you as everyone else. But then, on pain of incoherence, you should lose confidence in those beliefs. But (I hope you agree) you *shouldn't* lose confidence in those beliefs. So you shouldn't doubt rational psychology.

Now a slower, but still informal version. None of us can rationally believe what we think an ideally-rational version of ourselves (with all our evidence) *wouldn't* believe. We *defer* to ideally-rational versions of ourselves, in the sense that we try to approximate them. Thus it can't be rational to confidently believe something, while also suspecting that your confidence is driven (or exacerbated) by irrational factors. If you rationally think, 'If I were rational, I wouldn't be too confident of q', then you can't rationally be extremely confident of q. This is sometimes called an 'anti-akrasia' constraint on beliefs: don't believe against your better judgment.

We each have our own predictably-polarized beliefs: beliefs we are very confident it, but which also were plausibly influenced by the biases that lead to polarization—hindsight bias, confirmation bias, motivated reasoning, etc. For example: in his second term, do you think Donald Trump has violated the Constitution? I'm guessing that most readers will either confidently believe *yes* or confidently believe *no*. Suppose that you believe yes: you are extremely confident that *Trump violated the Constitution*.

Suppose, finally, that you buy into mechanistic psychology and doubt rational psychology. Now you're in a pickle. For if you doubt rational psychology, then you should suspect that your confidence that *Trump violated the Constitution* is driven in part by irrational factors. After all, this is *exactly* the sort of belief that the mechanistic psychologists find—across political, educational, and socioeconomic spectrums—is influenced by polarizing biases like confirmation bias and motivated reasoning. So since you doubt rational psychology, you should suspect that your confidence is exacerbated by such irrational factors. You'll find yourself thinking, 'Trump violated the Constitution; but if I were rational, I'd probably be significantly less confident of that.' But you can't rationally think that, for it violates the anti-akrasia constraint!

In short: if you should doubt rational psychology, then you're under normative pressure to lessen your confidence in your polarized beliefs. So if you are reasonable to maintain your confidence in your predictably-polarized beliefs, you should *not* doubt rational psychology.

There are various ways to press on this informal argument. What exactly do we mean by 'ideally rational', and is anti-akrasia a genuine norm? *How much* lower does your confidence need to be, if you suspect (to what degree?) that the ideally-rational confidence would be somewhat (how much?) lower. There are too many imprecise terms in the argument. So let's sharpen it up.

The first premise is our anti-akrasia constraint. Let $\mathcal{P}$ be the ideally-rational probability function to have given your evidence. We can be neutral about it; all I'll assume about $\mathcal{P}$ is that your *informed* probabilities—$\widehat{P}$, the probabilities you would have if your ambiguity were removed—defer to $\mathcal{P}$, in the sense that $\widehat{P}$ matches its probabilities to its expectation of $\mathcal{P}$'s probabilities.[2] Then:

(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 $\mathbb{E}_{P_\mathfrak{a}}(\mathcal{P}(q)) = x$, then $P_\mathfrak{a}(q) = x$.

Anti-Akrasia follows from Informed Reflection, i.e. factorability—the minimal constraint we've imposed on all Ambiguous-Bayesian models. It is valid on every model written down in this book, including those that have undergone predictable radicalization and polarization.[3]

Recall that your 'predictably-polarized' beliefs are those that have been influenced by the biases ('irrelevant influences') made famous by mechanistic psychology—confirmation bias and motivated reasoning, but also conformism, blind trust, and indoctrination, insofar as there's evidence of those.[4] The standard list includes political, religious, and ethical beliefs—but we can now see that it should include many other beliefs that have been influenced by biases, such as the belief that you do most of the chores.

Let t be the claim that *Trump has violated the Constitution*, and $\neg t$ the claim that *Trump hasn't violated the Constitution*. Whichever your belief, I'll assume that you're reasonable to be quite confident in it. Let's say you believe t; so my second premise is:

(P2) Political Confidence: You're reasonable to be (say, 95%-)confident in your predictably-polarized belief.
$P_\mathfrak{a}(t) \geq 0.95$

So if given a choice between betting on t and the claim that a fair coin flipped 10 times will land heads at most 7 times (probability ≈ 0.945), you'd prefer to bet on t.

Final premise: no special pleading. If you think that most people's confidence in their political beliefs is driven substantially by irrational factors, then you should suspect that *yours* has been as well. Thus you should suspect that your beliefs have been directionally-influenced in the ways that the mechanistic psychologists would predict. Since you're confident of t, that's probably in part due to confirmation bias, motivated reasoning, conformity to your peer group, (overly) selective exposure to congenial sources, and so on, all favoring t. If you were *ideally* rational, you would have avoided or corrected for these biases—lowering your probability for t. Stated precisely:

(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

[2] Formally, for all worlds w and propositions q, $\mathbb{E}_{\widehat{P}_w}(\mathcal{P}(q)) = \widehat{P}_w(q)$. As we've learned, this presupposes that $\mathcal{P}$ is clear. But—despite some arguments to the contrary (e.g. Christensen 2007a, 2010a, 2016; Williamson 2008) that's no matter. When we say '*ideally* rational' we are giving ourselves free reign to idealize $\mathcal{P}$; and as we know, for any higher-order uncertain probability function R, we can make it more ideal by *informing* it to get $\widehat{R}$; then we can let $\mathcal{P} = \widehat{R}$. Indeed, the argument would go through if we let $\mathcal{P}$ simply be your informed probabilities $\widehat{P}$.

[3] To see this, treat your probabilities P, your informed probabilities $\widehat{P}$, and the ideal probabilities $\mathcal{P}$ as stochastic matrices. For $\widehat{P}$ to match its expectation of $\mathcal{P}$ means that $\widehat{P}\mathcal{P} = \widehat{P}$. For your probabilities P to match their expectation of your informed probabilities (assuming Informed Reflection) means that $P\widehat{P} = P$. For Anti-Akrasia to hold, we need $P\mathcal{P} = P$. Substituting our above identities, we know that $P\mathcal{P} = (P\widehat{P})\mathcal{P} = P(\widehat{P}\mathcal{P}) = P(\widehat{P}) = P$.

[4] For philosophical discussions of irrelevant influences, see e.g. Cook 1987; Cohen 2000; White 2010; Srinivasan 2015b; Schoenfield 2017c; Vavova 2018

in your predictably-polarized beliefs.

