

Chapter 1

The Assumption of Convergence

Abstract

We often assume that when evidence is plentiful, reasonable people will converge to the truth. This assumption is buttressed by ‘Bayesian convergence theorems’, claiming to show that ideal reasoners would do so. Since *real* people often don’t do so, we conclude that they are often unreasonable. We’re wrong. Those theorems implicitly and essentially assume *clarity*: that reasonable people know exactly what they think. Yet our reasonable opinions are often *ambiguous*: we’re unsure what we think. This understanding of ambiguity as *higher-order uncertainty* matters. Theoretically, Bayesians with higher-order uncertainty will often fail to converge. Empirically, higher-order uncertainty offers a new, unified explanation of why we exhibit biases—one that dovetails with our best theories of the everyday feats of human cognition. Properly understood, ambiguity promises to reshape our view of both our ideological opponents, and ourselves.

1.1 Reasonable Convergence

We are in the grip of an idea. It guides our social science and pervades our politics. It is formalized in economics journals and weaponized by political pundits. It lives in the woodwork of our contested spaces—shaping public debates, office dramas, and couples’ quarrels. It is simple. It is compelling. It is wrong.

It is the idea that if evidence is plentiful, and people are reasonable, they will converge on the truth. At least, near enough. At least, in the long run. At least, usually—absent skeptical scenarios and wicked environments. Let’s call it:

Reasonable Convergence: When evidence is plentiful, if people take reasonable steps to figure out the truth, they’ll usually succeed.

This claim is usually regarded as an optimistic thesis about the power of rationality. It’s rarely recognized as a pessimistic thesis about the nature of humanity. For Reasonable Convergence implies that when evidence is plentiful but people *haven’t* converged, then they’re (usually) *not*

taking reasonable steps to figure out the truth. And just look around. For many of the questions we care about—from the grand ones of what we believe in and who we elect, to the communal ones of what we prioritize and who we hire, to the personal ones of what’s in the budget and who does more chores—we often don’t converge. We polarize.

The assumption of Reasonable Convergence shapes how we react to such disagreements. When we know that we’ve each had reasonable reactions to only part of the evidence, we tend to disagree respectfully and try to hash things out. Think of two scientists who work with different models, two colleagues who know different job candidates, or two friends who’ve heard opposite sides of a couple’s quarrel. We trust that, despite disagreeing, we’re each acting and thinking as we should. The other person is probably seeing things that we’re not.

But when we have plentiful evidence and yet persistently disagree, Reasonable Convergence warrants reactive attitudes. Day in and day out, we each get plentiful evidence about how many chores we each do. Given this, I think, it’s patently obvious that I do most of them. Then you say that *you* do most of them. Excuse me? Implicitly accepting Reasonable Convergence, I conclude that you aren’t taking reasonable account of who does what—or, worse, are self-deceived or acting in bad faith. Regardless, I think, you should know better—and so should *do* better: blame, resentment, and anger are warranted. Meanwhile, you’re thinking the same about me. So the next time you ask me to take out the trash, here come the fireworks.

That example is low stakes. But scale it up to our communities or nations, and the resulting animosities are anything but. People on the other side believe *that*? They voted for *him*? What’s *wrong* with them? They should know better. They should *do* better. And we have a mind to tell them. Reasonable Convergence is a load-bearing assumption in the pathological politics of our age.

The thesis of this book is that Reasonable Convergence is wrong. Not wrong around the edges. It is deeply, profoundly mistaken. It underestimates the ambiguities induced by trying to break down the big questions we care about into small pieces we can answer, and misunderstands the reasonable effects that doing so has on our beliefs. For *most* of our big questions, *most* of the time, taking reasonable steps toward the truth on those small pieces is no guarantee that we will converge to the truth overall. Very often, it’ll lead us to reasonably polarize.

My thesis is both normative and descriptive—making claims about both ideal agents and real people. On some readings, it isn’t very controversial; on others, it’s quite so.

Everyone can agree that, in some sense, *rational* people can fail to converge. After all, they might have no interest in getting to the truth, and great interest in being accepted by their in-group (Kahan et al. 2017; Klein 2020). For most of us, changing our beliefs about Donald Trump would not materially affect politics, but *would* materially affect *us*—it would strain or sever some of our friendships. So from the point of view of rationally pursuing their interests, Republicans and Democrats are rational to polarize.

This is true, but unsurprising. Distinguish *practical* rationality from *epistemic* rationality. Practical rationality is taking sensible steps to pursue your interests. Epistemic rationality is taking sensible steps to figure out the truth. Obviously practically-rational people—even in the face of plentiful evidence—often don’t converge. My thesis is that *epistemically*-rational people often don’t converge. That is: even if people have plentiful evidence and take sensible steps to figure out the truth, they will *not* usually converge to it; quite often, they’ll polarize. I prefer the term ‘reasonable’

because it more reliably evokes the epistemic side of rationality: it's not reasonable to believe that there's a (literal) elephant in my room, regardless of how much incentive you have to believe it.

So my claim is that taking (epistemically) reasonable steps to figure out the truth will often lead, predictably, to polarization. There are readings on which this is only somewhat-more controversial. For 'reasonable' suggests an *attainable* level of rationality; the sort we can expect from real people. And most will admit—when they think about it—that reasonable humans *do* often fail to converge in the face of plentiful evidence.

Consider. There is plentiful evidence that there are no talking donkeys. There is plentiful evidence that markets provide better standards of living than command economies. There is plentiful evidence that life in 2035 will look more similar to life today (in 2025) than life today does to life in 1800. I hope you'll agree that, on at least one of these claims, I'm not going out on a limb. But for each claim, there are smart people who have spent far more time thinking about the issue than me (or you) who confidently believe the opposite. David Lewis thought there were talking donkeys, since there *could have been* talking donkeys—and every way the world could have been is the way that some world actually is (Lewis 1986). Paul Cockshott and Allin Cottrell think that the problem with command economies was merely computational, and that with modern computers and better data they would match or outperform markets (Cockshott and Cottrell 1993). And some AI researchers believe that the societal impact of AI over the next decade will exceed that of the industrial revolution (Kokotajlo et al. 2025). The list could go on.

No one should deny that these are smart people, doing their best to reasonably think things through. But most should agree that they have radically diverged from the truth. So: when smart, reasonable people—in the non-ideal, humanly-attainable sense of 'reasonable'—embark on large research projects, it's normal for them to polarize. Witness academia.

More controversially, my thesis is that the processes that drive zany academics to polarize over their research are remarkably similar to the processes that drive regular people to polarize over politics, religion, hiring, and who does more chores. In each case there are innumerable-many small bits of evidence that people are remarkably *good* at assessing—but doing so induces ambiguities that lead, predictably, to reasonable polarization over the big question they care about.

Most controversially, my claim is not (just) about humanity. It's also about ideal rationality—applying to any new agents we might conjure up. (I'm looking at you, AI boosters.) The gold-standard of ideal rationality is Bayesianism. My claim is that if Bayesian agents both (1) started with the uncertainty ('priors') that we reasonably start with, and (2) ideally searched for and updated on the sorts of evidence that's available to us, then they would often predictably polarize in very much the way that *we* do.

In fact, I'll argue that the hypothesis that real people approximate such Bayesians offers a better explanation of many human foibles—including hindsight bias, confirmation bias, and overconfidence—than top-line models from psychology and behavioral economics. Moreover, it does so in a way that dovetails with top-line cognitive-scientific explanations of many human feats, offering a resolution to a decades-old tension in our understanding of human cognition. Assuming Reasonable Convergence, we've been misled into thinking that real people must be systematically irrational; how else could they wind up where they do? But Reasonable Convergence is false—and reasonable biases are inevitable—even for ideal reasoners. The bad news is that rationality buys us

far less than we've thought. The good news is that real people are far more rational than we think.

But wait. How could this be? When Reasonable Convergence is formalized in economics journals, it takes the form of *Bayesian* convergence results. It's a theorem that Bayesians who start out with reasonable uncertainty and receive plentiful evidence will converge to the truth. Right?

That depends. What type of 'reasonable uncertainty' do they start with?

1.2 Clarity and Ambiguity

Life is uncertain. We're constantly wondering who'll win the next election, what we'll have for dinner tomorrow, and how many chores we did in the past week. Bayesianism is a theory—perhaps *the* theory—for forming reasonable beliefs and decisions, given such uncertainty.

Bayesian agents are mathematical models of such uncertain scenarios. Their uncertainties—aka *subjective probabilities* or **credences**¹—are modeled with a probability function that makes fine-grained distinctions in their degrees of confidence, varying between 0–100%. How likely is a fair die to land on 1? $\frac{1}{6}$, or around 17%. How likely is it to land on 1, 2, or 3? $\frac{1}{2}$, or 50%.