If you reasonably doubt rational psychology and $P_{\mathfrak{a}}(t) = x$, then $P_{\mathfrak{a}}(\mathcal{P}(t) < x - 0.05) \geq 0.5$.[5]

Notice that '*No* Special Pleading' might be a misnomer. You are allowed some special pleading, just not a *ton*. All I'm assuming is that if you should doubt rational psychology, then you should think it's at least as likely as a coin flip (50%) that, if you were ideally rational you'd be a *bit* (5%) less confident than you are. For all No Special Pleading says, you might reasonably think you are much *less* influenced by irrational factors than all the other dunces out there.

That's all we need. Suppose—for reductio—that you should doubt rational psychology. That means the antecedent of No Special Pleading is satisfied, so it follows that if you should assign x probability to Trump violating the Constitution, you should be at least 50%-confident that if you were ideal, your probability that *Trump violated the Constitution* would be 5% less: $P_{\mathfrak{a}}(\mathcal{P}(t) < x - 0.05) \geq 0.5$. By Political Confidence, you assign at least 0.95 probability to t. Suppose you assign *exactly* 0.95 to t. (This is the best case for your belief, as we'll see.)

Since $P_{\mathfrak{a}}$ is probabilistically coherent, it thereby follows that your expectation for the ideally-rational probability in t is *below* 0.95, since. Why? Well, the maximum possible value is 1, and your expectation of $\mathcal{P}(t)$ is a weighted average of its various possible values. You have (at least) a 50%-probability that it's *less* than $0.95 - 0.05 = 0.9$, so you have (at most) a 50%-probability that it's as high as 1. And the 50-50 average of 0.9 and 1 is 0.95. So your expectation for $\mathcal{P}(q)$ should be less than 0.95: $\mathbb{E}_{P_{\mathfrak{a}}}(\mathcal{P}(t)) < 0.95$.[6] By Anti-Akrasia, it follows that your probability for t should be less than 0.95: $P_{\mathfrak{a}}(t) < 0.95$. But that contradicts Political Confidence. So we reject our supposition that you should doubt rational psychology.

(If you had more than 0.95 probability for t, the argument is the same. If $P_{\mathfrak{a}}(t) = 0.97$, then $0.5(0.97 - 0.05) + 0.5(1) = 0.96$, so again your doubts about rationality force your confidence lower.)

Notice that if we drop Political Confidence as a premise, we can still infer the disjunctive conclusion: assuming only Anti-Akrasia and No Special Pleading, it follows that *either* you should lose confidence in your political beliefs, *or* you should accept rational psychology. If you coherently follow through the implications of your beliefs, believing in mechanistic psychology forces you to a form of political agnosticism.

That's the argument. It is valid, and I think sound. But let's kick the tires a bit. Notice that we can toggle No Special Pleading by varying both (1) the upper bound on the ideally rational probability and (2) your level of confidence in that bound to get different constraints on how high your level of political confidence in t can be. Under various other choices, Anti-Akrasia would still force your probability for t below the threshold set by Political Confidence.

Generically, No Special Pleading says that you're at least y-confident that the ideally-rational credence at least z lower than you are: $P_{\mathfrak{a}}(\mathcal{P}(t) < P_{\mathfrak{a}}(t) - z) \geq y$. Holding fixed your probability for t at $P_{\mathfrak{a}}(t) = 0.95$, we can get a contradiction by either setting z tiny but y pretty high: if you're $\frac{2}{3}$-confident that your probability should be at least 2.5% lower, then your expectation for the ideal

[5] For simplicity, I've treated you as certain that you're at least 0.95-confident of t, so No Special Pleading is equivalent to assigning 50%-probability the ideal probability being less than 0.9. If we wanted to be extra careful we could write $P_{\mathfrak{a}}\big(\mathcal{P}(t) < P(t) - 0.05\big) \geq 0.5$, and then add an assumption about what you believe about $P(t)$. The argument would be similar.

[6] By the law of total expectation, $\mathbb{E}_{P_{\mathfrak{a}}}(\mathcal{P}(t)) < P_{\mathfrak{a}}\big(\mathcal{P}(t) < 0.9\big)(0.9) + P_{\mathfrak{a}}\big(\mathcal{P}(t) \geq 0.9\big)(1) \leq 0.5(0.9) + 0.5(1) = 0.95$.

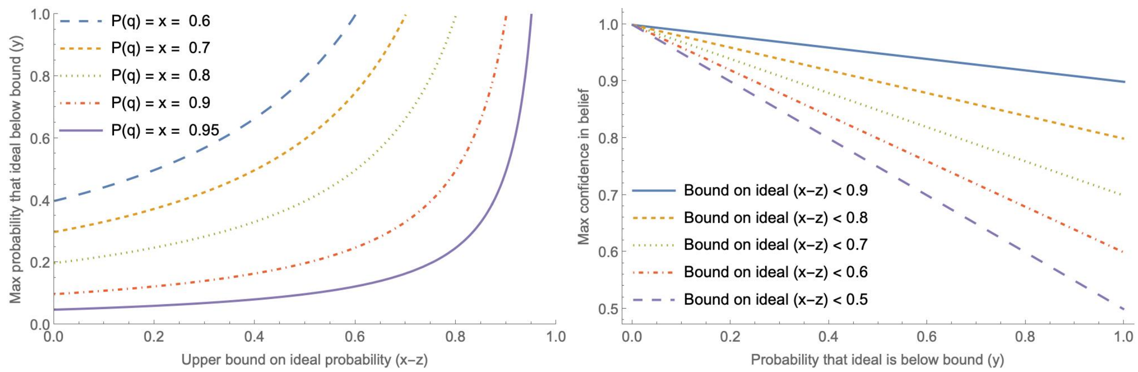

Figure 11.2: Doubts about ideal rationality constrain your confidence. ***Left:*** Given a level of confidence in your political beliefs (colored lines), we can plot the maximum probability y you can assign to the ideal probability being below a given bound. If you're quite confident, you cannot have much higher-order doubt. ***Right:*** Given a proposed bound on how confident the ideal probability $\mathcal{P}(t)$ would be, we can plot your maximum confidence in your belief, $P_{\mathfrak{a}}(t)$, as a function of how strongly you suspect that $\mathcal{P}(t)$ is below that bound. If you have much doubt about ideal rationality, you can't be very confident.

probability should be lower than $\frac{2}{3}(0.925) + \frac{1}{3} = 0.95$. Alternatively, we can set z large but y tiny: if you're merely 25%-confident that you should be 15%-less confident, then your expectation for the ideal probability is lower than $0.25(0.8) + 0.75(1) = 0.95$.

More generally, fixing a level of confidence x in your predictably-polarized belief, we can plot the maximum probability you can coherently assign to the ideal probability being below $x - z$, as shown on the left of Figure 11.2. For example, if you assign $x = 0.9$-confidence to t, then you can be at most 50%-confident that the ideal probability is below 0.8, and at most 25%-confident that it's below 0.6 (dot-dashed red line). Meanwhile, if you assign $x = 0.7$ probability to t, you can be at most 75%-confident that the ideal probability is below 0.6 (dashed yellow line).