Any claim e that a Bayesian learns is treated as evidence for hypotheses that make e more likely, and evidence against hypotheses that make e less likely. Learning that the die *landed on an even number* (2, 4 or 6)—call that e —is evidence for the hypothesis that *it landed on a high number* (4, 5, or 6)—call that h . Why? h would make it $\frac{2}{3}$ likely that e , since two of the three possibilities in h (4 and 6) are even, which is more likely than e 's prior probability of $\frac{1}{2}$. By the same token, e is evidence against the hypothesis that *it landed on a low number* (1, 2 or 3), since that hypothesis would make it $\frac{1}{3}$ likely that e .

Faced with a decision, a Bayesian chooses the option that would be best, on average, if done repeatedly with the probabilities that they have. (Best 'in expectation'.) Suppose they're offered a bet that pays +\$1 if a die lands on 1, 2, 3 or 4, and costs -\$1 if it lands on 5 or 6. If they think the die is fair then they'll take the bet, since they expect it to (on average) net out positive: if done 99 times, they expect to win around $\frac{4}{6}(99) = 66$ dollars and lose around $\frac{2}{6}(99) = \$33$, yielding around +\$33.

By the end of this book, you'll be all too familiar with Bayesianism. For now, here's what you need to know.

There are approximately a zillion arguments that (some form of) Bayesianism is the reasonable way to form beliefs and make decisions under uncertainty. In many senses, acting like a Bayesian is the optimal solution to responding to new evidence, given the beliefs you started with. Bayesianism is well-founded.

At the same time, Bayesianism is widely applicable. Real life is full of cases where we have degrees of uncertainty that vary subtly in response to small bits of evidence. Did your job application get passed over? If you haven't heard back after 2 days, you think it's possible. After 3 days, you think it's a tiny bit more likely. Each day increases your probability just a bit; after 4 weeks (or 4 months, in academia) you think it's certain. Likewise, your uncertainty about who the department will hire or whether Chipotle is too busy fluctuates with each conversation with a colleague and

¹Technical terms are bolded when defined; their definitions are collected in the Glossary.

each additional person in the line when you arrive at 12:04pm. No wonder that Bayesian models of real people's beliefs and decisions are so common in the social and behavioral sciences. As model of *rational* beliefs and decisions, they are dominant.

Their dominance has shored up the intuitive case for Reasonable Convergence. For an essential part of the social-scientific folklore is that Bayesian agents who start out with reasonable priors will converge to the truth, given enough evidence. For example, suppose the die is weighted so that one of the faces is slightly more likely to come up than the others. Or perhaps it has a shifting weight inside, so that certain *patterns* of faces are more likely to come up. Or perhaps something more-subtle still. So long as our Bayesian agent starts with a clear prior probability function that leaves open (assigns positive probability to) the true hypothesis, then if they observe enough rolls of the die they will almost-certainly become almost-certain of the truth (see e.g. Savage 1972, §3.6). The result can be generalized in many ways, leading to the common refrain: if evidence is plentiful, and people are reasonable (Bayesians), they'll converge to the truth. Right?

Wrong. It turns out that, for many scenarios, we've implicitly made unreasonable assumptions about our Bayesians' priors. Theorists standardly focus on the assumption that the prior must leave open the true hypothesis. They rarely notice—or even explicitly state—the assumption that the prior must be *clear*. What I mean is that the Bayesian's prior probability function is not itself an object of uncertainty. As they're standardly modeled, Bayesian agents are uncertain about the world, but (usually implicitly) *certain about their own beliefs* about the world. They don't know whether the die is biased toward 1 or toward 2, but they do know that their prior assigned (say) 18% to the former and 21% to the latter. This turns out to be crucial. As we'll see later: virtually every Bayesian convergence result assumes clarity. Without clarity, the convergence theorems fail.

And as we can see already, the assumption of clarity is where Bayesian detractors start to balk. We obviously *do* have subtly-varying degrees of uncertainty about whether Chipotle is too busy. But Bayesians seem to be saying that, if I'm reasonable, I must be able to tell you *exactly* how confident I am that it is. (36.4%, or 36.5%?) That's nuts.

Granted, *sometimes* our judgments under uncertainty are clear, in the sense that we know exactly how uncertain we are. But more often they're *ambiguous*, in the sense that we are unsure what we think. Consider some examples:

Clear Chances: I tossed a fair coin. How likely do you think it is that it landed heads?

That's easy: 50%. Not 60%, nor 40%, nor even 51%. *Exactly* 50%. Given just this information, you'd never give a different answer. Contrast:

Ambiguous Chances: ' _FO__ ' is a word-completion task that can be completed with an English word. Maggie—who plays the New York Times's 'Spelling Bee' every day—has 7 seconds to find it. How likely do you think she is to succeed?

That's hard. Maybe 50%. But maybe 60% (she does word searches every day), or maybe 40% (7 seconds isn't long). It's unclear—*ambiguous*—what you think. Given just this information, you could easily have said that you think it's 60% or 40% or 51% likely.

The distinction between clear and ambiguous judgments is not simply whether you know the objective probabilities (the **chances**, in philosophers' terminology). Example:

Clear Ignorance: The digits in the decimal expansion of π begin 3.1415...—the first is 3, the second is 1, the third is 4—and go on forever. There are 10 possible digits (0–9) for each spot. How likely do you think it is that the 99th digit of π is between 1–6 (inclusive)?

That’s easy: 60%. Not 70%, nor 50%, nor even 61%. *Exactly* 60%. You don’t know the objective chances—those are either 0% or 100%. Yet you know that the only fact you have to go on is that 1–6 covers 6 of 10 possible options. Given just this information, you’d always say 60%. Contrast:

Ambiguous Ignorance: How likely do you think it is that I own (at least) a dozen spoons?

That’s hard. Spoon ownership varies widely with age, demographics, etc. Maybe 60%? But who knows. Given just this information, you could easily have said that you think it’s 70% or 50% or 61% likely.

The distinction between clarity and ambiguity is also not determined by how much information you have. *Sometimes* adding information makes your judgment clearer:

Clear Deference: My brother thinks I own 8 spoons. Given this, how many do you think I own?

That’s easy: 8. Not 5 nor 17 nor even 9. After all: you know that my brother knows more about my spoons than you do, and you have no basis for second-guessing him. Given just this information, you’d always say 8. Yet adding more information can make your judgment *less* clear:

Ambiguous Deference: My sister-in-law thinks I have 18 spoons. For what it’s worth, my brother thinks that’s too high—he maintains his estimate of 8. She thinks that’s too low—she maintains her estimate of 18. They disagree. Given this, how many spoons do you think I own?

That’s hard. Maybe 13? But my brother has known me longer—maybe he knows more about my spoon habits. Then again, he’s an absent-minded philosopher—my sister-in-law might have a better eye for detail. Given just this information, you could easily have said that you think I own 10 or 17 or 15 spoons.

I intend the contrast between ‘clarity’ and ‘ambiguity’ to pick out an intuitive phenomenon illustrated by a set of cases and unified—at a first pass—by the difficulty and felt-arbitrariness in supplying precise degrees of confidence.

One way this contrast shows up is through *cognitive noise*. Clear judgments contain no whiff of randomness: every time you’re asked about how likely it is that a coin landed heads, if all you know is that it’s fair you’ll say 50%—never 60% or 51%. In contrast: when asked how likely Maggie is to find a word, you could easily have said a slightly different number; and if you reconsider the question later, you might well do so. Such cognitive noise is widely-studied within cognitive science (see Ch. 2).

The contrast also shows up through *self-doubt*. We’re confident in our clear judgments. We might not know that our decision will work out for the best, but we *do* know that we took reasonable risks, given our information. Would you rather win \$100 if (1) *the fair coin I tossed landed heads*, or if (2) *the 99th digit of π is between 1–6*? You’ll confidently choose (2), since you know that you should think it’s more likely to pay out. In contrast, we’re doubtful of our ambiguous judgments. Would you rather win \$100 if (3) *Maggie finds a word within 7 seconds* or if (4) *I own a dozen spoons*? Perhaps (4)—after, all you, said ‘60%’ to it and only ‘50%’ to (3). (Remember?) But maybe that’s taking your arbitrarily-chosen numbers too seriously. You might be making a mistake

by your own lights; perhaps you *don't* (reasonably) think it's more likely that I own a dozen spoons than that Maggie will find a word. Such *epistemic modesty* is widely-studied within epistemology.²

You get the idea. With fair coins and random samples, Bayesians are used to confidently supplying precise numbers for which subjective probabilities it's reasonable to have, as well as what estimates to make on their basis. But in everyday cases, it's extremely difficult to put precise numbers on our opinions (even experts are averse to doing so; see Tetlock 2005), and any particular number feels arbitrary. This discrepancy between Bayesian theory and everyday practice is a common source of skepticism about the whole approach. And notice that it's not just about the numbers. It's not merely that you're unsure what percentage-confidence to assign to Maggie finding a word; it's also that you're unsure whether you're more confident that (3) she'll find a word or that (4) I own a dozen spoons. That's what made you unsure which possibility you prefer to bet on. Ambiguity makes confidence-*comparisons*—and thereby *decisions*—difficult too.