Going the other direction, we can choose an upper bound on the ideal probability $(x - z)$ and an amount of confidence that you have that the ideal credence is below that bound (y), and infer the maximum confidence you can have in your political beliefs $(P_{\mathfrak{a}}(t))$, given those doubts about ideal rationality. For example, if you're 50%-confident that $\mathcal{P}(t)$ is below 0.9, you must be less than 0.95 in t (middle of blue/solid line); if you're 50%-confident that $\mathcal{P}(t)$ is below 0.8, you must be less than 0.9-confident in t (middle of dashed yellow line); if you're 50%-confident that $\mathcal{P}(t)$ is below 0.7, you must be less than 0.85-confident in t (middle of dotted tree line); and so on.

In short: your doubts about whether an ideal version of you would be as confident as you are puts strong upper bounds on how confident you can be. Since doubting rational psychology should increase your probability that an ideal version of you would be less confident, it puts strong upper bounds on how confident you can be in your predictably-polarized beliefs.

Since the argument is valid, the only response is to reject a premise. How might that go?

Some externalist epistemologists might want to reject (P1) Anti-Akrasia, the assumption that you should match your probability in q to your expectation of the ideal probability in q.[7] Note that the main arguments for the rationality of epistemic akrasia are not arguments against Anti-Akrasia

[7] E.g. Williamson 2014, 2019; Lasonen-Aarnio 2015; Christensen 2016.

as formulated here. They attempt to show that you can rationally fail to match your probability to your expectation of the *uninformed* (rational) probability—they are violations of what I've called 'Self-Martingale' (Ch. 5; see Salow 2019; Dorst 2019). And, more generally, they are built on cases where you are rationally but radically misled about what the (uninformed) rational credence is. This is perfectly consistent with my formulation of Anti-Akrasia, since it follows from Informed Reflection (aka 'New Reflection'; Elga 2013), which is consistent with such radically-misleading higher-order evidence. Indeed, Informed Reflection is valid on the standard models that are used to argue for the possibility of rational epistemic akrasia, such as Williamson's unmarked clock (Williamson 2014) and Horowitz's dartboard (Horowitz 2014). Since it also follows from minimal rationality conditions (like No Foregone Questions—see Ch. 5), the case for Anti-Akrasia as I've formulated it remains strong.

Of course, there are philosophers who argue against Informed Reflection (e.g. Lasonen-Aarnio 2015; Christensen 2016). I don't find their cases compelling; but if you do, what should you think? First, to run the argument we don't need the full strength of Anti-Akrasia; all we need is a thresholded, 'Trust-like' variant: conditional on the ideal probability for t being below x, your credence in t drops below x, $P_{\mathfrak{a}}(t|\mathcal{P}(t) < x) < x$. Objectors need to deny that there is *any* notion of 'ideal probability, given your evidence' that satisfies this constraint. And second, we don't need to assume that the constraint *always* holds. Suppose there are recherche counterexamples involving credible testimony that you're hypoxic or have taken a reason-distorting drug. No matter. All we need to assume is that in *normal* cases of predictably-polarized beliefs (say, about politics), when you're otherwise thinking clearly and have no paradoxical evidence, the constraint holds. That's enough to run the argument.

Can we reject (P3) No Special Pleading? This says that if you reasonably doubt rational psychology, then you must lend significant credence to the possibility that the ideal probability for you to have in your predictably-polarized beliefs is should be lower. Of course, you might try to do some special pleading and find differences between yourself and the people in psychology studies who exhibit polarizing biases. This will be hard lift. Although there are always moderators, there is strong evidence that people of all demographics, partisanships, education-level, and IQ are susceptible to substantial degrees of hindsight bias (e.g. Tetlock 2005), confirmation bias (e.g. Ditto et al. 2019; Stanovich 2020), and overconfidence (e.g. Moore et al. 2015). These biases are substantial enough that, on questions we can measure the truth of, people of all political partisanships regularly fail to converge to the truth—often with those on the political extremes mis-estimating the truth in their preferred direction, and doing so *more* than political moderates (e.g. Yudkin et al. 2019; Zmigrod et al. 2019; Zmigrod and Rentfrow 2020; Rekker 2024). Indeed, there is some (contested) evidence that those who are more educated or numerate are *more* susceptible to such biases (Kahan 2013, 2018).

Finally, and most importantly: it's one thing to find some differences between yourself and others; it's another to justify starkly different estimates of your own susceptibility. No Special Pleading doesn't say that you should expect that you're as irrational as the average psychology-experiment subject. Rather, it says you need to lend *enough* credence to being subject to the polarizing biases that you think it's as likely as a coin-flip (50%) that, if you were ideally rational, you'd be a *bit* (5–10%) less confident in your political beliefs. It's hard to see how you can seriously doubt rational psychology while denying that.

Can we reject (P2) Political Confidence? I suspect many people will claim to go this way—'I really am not that confident in my views!'. But I also suspect that rarely is this position coherently carried through to your other actions. Beliefs are functional states, not what we assert in the context of an academic seminar. If you *truly* were not confident that Trump has violated the Constitution—or of your other political beliefs—then there's all sorts of behaviors you'd likely take: (1) listen to conservative media, to find out what their view is of what Trump has done; (2) be curious about what's motivating your friends or family members who support his administration; (3) avoid talking or feeling or posting as if we're living through the collapse of American democracy. If you find yourself not doing any of these—let alone if you find yourself upset at those who *do* do these things—that strongly suggests that you remain extremely confident.

Moreover, I think, you are *rational* to do so. Should you really give up your confidence in your deeply-held political (and religious, and historical, and scientific...) beliefs on the basis of what you've learned about psychology? That sounds dangerously close to what undergraduates do when they (claim to) become philosophical skeptics after the first day of Introduction to Epistemology—over-indexing on persuasive intellectual arguments, while failing to credit the vast, hard-to-articulate sub-structure of their life experience that tells them otherwise.

Likewise for politics. If someone tried to make an anti-akrasia-plus-psychology argument against your political beliefs, and in favor of political agnosticism, the correct response is to reject one of their premises. You should be confident that something has gone wrong somewhere, even if you can't pinpoint where. The argument of this book is that the incorrect step is the normative interpretation of the psychological studies of biases. '*Yes* I am subject to the polarizing biases of the mechanistic psychologists. Yes, if you had asked me *before* I began to be subject to them, I would've predicted the polarized results, and would've believed that if I were ideally rational, I would not be so polarized. But I was wrong. I see now that, if anything, my ideally-rational self might well shift in an even *more* extreme direction than I have—for it would not be weighed down by higher-order doubts. My current beliefs *do* match my best estimate of what my ideally-rational self would do, despite the fact that I have been subject to such polarizing biases. For I have been *reasonable* to exhibit those biases and become polarized.'