These difficulties lead some to doubt the importance of such fine-grained degrees of confidence for understanding human beliefs and behavior. Tellingly, such skepticism rarely survives long in the behavioral sciences: when it comes to the empirical task of explaining and predicting how real people think, talk, and act, it's not an overstatement to say that probability(-like) judgments are the only game in town.³ So I invite you to suspend disbelief, and consider how a probabilistic theory of cognition might *explain* the discrepancy between Bayesian theory and everyday practice.

1.3 Higher-Order Uncertainty

What's my theory? Why would our opinions be difficult to articulate, hard to compare, and feel arbitrary when expressed precisely? *Because we don't know them.* Or, more carefully: *Because we're unsure what they are.*

In a sense, every theory of ambiguity agrees. Patently, we don't know what precise number to write down, or how to precisely compare our levels of confidence. So all agree: we don't know such facts. How to explain that?

The standard next step is to say that *there are no such facts.* We don't have opinions that are fine-grained enough to correspond to a number, or to warrant precise comparisons. Maybe we have 'imprecise' probabilities best modeled with a *set* of probability functions (Ellsberg 1961). Maybe we have only a comparative confidence ordering of propositions (Koopman 1940a,b; Scott 1964). Maybe we have only binary (on/off) beliefs, which can be about probabilities but often aren't (Holton 2017; Byrne 2022). Maybe we have nothing of the sort, and instead get by with a hodgepodge of heuristics (Tversky and Kahneman 1974). Maybe something else.

That step was a leap. It pointed out that we don't know certain facts about ourselves, and

²Along with its correlate, 'higher-order evidence'; see e.g. Christensen 2010; Lasonen-Aarnio 2013, 2014, 2015; Horowitz 2014, 2019; Schoenfield 2015, 2018; Sliwa and Horowitz 2015; Dorst 2019, 2020a; Fraser 2021. For a summary, see Dorst 2020b.

³This isn't to say that *probabilities* are the only game in town—just that every predictively-successful alternative has some comparably-fine-grained analogue of subjective probabilities. See e.g. Thurstone 1927; Luce 1956, 1959; Ellsberg 1961; Quiggin 1982; Buchak 2013; Loomes and Sugden 1982; Griffin and Tversky 1992; Laibson 1997; Koriat 2012; Vul et al. 2014; Woodford 2020. Chapter 3 (§3.3) will discuss Bayesianism's main rival, prospect theory (Kahneman and Tversky 1979).

concluded that there are no such facts. Often that’s silly. You don’t know your current blood pressure or heart rate or body-temperature set point. Any numbers you write down or precise comparisons you make would feel arbitrary. Patently, it doesn’t follow that there are no such correct numbers or comparisons (Easwaran 2024). Let \mathbf{P}_a be your subjective probability function at the actual state of the world \mathbf{a} . Let X be some quantity about you, like your blood pressure. Often you’re unsure of X ’s value—there’s no number x such that $P_a(X = x) = 1$ —despite the fact that X has a precise value: for some x , $X = x$.

My claim is that if we substitute *your subjective probability for q* in for the quantity X , this is what’s happening under ambiguity. To state this precisely, we need some notation. Let \mathbf{P} be a *variable* probability function that picks out the subjective probability function you have at each possible state of the world (‘world’) w , with \mathbf{P}_w a constant for the probability distribution you have at w . Think of ‘ \mathbf{P} ’ like a *description*, and ‘ \mathbf{P}_a ’ and ‘ \mathbf{P}_w ’ like *names* (Kripke 1963). \mathbf{P} needs to be a variable so that we can model people’s uncertainty about it. For example, when I’m unsure whether you assign 0.6 or 0.7 probability to me owning a dozen spoons—call that d —I’m unsure whether I’m in a world w where you assign 0.6 (where $P_w(d) = 0.6$), or a different world u where you assign 0.7 (where $P_u(d) = 0.7$). The value assigned to d by your probability function, $P(d)$, is an object of uncertainty for me—so it varies across possibilities I leave open. That’s why \mathbf{P} is a variable.⁴

My proposal is that your judgment about q is ambiguous iff you are (probabilistically) uncertain about what your own probabilistic judgment is:

Ambiguity as Higher-Order Uncertainty:

Your judgment about q is **clear** when you are certain of what subjective probability is—i.e. you have *higher-order certainty*. (There is an x such that $P_a(P(q) = x) = 1$.)

Your judgment about q is **ambiguous** when you’re uncertain of what your subjective probability is—i.e. you have *higher-order uncertainty*. (For any x , $P_a(P(q) = x) < 1$.)

You have a clear judgment about whether *the 99th digit of π is between 1–6* (p) in the sense that you are sure that your credence is 60%: $P_a(P(p) = 0.6) = 1$. You have an ambiguous judgment about whether *I own a dozen spoons* (d) in the sense that you are unsure what your credence is—you leave open that it might be (say) either 0.55 or 0.45: $P_a(P(d) = 0.55) > 0$ and $P_a(P(d) = 0.45) > 0$.

Some parts of this proposal are essential; others are incidental.

An incidental feature is that I’ll interpret \mathbf{P} as a description for ‘your *reasonable* subjective probabilities’, assuming that you know that your probabilistic opinions \mathbf{P} are reasonable, but are often (under ambiguity) unsure what they are. Thus when you are unsure what you think, you are thereby uncertain what’s *reasonable* to think and do—capturing discussions of ‘epistemic modesty’ and the possibility ‘higher-order evidence’ familiar to philosophers. But many lessons of this book hold up under different interpretations of \mathbf{P} .⁵

⁴See Chapters 4 and 5, as well as Schervish et al. 2004; Williamson 2008; Schoenfield 2016, and Dorst 2019, 2020b.

⁵What’s crucial is that there is a single probability function \mathbf{P} that is unsure of its own values. If we let \mathbf{P} be *the rational credence function for you*, then it can model someone who knows what their *actual* credences are, but is rational to be unsure about the rational ones to have (Dorst 2020a). Chapter 2 will show how, due to the difference between your true probabilities and your noisy elicitation of them, my interpretation can capture a version of this.

An essential feature is that you have the *same sort* of uncertainty about first- and higher-order claims—but it’s incidental that this uncertainty is modeled with exact (real-valued) probabilities. Presumably our degrees of belief are *also* inexact (‘imprecise’, in philosophers’ terminology) in the sense that they don’t admit complete comparisons, and so are often better modeled with (something like) a *set* of probability functions. What I’m proposing is that (1) the core features of ambiguity are better-explained by higher-order uncertainty than inexactness (Carr 2020), and (2) we often have higher-order uncertainty about what our (exact or inexact) probabilities are. I’ll use real-valued probability functions for simplicity and conservativeness; but many of the arguments of this book will generalize to the inexact case. I understand my use of real numbers as a scientific idealization—like point-particles on frictionless planes—not a normative one.⁶

It’s essential to distinguish my notion of ambiguity from two apparently-nearby alternatives. The first, made famous by Ellsberg (1961), Levi (1974), and Seidenfeld and Wasserman (1993), says that under ambiguity we *can’t assign* exact probabilities to events, leaving open a range of possible probabilities;⁷ or if we can, they don’t guide our action—instead we assign more ‘weight’ to some probabilities than others.⁸ My approach differs in both formalism and emphasis. Formally, my ‘Ambiguous Bayesians’ have exact probabilities for all events, but have (exact) probabilistic uncertainty about what those probabilities are. This permits stronger normative foundations (Chs. 5–6), while altering decisions by making them noisy (Ch. 2) and altering belief-updating by inducing biases (Chs. 7–10). While that approach focuses on ‘ambiguity aversion’—people’s preference to bet using known probabilities rather than uncertain ones—my approach focuses on why higher-order uncertainty generates inevitable biases (like hindsight bias and confirmation bias) that lead to failures of Reasonable Convergence. Though it is not the core focus, my approach can also explain forms of ambiguity aversion (see §5.7.2 and §[XXX]).

The second alternative says that under conditions of ambiguity, your subjective probabilities P are unsure what a *more informed* (or: *more rational, more reflective, more insightful...*) version of your probabilities—label that \mathcal{P} —would be.⁹ This includes *Hierarchical Bayesian Models* (HBMs) in cognitive science. For example, you might be unsure what the bias of a coin is, and so unsure what probability function \mathcal{P} you’d have if you learned the bias. This is consistent with being certain of what subjective probability distribution P over the possible biases you *currently* have. Indeed, it is standard to model \mathcal{P} as a variable but P as a constant, implicitly guaranteeing that you update as if you’re certain what probability function you have (see Ch. 3). In such a case, by marginalizing (averaging) over the possible biases using your current probabilities P , you will be certain of what your current probability is that the coin will land heads. For example, if your distribution P is uniform over all possible biases of the coin between 0 and 1, and you defer to them

⁶Suppose we use a set of probabilities \mathbb{P} to model your opinions (e.g. Ellsberg 1961; Levi 1974; Seidenfeld and Wasserman 1993; Joyce 2010; Schoenfeld 2012; Trautmann and van de Kuilen 2015; Moss 2018). Then (following Molinari 2023, 2025), my claim is that under ambiguity, for all sets Γ : $\mathbb{P}_a(\mathbb{P}(q) = \Gamma) \neq \{1\}$.

⁷E.g. Walley 1991; Seidenfeld 2004; Moss 2018; Konek 2019a,b, 2023

⁸E.g. Einhorn and Hogarth 1985; Camerer and Weber 1992; Klibanoff et al. 2005; Etner et al. 2012; Baliga et al. 2013; Machina and Siniscalchi 2014.

⁹So they say that for all x , $P_a(\mathcal{P}(q) = x) < 1$. See e.g. Frisch and Baron 1988; Schoenfeld 2012; Hedden 2019, as well as Hierarchical Bayesian Models that represent what your probabilities would be were they to be updated on the value of some unknown parameter(s) (e.g. Pearl 1988; Henderson et al. 2010; Perfors et al. 2011; Ullman and Tenenbaum 2020), or what your future probabilities might be (van Fraassen 1984, 1995; Gaifman 1988).

? change to critique? Critique is both normative and empirical

via the Principal Principle (Lewis 1980), then since the mean of that distribution is 0.5, you’re certain that your subjective probability that it’ll land heads on the next toss is 0.5.

Call cases like this (**mere**) **probabilistic uncertainty** about the value of some relevant probability function, and contrast it with genuine **higher-order uncertainty** about what your own subjective probability function is. The difference is crucial. Why?

1.4 Why Ambiguity Matters

Almost all models of mere probabilistic uncertainty are what I’ll call ‘Standard-Bayesian’ models (Ch. 4). This amounts to the claims that (1) their prior subjective probability function is clear (i.e. not uncertain), and (2) they update their beliefs by ‘conditioning’ this clear prior on the true answer to a question (i.e. the true cell of a ‘partition’, i.e. the true value of a variable).

The details don’t matter just yet. What matters is three things.

First, such Standard-Bayesian agents always have clarity: their priors are clear, and they always know how they’ve updated them in response to evidence—so their posteriors are clear too.

Second, such Standard-Bayesian models are the ones used to prove formalizations of Reasonable Convergence. With enough evidence, *Standard* Bayesians will converge to the truth (e.g. Savage 1972, §3.6; Blackwell and Dubins 1962), become well-calibrated in their estimates (Dawid 1982), avoid forms of confirmation bias (Kadane et al. 1996; Salow 2018), and—if they share priors—be unable to ‘agree to disagree’ no matter what (different) evidence they’ve received (Aumann 1976). We’ll discuss these results and more at length in Chapters 7–10. The crucial point is that no model of ambiguity as (mere) probabilistic uncertainty can explain the failures of convergence that call out for explanation. Whatever else such models get right—and they do offer much insight into the sorts of ‘biases’ that even ideal reasoners can exhibit¹⁰—they can’t explain real people’s tendency to fail to converge to the truth (or even agreement) despite plentiful evidence.

Third, clarity is essential to these results. Suppose we drop clarity, but otherwise keep our models exactly the same: our Bayesians have exact probabilities about all events, and they always update by conditioning on the true answer to a question. Then they will often *not* converge to the truth—very often, they will predictably polarize—despite being exposed to arbitrarily-large amounts of reliable evidence. Under such ambiguity, Bayesians inevitably exhibit hindsight bias (Ch. 7) and confirmation bias (Ch. 8), leading them to be at constant risk of polarizing away from the truth and then agreeing to disagree (Ch. 9), and of becoming radically miscalibrated in their estimates and forecasts (Ch. 10).

In short: the exact same Bayesian models used to prove Reasonable Convergence under clarity show why it *fails* under ambiguity. When evidence is plentiful but people’s opinions are ambiguous,

¹⁰For example, Standard-Bayesian models have been used to show that ideal reasoners will often think that arguments favoring their side tend to be stronger (Feeney et al. 2000; Hahn and Oaksford 2007; Jern et al. 2014; Benoît and Dubra 2019; Gershman 2019; Henderson and Gebharter 2021); will often seek confirmatory instances of their hypotheses (Oaksford and Chater 1994, 2003; Navarro and Perfors 2011; Hahn and Harris 2014); will sometimes ask ‘nondiagnostic’ questions (Feeney et al. 2008; Crupi et al. 2009); will usually be miscalibrated on tricky questions (Juslin 1994; Juslin et al. 2000; Moore and Healy 2008; Moore 2020); and will choose candidates in ways that systematically disadvantage less-legible groups, despite a pure concern for accuracy (Phelps 1972; Aigner and Cain 1977; Cornell and Welch 1996; Hedden 2021).

then even if they take reasonable steps to figure out the truth, they will *not* usually succeed.

Those are bold claims. They raise three questions:

1. Why do these theoretical claims matter?
2. How could they be true?
3. If true, why has ambiguity been overlooked?

It will take an entire book to fully answer these questions. But let's give it a start.

(1.) *Why does it matter that, for ideal Bayesians, Reasonable Convergence fails under ambiguity?*

Consider the sorts of opinions that real people polarize about:

Politics: How likely do you think Republicans with a trifecta are to cause a recession? Democrats?

Race: James and Jamal have similar resumes, but James is white and Jamal is black. How likely do you think each is to be hired by the Rowe and Flint law firm?

Gender: How likely do you think it is that biology influences the distribution of women and men's career interests?

Economics: Piketty et al. 2018 estimate that in 2014, the after-tax income share of the top 1% of U.S. earners was 15.7% and rising; Auten and Splinter 2023 estimate that it was 9.1% and steady. Given that disagreement, what's your estimate?

Religion: How likely do you think it is that God exists?

Character: What proportion of Professor X's comments do you think are insensitive? Insightful?

Labor: What's your estimate for the proportion of the household chores you do?

The behavioral sciences have had a field day with such judgments. Subfields have been built, books written, and Nobel prizes won on the basis of the claim that our judgments about such topics are riven with irrationality.¹¹

And who could disagree? We *are* prone to bias, overconfidence, and polarization on such topics. We *don't* converge to the truth, even when there's plentiful evidence. If Reasonable Convergence should be expected from such opinions, then they *must* be riven with irrationality.

But take a moment. Try to supply some numbers. Try to make some comparisons. Which do you think is more likely: that God exists, or that Jamal will be hired? Which proportion do you estimate is higher: the number of chores you do, or the after-tax income share of the top 10% (rather than 1%)? Obviously *you don't know* exactly what you think. Any numbers you supply or comparisons you make will feel arbitrary. Your judgments about such topics are ambiguous.

Thus when we compare the way we polarize to the way that Bayesians *under clarity* would converge—when we hold humans against the normative standard of Reasonable Convergence—we are using the wrong yardstick. To assess the evidence for human irrationality, the relevant question is whether ideal agents *who started with the opinions that we did* would converge. Our opinions were always ambiguous. So the relevant question is whether Bayesians *under ambiguity*—who started out unsure what their own opinions were, and had access to the sorts of evidence that we have access

¹¹E.g. Tversky and Kahneman 1974; Kahneman et al. 1982; Sutherland 1992; Lakoff 1997; Fine 2005; Ariely 2008; Thaler and Sunstein 2009; Hastie and Dawes 2009; Kahneman 2011; Haidt 2012; Thaler 2015; Lewis 2016.

to—would converge to the truth. If I’m right, they would not. The case for human irrationality rests on a mistake.

This isn’t just high theory. I will argue that Bayesians under ambiguity would exhibit the sorts of biases and fall into the sorts of polarization that we fall into. In fact, the *ways* that they’d do so turn out to be remarkably similar to the ways that we do so. Gathering lists of empirical trends from the vast literatures on risk attitudes, hindsight bias, confirmation bias, polarization, and overconfidence, I will run simulations to show that Bayesians under ambiguity would display most if not all of the trends in these biases that real people display (Chs. 3 and 7–10).

This also isn’t just retrodiction. Ambiguous-Bayesian models make many new and surprising predictions—for instance, that hindsight bias will be eliminated by reducing self-trust, that confirmation bias is driven by ambiguity asymmetries, that local accuracy improvements lead to global polarization, and that all of these trends will be reduced or eliminated when we induce clarity in people’s opinions. I will report on new (pre-registered, high-powered) experiments that confirm each of these predictions.

In short: when it comes to empirical coverage, the hypothesis that people approximate ideal Bayesians *under ambiguity* rivals top-line irrationalist models of bias from psychology and behavioral economics. When it comes to theoretical motivation and consonance with cognitive-scientific explanations of many human feats (§1.7), there is no comparison. Attending properly to ambiguity reveals that the case for human irrationality—even in politics, religion, and social judgment—is remarkably weak. That’s why ambiguity matters.

1.5 How Ambiguity Matters

(2.) *How could ambiguity matter so much to the behavior of ideal Bayesians?*

To answer that, we need to know why Bayesians *under clarity* inevitably converge to the truth. This can seem puzzling. The convergence theorems make no mention of a Bayesian’s motivations, nor how those might influence their search for evidence. Suppose that, in fact, you do 40% of the chores. Then the convergence theorems imply that, under clarity, *even a Bayesian who searched relentlessly for evidence that you did most of the chores* would converge to the truth that you don’t. Why?