This, I think, is the right response to the argument. We *should* (rationally) be confident of our political beliefs—even those of us who, due to ambiguity, are wrong. And as a result, we should doubt mechanistic psychology and believe that rational psychology offers the right explanation for our biases. If you don't, you are faced with a choice between (1) incoherence or (2) political (and religious, and historical, and scientific...) agnosticism. Pick your poison.

11.3 Defending Rational Psychology

These defenses of rational psychology would come to little if the arguments against it were overwhelming. Are they?

11.3.1 Bayesian complexity

There are two types of argument for the conclusion that we *couldn't* be approximating Bayesian inference, because it's just too hard.

The commonsensical argument is that numbers are just too hard to work with—give people a simple probability problem, and see how they flub it. The problem with this argument is that it obviously overgeneralizes—it would equally show that we can't see or walk or talk, since the computations involved in each of those feats are stunningly complex. Obviously we can't *consciously* crunch the numbers needed to recover a 3D shape from changes in luminance values. Equally obviously, our brain *does* subconsciously do this every waking second. Likewise, we can't consciously calculate how much more likely it is that your eyes would dart down and to the left if you were too busy to continue our meeting; but we all know instantly that that means you've glanced at your watch and so I should ask if we should wrap up.

Other times the 'too complex' argument is wrapped in computer-science jargon: people will point out that Bayesian inference is computationally intractable, or (if they're feeling sophisticated) 'NP-hard'. As the size of the problem grows (the number of possible states you're tracking), Bayesian inference quickly gets impossible for even supercomputers to perform over arbitrary problems.

The issue with this argument is that it's an extremely unreliable inference to go from a fact about computational complexity to a claim about what the human mind might be (approximately) doing. It's *also* NP-hard to solve a Sudoku puzzle, to schedule courses and meetings, to find the shortest route between cities, and to fill a backpack optimally. Moreover, similar arguments show that the same reasoning should show that vision is impossible—there are simply too many possible arrangements of objects and lightings (Brooke-Wilson 2023). But obviously real people do passably well at these problems all the time. *How* exactly they do so is a challenging question for re-engineering the mind; but no armchair argument can plausibly establish that we couldn't possibly solve such problems. Likewise for approximating Bayesian inference.

11.3.2 Structural biases

Forget the polarizing biases that this book has focused on. What about the *structural* biases that seem to show that people don't handle uncertainty in the way Bayesians would? For example, the conjunction fallacy seems to show that people will think representative and informative conjunctions ('Linda is a feminist and a bank teller') are more likely than their conjuncts ('Linda is a bank teller'), despite the fact that the probability of the latter must be higher than that of the former (Tversky and Kahneman 1983). For another, the base rate fallacy seems to show that people ignore base rates when forming estimates, rather than using priors as their starting point (Kahneman and Tversky 1973). For a third, pseudodiagnosticity seems to show that people often ask questions that are 'non-diagnostic' about the question they're interested in (no more likely if q than if $\neg q$), and yet update on the answers as if they were diagnostic (Doherty et al. 1979). And so on.

What to make of these results? They certainly put pressure on the idea that the mind is approximating Bayesian inference. But to assess how compelling they are, we need to look closely at the details of the results and the theories that can explain them.

For instance, the classic explanation of the conjunction fallacy says that people's judgments are driven by 'representativeness'. But representativeness is rarely made precise enough to make clear and general predictions. And when it is—as when, for example, it is formalized as how much the hypothesis was *confirmed* (made more likely) by evidence that was provided (e.g. Crupi et al. 2008; Tentori et al. 2013)—it makes clearly false predictions, predicting that people will think that it's more likely that (1) Linda is a feminist bank teller who had a hangnail on her left pinky finger last Thursday than that (2) Linda is a bank teller, since the further conjuncts on (1) are independent and thus (1) is also confirmed. Moreover, when you control for conversational relevance, representativeness understood as confirmation no longer drives the conjunction fallacy (Chung et al. 2025). And there are other theories, such as a tradeoff between accuracy and informativity or other goals in communication, that explain a wider range of the data than representativeness (Yaniv and Foster 1995; Dorst and Mandelkern 2022; Quillien et al. 2025)

Similar results hold for other structural biases. It turns out many findings of base-rate neglect are driven by conversational dynamics (Schwarz et al. 1991), and that people standardly *are* sensitive to base rates (Koehler 1996), often remarkably so (Tenenbaum and Griffiths 2006). Likewise, it turns out that often apparent pseudodiagnosticity is explained by using prior information to extract as much information as possible from the possible questions (Feeney et al. 2008; Crupi et al. 2009). And so on.

Suffice it to say that the debate about how to interpret these structural biases rages on, and it is far from obvious that the most plausible and general explanation of them is that people don't handle uncertainty in a Bayesian way. The jury is still out.

11.3.3 Moral dumbfounding

Infamously, if you ask people to provide reasons for their strongly-held political beliefs, they often can't supply many—and, often the ones they do supply will seem comically weak in comparison with the strength of their convictions. I once asked a well-regarded academic why he had shifted from liberal to conservative, and he said something to the effect of, 'I saw enough patterns in (functional vs. dysfunctional) communities that I concluded progressives had missed the mark'. A fine start, but further follow-up questions did not reveal much more than racially- and socioeconomically-charged anecdotes.

This dynamic often comes up in discussions of political rationality. If people really had good reasons for their beliefs, surely they'd be able to trot them out when pressed and debate the issues in an informed are nuanced way. They can't—indeed, when you press them on their beliefs, it quickly results in 'moral dumbfounding', where they simply insist on their moral (or political) judgments, without being able to articulate the reasons behind them (Haidt 2001). For example, Haidt et al. 1993 found that people judged it to be categorically wrong for people to eat their dog after it had died of natural causes, or for a person to use a chicken carcass to masturbate, even when the vignettes are designed so that no harm ensues. When pressed on why they make such judgments, and reminded that no harm had come, people often insisted on their judgments without being able to provide reasons for them. A common hypothesis to explain such patterns is that people's moral and political judgments are driven by affective reactions, and then are simply *rationalized* by conscious reasoning processes. An 'emotional dog and its rational tail', in Haidt's memorable phrase.