Suppose, for example, our clear Bayesian asks you to construct a compelling case that you *do* do most of the chores, and (in an unfair manner) doesn’t give your partner a chance to make their case. The Bayesian listens as you list off chores you did and chores your partner forgot to do. Will this predictably increase their estimate for how many chores you did?

No. Bayesians update on their total evidence—they incorporate everything that they can be certain of in their interpretation of new evidence (Ch. 4). Thus Bayesians always update on evidence *relative to their expectations*. If the case you construct is weaker than they expected you to be able to make (‘I did the laundry yesterday and the dishes the day before! ...hmm. That’s all I can remember.’), then hearing the argument will *lower* their estimate for the number of chores you do. This is true under both clarity and ambiguity.

But under clarity, one of the things a Bayesian is certain of—part of their total evidence—is *their own prior expectations*. How do they expect their search for evidence—asking you but not

your partner to make an argument—to skew the results they see? Clear Bayesians know the exact answer to that: they know exactly how strong they expect your argument to be if in fact you do 40% (or 50%, or 60%) of the chores. As a result, their prior expectations are a fixed standard against which they can judge the force of your argument. Nothing you say—including your excuses or distractions—will lead them to rethink what they expected. If you say, ‘I can’t remember whether I did the laundry last week. But I’m so absent-minded—that doesn’t mean much’, that will never lead a clear Bayesian to think, *Ah, I didn’t really expect him to remember that* if in fact they did expect you to remember it.

In other words: observing the result of a search for evidence for q never leads a clear Bayesian to exhibit *hindsight bias* on what they expected from the search (Ch. 7). Observing that you failed to find the sought evidence for q doesn’t shift their estimate about how likely they thought you would be to find the evidence if q were true. Clear Bayesians thus have a clear standard against which to judge the evidential import of your failure to find the evidence. When you fail to report information that they expected you to be able to report if q were true, they substantially lower their probability for q . This implies that even searching for evidence in a biased manner doesn’t, on average, predictably skew a clear Bayesian’s beliefs (Ch. 8). That fact is essential to the convergence results (Chs. 9–10).

The through-line of this book is that ambiguity induces cascading failures in this line of reasoning. Ambiguity inevitably leads Bayesians to exhibit hindsight bias (Ch. 7); such hindsight bias induces confirmation bias (Ch. 8); and such confirmation bias leads to failures of convergence (Chs. 9–10). The detailed reasoning will have to wait. Here’s a compressed, informal version.

Consider Amber, an ideal Bayesian who has ambiguous opinions. Due to ambiguity, Amber *isn’t sure what she herself expects*. Thus she doesn’t have a clear standard against which to judge the force of the evidence she receives—since she isn’t sure what her prior P is, evidence she receives might shift her estimate for what it was. As a result, when she observes that you failed to find evidence you were looking for, that (i) lowers her estimate for how likely *she* thought you would be to find the evidence, exhibiting hindsight bias. This in turn (ii) dampens the disconfirming effect that your failure to find the evidence has on her probability in q , leading to confirmation bias.

Why does she (i) exhibit hindsight bias, lowering her estimate for how likely she thought you would be to find the evidence? Suppose she asks you find evidence that you did the laundry last month—a clear memory of when and how you did it, a text message mentioning that you did it, or something like that. Suppose you fail to come up with any such evidence. What effect does this have on her beliefs?

Consider what would happen in a *third*-person case. Suppose Amber begins unsure how likely *Chris* thinks it is that you would find the evidence. Amber trusts *Chris*’s judgment—she knows that he knows your texting habits and how good your memory is, so she thinks that his estimates are correlated with the truth on such matters. That is: she were to learn that *Chris* is *confident* you’ll be able to find the evidence, that would boost her probability that you will; if she were to learn that *Chris* is *doubtful* you’ll be able to find the evidence, that would lower her probability that you will. Correlations are symmetric. So if—without consulting *Chris*—Amber learns that *you didn’t find the evidence*, that provides Amber with (some, inconclusive) evidence that *Chris didn’t expect you to find the evidence*. (‘*Chris* probably suspected he wouldn’t have a clear memory.’)

Too heavy-going?

Under ambiguity, a *Bayesian's own expectations* are—even to themselves—a bit like those of another person's, like Chris's. Amber is often unsure what her own expectations are, but she also (to some degree) trusts those expectations: she thinks they are correlated with the truth. In such a scenario, the exact same reasoning implies that observing that *you didn't find the evidence* provides Amber with (some, inconclusive) evidence that *Amber herself* didn't expect you to find the evidence (see Ch. 7). That's hindsight bias.

Why does this in turn (ii) dampen the disconfirming effect that your failure to find the evidence has on Amber's probability in q ? In general, the effect that you failing to find evidence for q has on Amber's probability for q depends on how likely she thought you were to find the evidence if q were true vs. if q were false. For example, if she asks you to find evidence that you did a chore within the last week and you can't remember any, she'll dramatically lower her probability that you do most of the chores—if you did, almost certainly you would've had a record or clear memory of doing a chore within the last week. On the other hand, if she asks you to find evidence that you did a chore on this day *last year*, and you can't remember any, she'll only slightly lower her probability that you do most of the chores—even if you do, you almost certainly wouldn't have any record or memory of last year.

Under ambiguity, whenever you fail to find evidence for q , that provides Amber with some ('higher-order', inconclusive) evidence that she didn't expect you to find such evidence. In other words, it provides her with (misleading or probative) evidence that your failure to find evidence *wasn't very strong evidence* against q . It makes her wonder whether the search was more like searching for evidence of chore-doing *last year* than *last week*. Thus, unlike clear Bayesians and just like real people, searching for evidence for (vs. against) q can on average predictably skew a Bayesian's beliefs under ambiguity—exhibiting confirmation bias (Ch. 8).

This in turn means that, under ambiguity, people's searches for evidence—and, through social and media networks, their exposure to *other* people's searches for evidence—can have predictable and dramatic effects on whether and how they'll polarize (Chs. 9–10). Ambiguous Bayesianism gives a theoretical basis for the fact—emphasized by recent 'zetetic' epistemology,¹² and appreciated throughout much of the social sciences—that how we search for evidence is deeply important for where we end up. That's how ambiguity matters.

1.6 Why It's Been Overlooked

(3.) I've claimed that ambiguity—understood as higher-order uncertainty—has a sound basis but radical implications. That implies that others have missed it. How could that be?

An economist and a philosopher are walking down the street, when the philosopher excitedly points to the ground: 'There's a \$100 bill!'. The economist doesn't break his stride. 'Wait, don't you want the money?', asks the philosopher. Over his shoulder, the economist replies, 'It's not real. If it were, someone else would've picked it up by now.' So the philosopher picks it up.

The economist is the butt of the old joke, but it'd be less damning if we were to follow our pair for a full afternoon. Every few minutes, the philosopher would exclaim that some fantastical,

¹²E.g. Friedman 2017, 2019, 2020, 2024; Thorstad 2021; Falbo 2023a,b; Steglich-Petersen 2024; Pettigrew 2025.

unlikely object is right before their eyes. In 99% of the cases, the economist would be correct not to break his stride. The radical conclusions of philosophers are infamously unreliable.

I claim that ambiguity is in the 1%. Despite economists (and psychologists, and philosophers, and others) working on ambiguity for decades, they have missed both the most-conservative and well-grounded way to model it, and the radical implications of doing so for Bayesian convergence. Like the over-excited philosopher in the joke, I am claiming that money is hiding under our noses. And like the economist in the joke, it's reasonable for you to be skeptical.

As in the joke, the most reliable way to tell is to stop and look closely. But unlike the joke, that isn't easy. Despite my best efforts, this manuscript remains... let's say, intricate. You are right to ask for an explanation of how so many smart people could've missed a \$100 bill lying in plain sight, before you decide it's worth getting out your magnifying glass.

My answer is that it hasn't been in plain sight. Ambiguity as Higher-Order Uncertainty has been obscured by a confluence of confusion and dissension around the idea of 'higher-order probabilities' at its core. Ever since people started theorizing about probabilities, they've been theorizing about higher-order probabilities (e.g. Hume 2000 [1738], Part IV, §1). But—as anyone who's taught it can attest—the interpretation of even first-order probabilities has always been fraught, and it's easy for discussions to fall into confusion. The difficulties ramify for *higher-order* probabilities. Let me offer you a menu:

Actions reveal probabilities: Subjective probabilities are supposed to explain behavior (Ramsey 1931; Von Neumann and Morgenstern 1944). So couldn't a reflective person figure out what their probabilities are just by acting (or thinking)? No. Chapter 2 explains how *cognitive noise*—stochasticity between our underlying mental states and actions—prevents this.

Map, not territory: Your probabilities capture your uncertainty about the world; is it a mistake to reify them as a variable, P ? No. Chapter 4 will explain why even bog-standard Bayesian models treat your (future, which will later be your present) probabilities as a variable.

Collapsing expectations: If you're unsure about your own probabilities, can't you just take your *expectation* of them (Savage 1972, §4.2) and use that instead? Isn't that the 'tower law' of expectations? No. Chapter 5 will show why in models where your opinions are probabilistic (obeying the tower law) but higher-order uncertain, the 'collapsing expectations' argument fails. What it overlooks is the fact that when you don't know what your opinions are, learning what they are would provide new information—thereby *changing* them.

Infinite hierarchy: Higher-order probabilities involve an indefinite hierarchy. Does that make them intractably complex? No. Interpersonal uncertainties *also* generate indefinite hierarchies: I'm unsure how unsure you are about how unsure I am about... (etc.). Chapter 5 will use epistemic logic (Hintikka 1962) and variants of type spaces (Harsanyi 1967), to show that the solution is to use an outcome space with states fine-grained enough to settle what your opinions about those states are. There *is* an indefinitely hierarchy, but models of it can be finite and tractable.

Irrational behavior: Are agents with higher-order uncertainty inevitably subject to sure losses (Uchii 1973)? Not on the right notion of 'sure' loss. Chapter 6 will show that higher-order uncertainty can be constrained to satisfy iron-clad normative credentials, such as the value of information and the avoidance of sure losses.

Idle theory: Even if theoretically sound, how could higher-order probability have anything to do with real people? Because real people track how confident they are in their judgments under uncertainty. Chapters 3 and 7–10 show that higher-order uncertain Bayesian models can both retrodict and predict large swaths of how real people reason.

I don't claim to see \$100 bills often. I have done my homework.

Fair enough. But even if a Bayesian theory can indeed do all that, surely even I don't really believe it—right? Haven't we learned from decades of empirical research that people are systematically irrational, unable to do even the most basic Bayesian reasoning when asked? The lesson of behavioral economics was that we need to *de-idealize* our models of human beliefs and behavior. Don't we *know* that people's reasoning is not (even approximately) Bayesian?

1.7 Two Faces of Human Cognition

Pick up a modern book about the human mind. What image of humanity will you walk away with? It depends.

If the book is labeled 'psychology', you will likely walk away with an image recognizable from public discourse. The one-word summary? Disdain.

The author will bemoan the pervasive foibles of human cognition. You will be told that people are inept at reasoning under uncertainty, failing the most basic comprehension tests of statistical reasoning. People commit the *conjunction fallacy*, thinking that social-justice-oriented Linda is more likely to be (1) a feminist bank teller than (2) a bank teller, despite the fact that every possibility in which (1) is true is also one in which (2) is, but not vice versa (Tversky and Kahneman 1983). They commit the *base rate fallacy*, thinking that math-puzzle-loving Jack is more likely to be an engineer than a lawyer, ignoring the fact that he was picked at random from a group of mostly lawyers and few engineers (Kahneman and Tversky 1973). They don't even understand the simplest probability problems, committing the *gambler's fallacy* and thinking that after a streak of tails, a fair coin is more likely to land heads (Tversky and Kahneman 1971).

You'll be told that with foibles like these, it's no wonder that people so readily go off the rails. They exhibit *confirmation bias*, accepting without question information that fits with their beliefs, but being relentlessly critical of information that tells against them (Nickerson 1998; Fine 2005). They approach inquiry with the *motivated reasoning* of a lawyer rather than the impartial balancing of a scientist, arguing themselves into desired conclusions (Kunda 1990; Mercier and Sperber 2017). As a result they make bold, *overconfident* forecasts that are predictably miscalibrated: things experts are 100%-confident in turn out to be true around 80% of the time, while those they are 80%-confident in are true less than 60% of the time (Lichtenstein et al. 1982; Tetlock 2005). But despite being regularly surprised, they fail to learn from their mistakes due to *hindsight bias*, adjusting their memory of what they forecasted based on what happens. When they learn that something happened, they over-estimate how likely they thought it was; when they learn that it didn't happen, they under-estimate how likely they thought it was (Fischhoff 1975; Tetlock 2005).

You'll be told that with biases like these, it's no wonder human forecasters perform so poorly. Meehl (1954) famously compared the forecasts of simple linear models based on a few cues with those of human experts using a broad range of evidence, and found that the simple statistical models

often outperformed (and rarely underperformed) the experts. The result has been replicated many times across many domains—including clinical predictions, college admissions, parole decisions, and political and economic forecasts (e.g. Grove et al. 2000; Tetlock 2005). So it's not simply that the world is complex and prediction is difficult. They are—but even so, humans seem to be using their information sub-optimally. A disturbing corollary is that for some ('Hedgehog') styles of reasoning—those more inclined toward confirmation bias and hindsight bias—people actually make *worse* predictions in domains where they are experts than domains in which they are dilettantes (Tetlock 2005). So much for the Bayesian 'value of information' theorem (Blackwell 1951; Good 1967) that more information about a topic leads (on average) to better beliefs and decisions!

Finally, it will be suggested that these foibles explain the pathologies of human society. If people are such poor reasoners, it's no wonder that they vote against their interests (Achen and Bartels 2017), fall for demagogues (Levitsky and Ziblatt 2019), dismiss science (Oreskes and Conway 2011), polarize into political tribes (Mason and Wronski 2018; Klein 2020), and radicalize themselves into believing conspiracy theories (Ariely 2023). If only people could think straight, an untold number of societal problems would be solved. Daniel Kahneman put it bluntly: 'What would I eliminate if I had a magic wand? Overconfidence.' (Shariatmadari 2015).

Call this *mechanistic psychology*. The image you'll walk away with is that human cognition consists of a hodgepodge of heuristics that passed muster in our evolutionary past, but is ill-suited to our modern society.

But suppose, instead, the book you pick up is labeled 'cognitive science'. Curiously, you'll likely walk away with a very different image of the human mind. The one-word summary? Awe.

The author will marvel at the everyday feats of human cognition. They will approach the problem from the perspective of 'reverse-engineering': instead of evaluating human performance against ideal reasoners, they will set about the onerous task of trying to design machines that can learn, and think, and walk, and talk like we do. They will point out that once you try to design a robot that can do everyday tasks like unload the dishwasher or clean up your apartment, you realize that we are far, far too blasé about our own mental lives:

The reason there are no humanlike robots is not that the very idea of a mechanical mind is misguided. It is that the engineering problems that we humans solve as we see and walk and plan and make it through the day are far more challenging than landing on the moon or sequencing the human genome. (Pinker 1997, 4)

We open our eyes, and familiar articles present themselves; we will our limbs to move, and objects and bodies float into place; we awaken from a dream, and return to a comfortably predictable world... But think of what it takes for a hunk of matter to accomplish these improbable outcomes, and you begin to see through the illusion. Sight and action and common sense and violence and morality and love are no accident... each is a tour de force, wrought by a high level of targeted design. Hidden behind the panels of consciousness must lie fantastically complex machinery—optical analyzers, motion guidance systems, simulations of the world, databases on people and things, goal-schedulers, conflict-resolvers, and many others. (Pinker 1997, 18–19)

They will point to Moravec's paradox, that programming a machine to do the tasks that we think

are easy turns out to be maddeningly difficult. That is the reason that there are (*still*, in 2025) no humanlike robots: ‘it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility’ (Moravec 1988, 15). Indeed, despite decades of trying, [we still can’t reverse-engineer the mind](#) of microscopic, 302-neuron nematode *C. elegans* (Evans 2025).

They will point out that every time you open your eyes and immediately know what is where in front of you, you solve an ‘ill-posed problem’ (Marr 1982): a 2D projection onto your retina never uniquely determines the 3D scene that generated it. A white splotch could be a bit of paint on the window 3 feet away or a mile-wide cloud 3 miles away or a 1000-mile-wide moon 250,000 miles away. And that’s just the beginning of the problem (Brooke-Wilson 2023). No wonder modern computer vision still misclassifies people and dogs and cars and trucks at an alarming rate—as anyone who’s watched the internal model of a ‘full self-driving’ Tesla can attest.

They will point out that not only can you know what is where by looking, but you have a good idea of both where it will be in a moment, and what will happen if you prod or push or shake it in indefinitely-many different ways (Battaglia et al. 2013; Smith et al. 2024). In under 1 second, the trajectory of a milk-carton’s projection on your retina as it tilts off the table will instantly modify your reach angle and velocity, while your memory of how the milk sloshed when you picked it up two minutes ago will instantly adjust your grip strength. Performing that bleary-eyed feat before biking into work requires motion-sensors to immediately update an intuitive physics-engine, which in turn combines its information with medium-term memory to modify an action-plan subroutine. The author will point out, as you hop on your bike, that self-driving cars are hard enough—don’t get them started on what it would take to build a robot that can manipulate a bicycle deftly enough to get itself through the streets of Boston in rush hour.

Speaking of memory, they will point out that while you’re kicking yourself for not remembering a friend’s partner’s name, the problem of memory is more one of timely retrieval than of storage. Supposing you’re 30 years old, perfectly storing (just) your visual experience would take on the order of tens of exabytes of data—about a month of global internet traffic today. Searching a database like that is intractable, and our memory displays patterns—from recency effects to power laws of forgetting—of a system optimized to quickly recall information that’s likely to be useful while forgetting information that’s not (Anderson 1990).