To see that something has gone wrong here, try it on yourself. Consider a good friend of yours, who you know fairly well. Suppose you discovered that he had once eaten his dog after it died. Suppose you were assured that no harm came of it, and it was a one-off thing. Would you shrug it off, carrying on as if you had found out some other mildly-surprising fact about him—like that he had once vacationed in South Dakota? *Should* you?

Obviously not. You would and should be deeply disturbed, re-thinking what you thought you knew about your friend and what that means for how you should interact with him going forward. Of course, if you were pressed to elaborate using purely-utilitarian considerations why you would be so disturbed, you would quickly run out of reasons. But that is like asking you to explain, on purely utilitarian reasons, why you shouldn't murder your aging grandma in order to give momentary sexual pleasure to a shed full of rabbits. (If the shed is big enough, the pleasure/pain calculus might not work out for Gran.) It is an *objection* to simple-minded utilitarianism that it cannot provide the resources to explain the richness of our moral life—including hard-to-formalize notions like respect for persons and the sanctity of life.

Of course, with enough argument you might be able to convince yourself that those richer moral notions rest on some sort of mistake. But likewise, with enough argument you might be able to convince yourself that we don't know anything, or that nothing is morally wrong, or that tables and chairs and people don't exist, or that all arguments are simply a mask for power, or that since modal realism is true, there are talking donkeys. Spend enough time in philosophy departments, and you'll cease to underestimate the madness of the conclusions that reasonable people can be convinced of.

Suppose, having convincing yourself of such conclusions, you then return regular society and ask ordinary people to explain how we know anything, or how anything could be morally wrong, or how there could be tables and chairs if really everything is just atoms and void, or whatnot. Dumbfounding will shortly ensue. As any parent of a 5-year-old or a philosophy major can attest, when someone insists on asking 'Why?' sufficiently persistently or cleverly, most people will soon run out of answers. That does not mean that they were not justified in believing the answers they started with; what it shows is that having good reasons to think things is a distinct ability from being able to remember or articulate those reasons. And that if a clever person sets their mind to it, they can make pretty much anyone look confused.

Here's a simple example:

A: 'I'm going to tell you about a new type of number. If a number is prime and greater than 3, it's a flarp.'

B: 'Okay, got it.'

A: 'Is a 5 a flarp? What about 9?'

B: 'Yes and no.'

A: 'Why no to 9?'

B: 'Because it's not prime.

A: 'I didn't say anything about when the number isn't prime. All I said was that *if* a number is prime and greater than 3, it's a flarp. I didn't say 'if and only if'.'

B: ‘Oh, well I thought you meant the stronger thing I guess.’

A: ‘Why in the world would you think that? Can’t you see how silly that is? If I say, ‘If that guy over there is drinking alcohol, he’s over 21’, do you infer that if he’s over 21, he must be drinking alcohol?

B: ‘Well, no...’

A: ‘But it’s the exact same inference. Do you not understand how conditionals work?’

B: ‘Ummm. I don’t know, maybe not...’

Should we conclude that there was no reason behind B’s answer, and that he simply misunderstands how conditionals work? Obviously not—A is the one who is confused. He has spent too much time in Logic 101 and too little time in ordinary conversation or linguistics classes. The phenomenon of saying ‘If X, then Y’ to mean ‘X if and only if Y’ is called ‘conditional perfection’, and it is a widely observed regularity especially in the contexts of definitions (Horn 2000). When a mathematician says ‘If a number is divisible by only 1 and itself, it is prime’, they mean to be giving us a definition of being prime, not merely a sufficient condition for it; and we pick up on contextual cues to realize that. Indeed, getting artificial language systems to mimic such pragmatic inferences is one of the achievements of modern AI.

The same is true for most of our beliefs and inferences. Try to explain to a skeptical audience how you came to believe that you do most of the chores or that your friend is a nice person or that climate change is human caused, and you will run out of room real fast. Such beliefs are built upon a vast substructure of evidence and reasons that we have imbibed and then largely forgotten. This information is often *encoded* in our affective system, tracking what we have reason to think is valuable or plausible (Railton 2014). Our gut reactions to things, whether moral, political, or philosophical, encode our lifetimes of experience with them. Such intuitions are far from perfectly reliable—as I have been at pains to show, ideally responding to reasons in the face of ambiguity is no guarantee of truth. But showing that they are sometimes unreliable and often inarticulable falls far short of showing that they are not responsive to or guided by reasons. Indeed, drawing this conclusion carries heavy risks of massively over-generalizing, leading you to think that people don’t have reasons to believe that the sun contains hydrogen or that Venus is a planet.

11.3.4 Bad reasoning

You might think this is all too sophisticated. Forget the general dumbfounding argument, or the high-level facts about polarization and biases. Just look at the political arguments people find convincing! People love hearing and repeating talking points on their side. But the inferential leaps from these points to their conclusions are so vast that we can’t understand them as trying to get to the truth—they must be doing something else in listening to arguments (Mercier and Sperber 2017).

Let’s be concrete. Consider two political arguments offered on August 25, 2025 on (opposing) popular podcasts. First, here’s Ben Shapiro arguing that the Democratic party is overly feminized:

> **Conservative.** Well, perhaps Democrats are learning that soy is actually not an amazing nutritional base for all things. They might be learning this from Zoran Mamdani,

> who is, of course, the likely next mayor of New York City....
>
> But if we're speaking about power, if we're speaking about socialist strongmen, Zoran Mamdani is going to be hard pressed, no pun intended, to actually be a strongman, pun intended. Zoran Mamdani, yesterday, showed up at the annual Men's Day Open Streets event in Brooklyn on Saturday, where a bunch of dudes were lifting, apparently. And the 33-year-old socialist... tried to bench...
>
> 135 pounds is not a lot of weight, like not a lot of weight at all... He gets two very assisted, very, very assisted bench presses at 135 pounds....
>
> So here's the thing. I think that for the vast majority of normal men, this is very, very embarrassing. But the left has decided that they are done with male normality, like truly done with it, that the future of the Democratic Party is apparently going to be an extraordinarily feminized party, not just because an outsize share of Democrats are now single women, but also because the kinds of men the Democratic Party wants to attract are not the kinds of men who lift heavy weights...
>
> The idea that you should stay healthy and that you should lift is Republican coded. And that goes to: why do you want a sort of weak and feminized America?
>
> Why do you? I mean, the media clearly do. There's a piece from Axios that is sort of fascinating about the supposed redefinition of identity under President Trump...
>
> American identity, you're telling me, has nothing to do with biblical values, meritocracy, traditional gender roles, and the nuclear family? That comes as a shock to literally every one of the founders...
>
> Axios is expressing a traditional left wing view of what it means to be an American... Apparently, no Judeo-Christian foundations to our civilization, no Western civilization at the foundation of America, no meritocracy, no traditional gender roles, and no nuclear family.
>
> By the way, the results of this are very clear. If you get rid of things like biblical values, traditional masculinity, the nuclear family, what do you end up with? You end up with a bunch of rootless and useless young adults. [CITE]