They will point out that despite failing to recall your friend’s partner’s name, you’ll be instantly able to understand what he’s thinking, wanting, or wondering from a curve of his lip or a dart of his eye or shift in his posture—figuring out where his hand will be down to a fraction of a second to meet it in a handshake, whether he’s relaxed or rushed so that you know whether to make a joke or instead ask how he’s doing, and whether he’s happy to see you or awkwardly trying to remember *your* name. Forget the predictable behavior of milk cartons—we spend much of our waking hours deploying a ‘theory of mind’ and ‘inverse planning’ to figure out what other black boxes filled with a hundred billion neurons and untold numbers of unobservable beliefs and desires are about to do. The fact that we *do* so often figure it out—instantly and effortlessly—is an ongoing mystery (Baker et al. 2009; Jara-Ettinger et al. 2024).

And then there’s language. The author will point out that, due to a combinatorial explosion

of possibilities, almost every sentence that is longer than 10 words has never been spoken before—and yet if you say to a 5-year old, ‘As the blue dragon twisted its neck, it spotted Jane’, he will know exactly what you mean (Chomsky 1957). If you then point to a person carrying a shopping bag and a guinea pig and say, ‘That’s a guinea pig’, the 5-year old will (in ‘one shot’) learn that guinea pigs are four-legged furry animals that are probably rodents, and avoid misclassifying dogs or people or shopping bags or people-carrying-shopping-bags as guinea pigs—despite the fact that all of those possibilities are consistent with their experience so far (Quine 1960; Xu and Tenenbaum 2007; Chater et al. 2024b).

You will be asked to consider why it has taken decades of research, an internet’s-worth of information, and trillion-parameter models trained over years at the cost of hundreds of millions of dollars to create modern AI systems that can approximate human fluency with language and common sense. And to notice that such models *still* fail to learn and think like humans—that humans get much more from much less (Chater et al. 2024a). There’s a reason that we each walk around with a hundred billion neurons and hundreds of trillions of synapses packed between our ears. There is no theory of how we see, infer, learn, walk, and talk—of the everyday feats of human cognition—that does not require us to be deploying enormous amounts of information and computational power every waking minute.

And deploying it well. In stark contrast with the mechanistic, heuristics-driven approach to high-level cognition, resource-constrained optimality—sometimes called ‘resource-rational analysis’ (Lieder and Griffiths 2019; Icard 2023)—has long been the dominant approach in explaining low-level cognitive processes like vision (Geisler 2011), sensorimotor control (Todorov 2004), causal reasoning (Anderson 1990), and memory (Anderson 1990). In recent decades the program has had surprising successes in explaining aspects of high-level cognition as well.¹³ That’s why ‘Bayesian models of cognition’ are all the rage in cognitive science.

Call this *rational psychology*. The image you’ll walk away with is that human cognition consists of a set of finely-tuned strategies for sensibly navigating intractably-difficult problems (Marr 1982; Anderson 1990; Griffiths et al. 2024).

If you read both books on ‘psychology’ and ‘cognitive science’, you’ll be left with a strange double-image of who we are. Mechanistic psychology is right that people exhibit biases that put them at constant risk of going off the rails, but leaves it utterly mysterious how they perform the everyday feats of human cognition. Rational psychology is right that we are adept at flexibly deploying finely-tuned strategies for solving intractable problems, but makes it surprising how readily we go off the rails. Granted, resource-limitations and approximation-strategies likely explain some of our foibles (Lieder and Griffiths 2019; Icard 2023; Chater et al. 2024a). But no ‘approximation’ to Standard-Bayesian updating worthy of the name would so readily violate Reasonable Convergence in the face of plentiful evidence. Something is missing.

This double-image is not just a theoretical problem—it shows up in our personal lives. Mechanistic psychology tends to talk about *them*, while rational psychology tends to talk about *us*. This is no doubt because it’s more tempting to believe that *they* are overconfident, biased, and liable to

¹³E.g. Gopnik 1996, 2020; Chater and Oaksford 1999; Tenenbaum and Griffiths 2006; Griffiths et al. 2008; Tenenbaum et al. 2011; Bonawitz et al. 2014b,a; Vul et al. 2014; Ruggeri and Lombrozo 2015; Icard 2016, 2017, 2023; Bhui and Gershman 2018; Bhui et al. 2021; Gershman 2021; Hartley 2022.

follow their tribe off the rails.

But they are us. Any defensible theory of how the mind works must be one that we can reasonably believe about how *our own* mind works. So take a moment, and look inward. Ask yourself what your own predictably-polarized beliefs—the ones influenced by the biases of the mechanistic psychologists—might be.

When I do this, I see the patriotism of a boy raised in middle America, the progressivism of a teenager with liberal parents, the secularism of a young contrarian intellectual, and the bothsidesism of an academic who’s worked on the rationality of polarization for a decade. If the mechanists are right—as surely they are—that early choices of orientation in people’s search for evidence influences where they end up, then those choices have undoubtedly influenced me. Philosophers are divided over how to react to such ‘arbitrary influences’ on beliefs.¹⁴ But one thing is clear: if I should accept the mechanists’ claims about irrationality—thinking to myself, ‘If I had ideally responded to my evidence, I would be much less confident in my progressive politics’—then I would be under normative pressure to lower my confidence in my progressive politics. No one can reasonably be confident that ‘*q is true, but if I were ideally rational I would be much less confident of it*’.¹⁵

But how to avoid it? I *am* a patriot, and a progressive, and a nonbeliever, and a bothsidesist. And although reading decades of mechanistic psychology has influenced many parts of my thinking and cognitive habits, it hasn’t changed that.

I’m guessing that you’re like me. We all see ourselves, at some level, the way the rational psychologists do: as using sensible strategies to answer the questions we face. *We* are not (simply) pigheaded, or conformist, or dogmatic. And yet the mechanists’ psychology works: if we were to share our biography, they could make good guesses about our beliefs. When we’re reflective, the double-image of human cognition leaves us seeing our own beliefs in double. The evidence supports progressive politics. But it also suggests that I formed that belief through predictably-polarizing mechanisms. What am I to do?

This book is my attempt to square this circle. The goal is to resolve both the theoretical double-image of the feats and foibles of human cognition, and my personal double-image of my own predictably-polarized beliefs.

My story, in a nutshell, is this. The rational psychologists are right: people *are* approximating Bayesian solutions to the inferential and decision problems they face. Under conditions where they can achieve sufficient levels of *clarity* about what they expect about the relevant evidence, they approximate Standard Bayesians and so converge to the truth. That is what explains our feats. But this is compatible with—in fact, explains and predicts—our foibles. For when our opinions about the relevant evidence are sufficiently *ambiguous*, then even approximating sensible (Ambiguous-) Bayesian solutions will leave us exhibiting biases that put us at risk of going off the rails. Yes, the mechanists are right that we exhibit biases and are prone to error. But—since they misunderstand the effects of ambiguity—they are wrong about what that means about us. We are not pigheaded, or

¹⁴E.g. White 2005, 2010; Schoenfeld 2014, 2022; Srinivasan 2015b; Mogensen 2016; Ye 2021; Dorst 2023b.

¹⁵This is a weaker claim than the ‘anti-akrasia’ norms that are hotly debated in epistemology (E.g. Horowitz 2014). It appeals only to the ‘New Reflection’ (aka ‘Informed Reflection’) principle of Elga 2013 and Chapter 5—which is valid on Williamson’s (2014) models of the unmarked clock. The principle is that your probability in q must equal your estimate of the *ideal* (‘informed’) credence to have given your evidence: $P_a(q) = \mathbb{E}_{P_a}(\hat{P}(q))$. See Chs. 3 and 11.

dogmatic, or irrational. We are, on the whole, relentlessly reasonable—and polarized nonetheless.¹⁶

In short, the hypothesis of this book is 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.

1.8 The Role of Rationality

From a certain perspective, this view of rationality is disheartening. If ambiguity makes bias and polarization inevitable, then thinking properly offers far fewer guarantees of accuracy or agreement than we’ve hoped. Rationality buys us less than we thought.

But—hold on. How much was our theory of rationality really buying us? Our best theory said that people who are minimally rational will avoid bias, become well-calibrated, and converge to the truth. Great news!

For them. What about *us*? If we believe the standard theory of rationality, the inevitable conclusion is that people like us—or, at least, like *them*, over there on the other side—aren’t even minimally rational. An overly optimistic view of rationality leads to an overly pessimistic view of humanity.

It also fractures the study of human reasoning. If minimal rationality is out the window, all bets are off when it comes to explain people’s cognitive feats and foibles—theoreticians have little to offer empiricists. Those interested in *humans* are left probing a black box full of 100 billion neurons, facing a hopeless underdetermination problem with no solution in sight. Meanwhile, those interested in *reasoning* declare themselves uninterested in how the mind works, retreating to mathematized idealizations that are insulated from what their neighboring disciplines are up to.

That’s what we get *if* we build Reasonable Convergence into our theory of rationality. This book shows that we needn’t. Despite a history of skepticism, higher-order uncertainty offers a theory of rationality under ambiguity that explains when and why Reasonable Convergence fails. Part I makes the conceptual case: Chapter 2 argues that cognitive noise makes higher-order uncertainty inevitable for agents like us, while Chapter 3 explains the concepts and distinctions needed to think about it rigorously. Part II makes the theoretical case: Chapter 4 explains why Standard Bayesianism implicitly presupposes higher-order *certainty*; Chapter 5 shows how to coherently and tractably model higher-order *uncertainty*; and Chapter 6 shows how such models can be given iron-clad normative credentials.

Many technical and philosophical subtleties emerge. Part II is for the theoreticians—those interested in the foundations of rational ambiguity. It explains what has been worked out (by me and others) so far, but also points to many open questions: once we recognize the coherence and

¹⁶No book could rebut the entire mechanist case for irrationality. This book will make little attempt to explain structural biases that purport to show that people’s beliefs *at a time* don’t approximate probabilistic coherence, like framing effects, the conjunction fallacy, and the gambler’s fallacy—though Chapters 2 and 10 will dabble. That is not because I think rational psychology doesn’t have (more) plausible accounts of what’s going on (see e.g. McKenzie and Nelson 2003; Sher and McKenzie 2006; Dorst and Mandelkern 2022; Quillien et al. 2025; Griffiths et al. 2018; Xiang et al. 2025; Dorst 2025), but because I’m less sure whether ambiguity is a driving force.

prevalence of higher-order uncertainty, the foundations of Bayesianism turn out to be far less settled than you'd think.

But—as with bog-standard Bayesian models—you don't need a Ph.D. in the foundations of Ambiguous-Bayesian models in order to *use* them. Part I will give you everything you need to jump to Part III, which is for the empiricists. There I'll use Ambiguous-Bayesian models to explain why Reasonable Convergence fails. I'll propose new, unified theories of hindsight bias, confirmation bias, polarization, and overconfidence. I'll show that these theories explain large swaths of the trends in the existing empirical literature; make new, concrete, and correct experimental predictions; and *make sense of* what we're up to when we exhibit these biases—as we all do.

Part IV will then zoom out, reflecting on the role of rationality in philosophy, science, and politics. I'll argue that rationality *matters*, and that implicit (and incorrect) assumptions about it exacerbate our fraught social and political ties. Not only are epistemology and behavioral science tightly linked, but both have a role to play in shaping our politics.

That is the plan.

Here is the picture. Higher-order uncertainty—i.e. ambiguity—has been unduly neglected. Properly understood, it is inevitable, rational, and significant. As a result, the foundations and consequences of our theories of rationality are far less settled than we've thought. That means, in turn, that theoreticians and empiricists are in this together. Rationality buys us *more* than we thought.

1.9 Methods

I'm often asked: Is this a book for 'formal epistemologists'? It's not. Or, at least, it shouldn't be.

It's an attempt to do *empirical* epistemology—epistemology firmly placed within the interdisciplinary study of human reasoning. As anyone who practices it will tell you: behavioral science is hard. If we want to make progress, we can't afford to be close-minded—all tools are welcome, provided they prove their worth. This book will be unapologetically ecumenical in its methods.

Often that means using math. Not because it gives our work a veneer of rigor, but because—as the behavioral sciences have long since learned—it's often the only way to state theories that are sufficiently predictive to be testable.

At some points—for example, when explaining higher-order uncertainty—we'll need to be especially mathematical if we want to think straight. The history of higher-order probability is riddled with failure in part because it's easy to write down constraints that trivialize it—theories for which there are no models. In Part II especially, I'll make extensive use of *model-based epistemology*—an approach that uses the methods of modal logic and model theory to check the coherence and breadth of the distinctions we're making and the principles we're proposing.¹⁷ This is essential: trying to reason about higher-order probability without models is flying without a net.

At other points, analytical methods will give out—there'll be no neat proof that establishes a prediction or set-of-constraints that specifies the plausible range of outcomes. Rather than wheeling out ever-more-arcane mathematics to try to get an analytic solution, we'll change tack, taking a

¹⁷Others working in this tradition: Williamson 2000, 2008, 2014, 2017; Titelbaum 2013, 2020; Goodman 2013, 2016; Goodman and Salow 2023; Salow 2018; Carter 2019; Goldstein and Hawthorne 2022.

page out of recent social epistemology by using simulation- and agent-based methods.¹⁸ This will let us explore otherwise-intractable questions, and draw out the consequences of our theories in a less-exact but more-robust way.

When this goes well, it'll lead to testable predictions. Then we can't shy away from it. First, we need to see how well our theories fit with the empirical trends already identified by behavioral scientists. The replication crisis calls for caution (Ioannidis 2005; OSF 2015), but not unbridled skepticism—there's a lot that we *do* know about human reasoning. Remembering to never put too much weight on a single finding, it's important to ask what empirical effects our theories can (and can't) explain.

Second, we need to make *new* empirical findings. To do that, we'll need to use established experimental and statistical tools to run severe tests of our theories' predictions (Mayo 2018). This will again require using mathematics and methods that most philosophers (myself included) don't specialize in.¹⁹ But there really is no substitute—any empirical epistemology worthy of the name needs to make empirical contributions to our understanding of human reasoning.

That probably sounds like a lot. Maybe too much. My hope is that many people will be interested in the broad arguments of this book, even though few will want to work through all the details. I've thus done my best to make it modular so that different audiences can quickly get to what they're most interested in:

Empiricists: If you want to see how ambiguity bears on psychology—without doing much math—read Part I ('Concepts') and then skip to Part III ('Uses'). Most psychologists and many economists will prefer this route.

Humanists: If you want to see how ambiguity bears on our understanding of ourselves and others—without *any* math—read Chapter 2 and then skip to Part IV ('Upshots'). Most social theorists will prefer this route.

Theoreticians: If you want to understand the normative and mathematical foundations of Ambiguous Bayesianism, read Part I and then continue onto Part II ('Theory'). Most mathematicians and formal philosophers will prefer this route.

Throughout the book, some sections are skippable—read them only if you care about the relevant details. Sections marked with an 'ε' contain empirical and experimental details, while those marked with a '†' contain †heoretical and mathematical ones. I have done my best to make the book (and its supplementary materials—like Mathematica notebooks, experimental data, and R scripts) self-contained. All of the mathematics—and most of the simulation-, experimental-, and statistical-methods—will be explained as we go.

Still, new methods are always intimidating. As someone who started graduate school with no knowledge of mathematics, programming, or experiments—believe me, I know. But at a high level: research is an explore-exploit dilemma, and I've personally found that there are few ways to spend your time that have higher payoffs than learning new methods that might open up entirely

¹⁸Others working in this tradition: Bala and Goyal 1998; Zollman 2007, 2013, 2024; Mayo-Wilson and Zollman 2021; O'Connor and Weatherall 2018, 2019; Weatherall and O'Connor 2020; Weatherall et al. 2020; Rubin and O'Connor 2018; Mohseni et al. 2021; Wu 2023; Freeborn 2024; Huang 2024.

¹⁹But we can learn from the growing contingent of experimental philosophers, e.g. Knobe 2003, 2007; Knobe et al. 2012; Machery et al. 2004, 2017; Sytsma and Livengood 2015; Chituc et al. 2016; Henne et al. 2017; Cova et al. 2021.

different ways of doing research. And at a practical level: the rise of large-language models—which can serve as personal tutors and coders—is a boon to anyone learning or deploying new mathematical or computational methods. (How do you think I managed to code anything in R?)

No one should be expected to have mastered all of these techniques. (I certainly haven't.) But everyone who wants to study human reasoning should be willing to try them. This book is written in the belief that philosophy has more to gain than to fear by being ecumenical. That if we learn enough from our neighbors, we might be able to teach them, too. That epistemology can be rigorous while also being relevant. That it can be formal while also being applicable. And that it can be abstract and principled, while also teaching us about ourselves.

That is my hope, at least. Let's see if you agree.

1.10 Looking Ahead

Failures of convergence pervade our lives; we are constantly navigating polarized opinions. Likewise, ambiguities pervade our lives; we are constantly uncertain what we ourselves think. These two facts are linked. To explain why we polarize—and to know what to make of it, and to do about it—we need to know how uncertainty about what we think *affects* what we should think. We need a theory of rationality under ambiguity.

I'll offer one. I'll permit higher-order uncertainty, but otherwise be maximally conservative—using exact probabilities, bog-standard Bayesian conditioning, and traditional expected-values. The point is to show how *minimal* changes to our theory of rationality make for *radical* changes in our verdicts about rationality.

Or so I'll argue. But I don't claim to have gotten to the bottom of this. What I *do* claim is that this rabbit-hole goes deep: the possibility and prevalence of higher-order uncertainty opens up a rich set of empirical, computational, mathematical, and normative questions.

Would you like a shovel?