Second, here's Bill Kristol on The Bulwark arguing that what's happening in the U.S. isn't 'democratic backsliding', but a march toward dictatorship:

> **Liberal.** This term 'democratic backsliding' has become pretty common in the last 10 years... but it really was invented... to describe the troubles in Central and Eastern Europe, and actually Russia and the former Soviet Union itself. It's the difficulty of getting rid of the old authoritarian habits, of getting democratic institutions solidly embedded.
>
> But the impression it gives, and why I think it's misleading now to apply to the U.S., is there's just this pull away from democracy in countries that haven't had it for a long time. It's kind of 'old habits die hard'.
>
> But that's not the real story here. The story here is less backsliding and more a genuine

> purposeful march towards despotism, dictatorship... You go through a list of various things that have happened over the past couple of weeks. Military leaders, intelligence professionals purged, critics' homes raided, obviously, John Bolton, that happened on Friday. The continued Epstein cover-up. The major corporations extorted. Presidential control of law enforcement in the nation's capital intensified, with promise that it will expand to other places. The list goes on...
>
> I do think that the takeover of the DC police force, the guard, the use of troops in LA, and then the open-ended promise by Trump that he's just going to do this elsewhere in the country. That's a pretty standard marker. That's not from the backsliding. That's Trump choosing to embrace an authoritarian and pre-dictatorial method...
>
> Trump explicitly said they're planning to expand this into Chicago and New York. In New York, there's a subtle Zoran threat there, and we'll see what they decide to do. But all of that, to me, this is what a lot of us that were sounding the alarm had said was going to happen and warned about. I think that the playbook here is the most straightforward of all of the things he's doing...
>
> A lot of what's happening in DC, it strikes me, is basically a giant ICE raid with some other stuff going on on the side, because that is so central to their agenda. And kind of like a fascist play. It's like play acting with ICE raids together. [CITE]

I'm willing to bet that you think at least one of Conservative and Liberal is a *bad* argument—in the sense that *no one* should find it convincing. The jumps from the premises (Zoran Mamdami can't lift; Trump is sending the national guard to D.C.) to the conclusions (the left hates Western civilization; Trump is instituting fascism) are simply too large, the steps along the way too tenuous.

Yet these are a couple of the most popular political podcasts on the right and left—Shapiro's show currently has 7.2 million subscribers on YouTube, while the Bulwark has 1.5 million. So what are people doing, if so many of them are listening to—and being convinced by—such bad arguments?

The default hypothesis for why people listen to arguments—which we know is the dominant force in many contexts—is that people seek out information because they want to be informed, so they listen to sources they have reason to think are trustworthy, and are convinced mainly by arguments that provide them with genuine evidence. But this rationalist picture of argumentation seems to fit poorly with the observation that (many) people like listening to arguments that (many of us) think are clearly bad.

A natural conclusion is that they are not listening to arguments for the sake of figuring things out, but instead for the sake of cheerleading their own tribe—like sports fans cheering on their team.[8] So we hypothesize that they are engaged in motivated reasoning and partisan cheerleading, rather than taking reasonable steps to become informed. That would explain this observation, and also the fact that people fail to converge to the truth about politics. It's a live hypothesis.

But it is a hypothesis. Call it the *cheerleading explanation* of partisan argumentation. We should assess it the way we assess any other hypothesis—by seeing how well it explains the totality

[8] E.g. Klein 2014; Kahan et al. 2017; Kahan 2018; Mandelbaum 2019; Williams 2021, 2023

of evidence, in comparison to its rivals. So what is the relevant evidence, and what are the rival explanations?

Let's bin the evidence into three broad trends:

1) Many people listen to arguments with big gaps in them when they favor their political side;
2) Over the long run, many of them end up being convinced by such arguments; and
3) In doing so, many of them polarize away from the truth.

The rival, *rationalist explanation* is that people *are* taking reasonable steps to try to be informed on the basis of their beliefs about who is trustworthy and insightful. As I've argued at length, the best version of this hypothesis is built on the fact that political arguments are conducted under conditions of high ambiguity—people are unsure what they expect to hear if the argued-for claim is true vs. false. Attempting to avoid ambiguity and get reliable information leads them to engage in Selective Search, tending to listen to and construct arguments favoring their side. This means that people on opposite sides will come away with *different evidence* when they listen to opposing arguments, since each will scrutinize the opposing argument more than the supportive one, generating internal arguments against it (Kelly 2008). Since failing to find evidence generates more ambiguity than finding such evidence, Arguments Work: searching for evidence for q tends, on average, to increase people's confidence in q (Ch. 8). Although this process *increases* the accuracy of their beliefs about the local subject-matters they're learning about, these accuracy-increase are asymmetric in a way that leads, in the long run, to polarization away from the truth (Ch. 9).

How do the cheerleading and rationalist explanations explain our empirical trends?

The cheerleading explanation says that people ignore the gaps in their own side's arguments, since they don't (much) care whether those arguments are tracking the truth. But since their motivations are flipped when listening to opposing arguments, they find it cognitively taxing (they feel compelled to scrutinize each step) and unpleasant, but are able to spot their flaws. That explains (1). Since they have their defenses down when listening to supportive arguments, they eventually are convinced, explaining (2). And since they are not assessing arguments for their merits, this process leads many of them away from the truth, explaining (3).

The rationalist explanation points out that *all* arguments have gaps in them; what's happening when someone gives an argument is they are searching for a fact e such that successfully reporting that fact (*found* e) raises their audience's probability for a target claim q. Due to ambiguity asymmetries, such arguments work *on average*—across people and instances—to nudge people in the intended direction, but many individuals in fact 'backfire' along the way, thinking 'That's the best they can do?' Whether someone does so is correlated with whether the argument is congruent with their prior beliefs, since incongruence is both an epistemic and practical motivation to scrutinize the argument. That explains (1). Of course, no one finds these arguments 'convincing' in the sense that they go from a low probability to a high probability in a big-picture political claim q. But since Arguments Work on average, many people *do* shift substantially in the argued-for direction over long time periods, explaining (2). And since this is the evaluation of evidence under ambiguity, reasonably performing this process leads to local improvements in accuracy but global polarization away from it, explaining (3).

At a high level, as far as I can tell, it's a draw—it's far from obvious which explanation is better. To assess that, we need to dive into the details. The cheerleading explanation will point out that my theory requires speculative Bayesian machinery and involves complex models that may be *too* flexible, able to explain too much. I will point out that the cheerleading explanation is informal and so fairly un-predictive on its own, has trouble explaining why Motives Work Indirectly (Ch. 8), why political arguments involve facts rather than literal cheerleading ('Obama sucks!! Go Trump! Yeah!!'), and how to avoid special pleading in explaining why *you* listen to political arguments. That is what I have tried to do throughout this book; I think the case for the rationalist explanation is strong.

Cite Rand adversarial collab, when out

But my point here is simply that the devil is in the details. We cannot simply appeal to high-level facts about who listens to which arguments to conclude that the rationalist explanation is wrong. And what of the low-level facts involving the particular arguments that Shapiro or Kristol make? I think the rationalist explanation explains people's reactions rather well.

Shapiro's target audience walks in with a widely-shared set of background beliefs that workplaces and popular culture have become overtly feminized or otherwise challenged masculine gender roles, and that boys and men are increasingly struggling in professional and personal life. They will likely have come across facts which support this claim—for example, that across income and race, significantly more women now enroll in college at men (around 47% vs. 63% at median incomes; cite); that the gap in college degrees amongst 25–34 year olds was 20% to 12% favoring men in 1970 and yet was 46% to 36% favoring *women* in 2021 (cite); that marriage rates are falling much faster for those (more of whom are men) without a degree; that from the Silent Generation to Gen Z, the rates of people identifying as LGBT have risen an order of magnitude, from 1.7% to 19.7% (cite); that from 1990 to 2021 the proportion of men who said they had 'no close friends at all' rose from 3% to 15% (cite); and that median real wages for men were higher in 1979 than the decades up till 2017, while such wages have risen steadily for women throughout that time period (cite).

Add citations

This audience then watches as Shapiro's show starts with the funny video of Mamdami as the lede, and are wondering what Shapiro is going to do with it. They see how he connects this anecdote to a broader trend. The culture *has* gotten much more feminized, and the Democratic party *does* seem to be celebrating that fact, along with the fall of other traditional values—while dismissing the consequences for them and theirs. From those background beliefs, Shapiro's anecdotes *are* evidence that they should not support the Democratic party.

Of course, if you walk into Shapiro's show skeptical, you will probably both have different background knowledge. You'll know that 'feminization' of culture has come along with massive declines in the rates of wars and violence (Pinker 2012); that the median woman still earns only 85 cents on the dollar compared to the median man (cite); that the majority (73%) of the increase in LGBT identification has come in the 'Bisexual' category (cite), providing evidence that this is largely due to the relaxation of societal strictures; and so on. As a result of your skepticism, you'll scrutinize Shapiro's argument as he spools it out. You'll point out that many Democratic men (and women) *do* work out—and in fact that exercise is correlated with education level, which is in turn correlated with being liberal. You'll point out that often people show strength by forbearance—that our country looks *weaker*, not stronger, when it extorts our allies for tiny investments and changes in trade policies. And you'll point out that the goal of liberal cultural and economic changes is to achieve *true* meritocracy. Having walked in more skeptical, and pointed to half a dozen objections to Shapiro's reasoning, you'll conclude 'if that's the best he can do to argue against the Democratic party—while Trump is sending the military into the streets, no less!—then I really *should* support Democrats.'

The same can be said for Kristol's argument. Most of the audience will walk in worried about the 'march to dictatorship' he described, but only having heard of one or two of the developments he mentions. Then Kristol hits them with a list of half-a-dozen new steps in only the past few weeks, including military and masked ICE agents in the streets. At a high level, you *would* expect to see that if Trump really were trying to implement fascism. So liberals nudge their confidence in that hypothesis upwards.

Meanwhile, a skeptical conservative will scrutinize the argument—pointing out that Trump has followed supreme court orders, that the IRS during the Obama administration infamously went after conservative organizations, that the president uncontroversially has plenary power over Washington D.C., and that Trump is always blustering and posturing, so he's probably just doing that with Chicago. Having walked in skeptical, and pointed out half a dozen objections to Bill's reasoning, they'll conclude that if that's the best he can do, Kristol probably just has 'Trump Derangement Syndrome'.

In short, neither the high-level empirical trends nor the low-level facts about particular arguments clearly favor the cheerleading explanation over the rationalist one. The fact that your political opponents nod along to arguments that you find unconvincing is *exactly what we should expect* if people are responding reasonably under ambiguity.

In sum: the arguments against rational psychology are far from overwhelming.

11.4 Debunking Mechanistic Psychology

But there is one argument I have omitted. The vast majority of academic psychologists (though not 'cognitive scientists') have been convinced of mechanistic psychology. Why?

The answer, I think, is simple. The point of *rational* updating is to figure out the truth. So, from the armchair, people are far more likely to converge to the truth given rational psychology than given mechanistic psychology. Polarization is incompatible with convergence. Therefore, they are far less likely to predictably polarize conditional on rational psychology than on mechanistic psychology. Semi-formally, our prior probabilities should obey the following inequalities:

$$\begin{aligned} P\Big(\mathit{converge}\Big|\mathit{rationalism}\Big) &\gg P\Big(\mathit{converge}\Big|\mathit{mechanism}\Big), \text{ therefore} \\ P\Big(\mathit{polarize}\Big|\mathit{rationalism}\Big) &\ll P\Big(\mathit{polarize}\Big|\mathit{mechanism}\Big) \end{aligned}$$

But we *do* predictably polarize. That is far, far more likely given mechanistic psychology than rational psychology. Things are evidence for things that make them likely. So the high-level patterns of human belief and disagreement are near-overwhelming evidence for mechanism.

Indeed, we can strengthen this argument. For when we look at the details of what causes people to polarize, we uncover the polarizing biases of hindsight bias, confirmation bias, motivated reasoning, overconfidence, and so on. From the armchair, these are even *less* to be expected given rational psychology than given mechanistic psychology—as we've seen, Standard Bayesians will never exhibit them. Summarizing the facts about these findings as '*biases*', we get that

$$P\Big(polarize \;\&\; biases\Big|rationalism\Big) \;\lll\; P\Big(polarize \;\&\; biases\Big|mechanism\Big)$$

Rational psychology would make the fact that we predictably polarize along with the way we do so much less likely than mechanistic psychology. So realizing that we *do* polarize and exhibit such polarizing biases is overwhelming evidence for mechanistic psychology. All that remains, in the eyes of most psychologists, is to uncover what exactly the mechanisms are.

This is a strong argument. I think it explains—and to a great degree justifies—the widespread acceptance of mechanistic psychology. I think it is why when people give arguments for mechanism, and others respond with detailed empirical and normative critiques of them, the latter often feel hollow—like the rearguard action of a flailing research program. Sure, there are empirical problems with the details—replication failures, alternative explanations, and the like. Sure, Bayesian models of cognition can do some impressive things. But take off your theoretician's glasses, and just look at what people believe! Look at politics! Look at religion! Look the denial and obfuscation of science (about climate change or gender—pick your poison). If rational psychology were right—if people were approximating rational Bayesian updating—surely they wouldn't so regularly end up so far off the rails.

Suppose you're a social scientist, starting your career. You find it a priori plausible that rational methods of forming beliefs are likely to lead to the truth. You are confronted with a wide variety of theorems and simulations—the Bayesian convergence results—that seem to show in detail why this is the case. Of course, you know that these results have idealizing assumptions built in. But there's no reason to expect that those idealizations *radically* distort the results—surely properly taking account of resource constraints or inexact attitudes or ambiguous opinions will still lead to similar conclusions. So surely, you think, even once we add in all the complications, *still* rational psychology makes polarization much less likely than mechanistic psychology.

This is a plausible, rationalizing story for why so many psychological researchers have justifiably been convinced of mechanism. Perhaps you find it convincing, too. Perhaps you think this is the problem at the heart of rational psychology. Very well.

It was a trap.

This was once a rationalizing story. It is now a *debunking* story—a story that undermines our justification for our beliefs once we accept it. For this book has undermined the inference on which it is based. It is indeed true that, from our prior perspective, rational psychology made polarization much less likely than mechanistic psychology. But our prior perspective reasonably but incorrectly failed to account for the importance of ambiguity.

As I have shown, following through the consequences of uncontroversial facts about ambiguity reverses our conclusions. Taking the *exact same* rational Bayesian models and introducing the *most conservative* model of ambiguity—namely, failing to be certain of what our own prior opinions are—predicts that people will *not* converge to the truth. Granted, it will be controversial whether these models are fully 'rational'—we can imagine certainly *more ideal* agents who do not suffer from higher-order uncertainty or cognitive noise. But *we* do suffer from these limitations—and sensibly so. Ambiguous-Bayesian agents are uncontroversially instances of (bounded-optimal) rational psychology, since they respond optimally given their sensible limitations.

Thus, conditional on the conceptual and mathematical facts about ambiguity I have pointed out—summarize those as '*ambiguity*'—polarization is *just as much* to be expected given rational psychology as given mechanistic psychology:

$$P\Big(\textit{polarize}\ \Big|\textit{ambiguity}\ \&\ \textit{rationalism}\Big)\ \approx\ P\Big(\textit{polarize}\Big|\textit{ambiguity}\ \&\ \textit{mechanism}\Big)$$

Given what we now know about ambiguity, the high-level facts about predictable polarization are no longer more to be expected given mechanism than rationalism. As a result, those facts no longer provide strong evidence for mechanism, let along settling the debate. The strongest argument for mechanism has been undermined.

Consider an analogy. We look at a wall, see that it appears to be red, and conclude that it is a red wall, not a white wall. This is a completely rational, good inference. But suppose then we find out that the wall is surreptitiously bathed in red light—so it would've appeared red *even if* it were white. Then we would lose our justification for our confidence that it was red. We initially thought it was much more likely to appear red if it was red than if it was white—$P(\textit{appear red}|\textit{red}) \gg P(\textit{appear red}|\textit{white})$. But we also realized that *if* it were bathed in red light, it would look red whether or not it actually was red—its appearance wouldn't be diagnostic about its color: $P(\textit{appear red}|\textit{red light}\ \&\ \textit{red}) \approx P(\textit{appear red}|\textit{red light}\ \&\ \textit{white})$. So when we learn that it *is* bathed in red light, we lose our reason for thinking the wall is white—it would appear red either way.

This, I claim, is what has happened to our belief in mechanistic psychology. We initially thought we were much more likely to see polarization if mechanism were true than if rationalism were true. We looked around, saw polarization, and concluded that mechanism is true. But we also knew that *if* ambiguity broke the convergence results, we would see polarization whether or not mechanistic psychology was right—the prevalence of polarization wouldn't be diagnostic about our rationality. So when we learn that ambiguity *does* break the convergence results, we lose our reason for thinking that mechanism is true—people would be polarized either way.

Indeed, I've argued for a stronger conclusion. Chapters 7–10 claimed that when we look into the details of the empirical trends around the polarizing biases, they are *better* predicted by rational psychology under ambiguity than by mechanism under ambiguity. That is:

$$P\Big(\textit{polarize}\ \&\ \textit{biases}\Big|\textit{ambiguity}\ \&\ \textit{rationalism}\Big)\ >\ P\Big(\textit{polarize}\ \&\ \textit{biases}\Big|\textit{ambiguity}\ \&\ \textit{mechanism}\Big)$$

In other words, once we bring to bear more of the evidence—both uncontroversial conceptual and mathematical facts about ambiguity, and the detailed empirical trends surrounding polarization—we find that those trends in polarization are *better* predicted by rational psychology than by mechanistic psychology.

Not only properly accounting for ambiguity undermine the strongest arguments for mechanism—it generates a well-founded, testable, rational psychological theory that *better* explains why we're polarized.

11.5 Upshot

Mechanistic and rational psychology have been duking it out for decades. On many measures, mechanism seems to have won the day.

But this has been a mistake—a reasonable failure of convergence. It is mistaken methodologically: it unmoors behavioral science from theoretical constraints, exacerbating the replication crisis and generating a proliferation of proposed mechanisms and biases with little unification or coherence (e.g. Fudenberg 2006). It is mistaken on the merits: it fails to account for the range of empirical data, including not only our foibles but also our feats (e.g. Griffiths et al. 2024). And it is mistaken about what it needs to explain: yes, we are studying computational cognitive systems that talk and walk and sometimes slip; but we are also studying *ourselves*—and the theories we accept with need to be ones that we can sensibly believe about why *we* do what we do.

Whether judged on theoretical motivation, empirical adequacy, or first-person believability—let alone all three—a rational theories outperform mechanistic ones, once they take proper account of ambiguity. As with any theoretical dispute, further data or work might overturn these verdicts. But it is high time we put to bed the simplified story that psychology has shown that people—especially in politics, religion, and couples' quarrels—are hopelessly irrational.

Why does it matter? Because this simplified story—this 'irrationalist narrative'—has badly distorted our view of human judgment, especially of those humans who judge differently than ourselves. The next chapter will argue that we have drawn the wrong lessons from the prevalence of bias—and this is exacerbating the pathological politics of our age.