

Chapter 3

What You Need to Know

Abstract

This chapter explains what you need to know to assess my theory of bias. I'll introduce the distinction between *random variables* (which a Bayesian can be uncertain about) and *constants* (which they cannot). Standard-Bayesian models use a constant for your prior—hence they implicitly presuppose higher-order certainty. I'll show how to use a variable for your prior to permit higher-order *uncertainty*, and highlight why it matters. It allows (partitional) Bayesian conditioning to lead to confirmation bias (as failures of the 'Martingale' or 'Reflection' principle). And it forces a separation between your *subjective* probabilities and '*informed* probabilities'—those you'd have were your higher-order uncertainty removed. I'll then show how to simulate Ambiguous-Bayesian models to make empirical predictions: the distinction between subjective and informed probabilities predicts the (in)famous 'probability-weighting' curves of prospect theory.

This chapter will explain the concepts and notation you need in order to read the rest of the book. I anticipate three different audiences:

Empiricists: If you trust me on the mathematics and want to see how ambiguity explains our biases—without doing much math—read this chapter (maybe skipping §3.2), and then jump to Part III ('Uses'). Most psychologists and economists will prefer this route.

Humanists: If you trust me on the mathematics *and* psychology, and want to see how ambiguity changes our self- and political-understandings—without doing *any* math—skip this chapter and jump to Part IV ('Upshots'). Most social theorists will prefer this route.

Theoreticians: If you don't trust me on anything—so want to see the mathematical and normative foundations of ambiguity—read this chapter then continue on to Part II. Most formal philosophers and mathematicians will prefer this route.

Humanists: catch you later. Empiricists and theoreticians: buckle up.

I'll begin by explaining Standard-Bayesian models, including the important distinction between *random variables* (which a Bayesian can be uncertain about) and *constants* (which they cannot). Priors in Bayesian models are usually represented with constants, which is why they implicitly

presuppose clarity. This turns out to be a crucial part of why Bayesian updates satisfy the ‘Martingale’ property—a property essential to the convergence theorems (§3.1.1). Once we represent your priors as a variable that your priors themselves can be uncertain about, (partitional) Bayesian conditioning can lead to Martingale failures and thereby failures of Reasonable Convergence (§3.1.2). The optional §3.2 gives a simple formal model of how this can happen. I’ve streamlined things for readability, omitting most references, some qualifications, and many follow-ups. Part II gives a more rigorous exposition of (Standard- and Ambiguous-)Bayesian models.

§3.3 then shows how simulations of Ambiguous-Bayesian models can be used to make empirical predictions. Under ambiguity, subjective probabilities inevitably differ from *informed* subjective probabilities—the probabilities you’d have if your higher-order uncertainty were removed. I show that if we use ambiguity as a model of ‘cognitive uncertainty’ (Enke and Graeber 2023)—how confident people are in what their own opinions are—then attention to this distinction can explain *probability-weighting*: the apparent distortions in subjective probabilities made famous by prospect theory (Kahneman and Tversky 1979; Tversky and Kahneman 1992). This illustrates how ambiguity changes Bayesianism’s empirical predictions. §3.4 will conclude with a word on how to understand the statistical methods I’ll use when running experiments to test those predictions.

3.1 The Bayes Necessities

The base of any Bayesian model is an **outcome space**, usually referred to with W in this book but often with ‘ Ω ’ in others. This is a set of possible states of the world, or ‘worlds’. When I want to mention an arbitrary member of W , I’ll use lowercase Roman letters like w , u , or v . These aren’t maximally-specific metaphysically possible worlds, but rather states of the world that are specific enough to answer every question we want to model uncertainty about. To model your uncertainty about whether you (vs. your partner) did the laundry a month ago, we need an outcome space W that has some worlds in which you did, and others in which you didn’t.

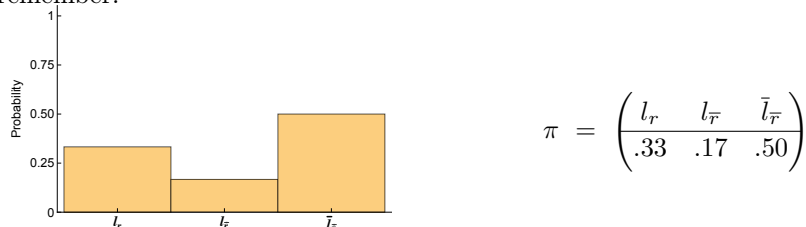
Sets of worlds then can be used to represent **propositions** (or ‘events’ or ‘claims’), denoted with other lowercase Roman letters like p , q , r , or others that are made clear by context. For instance, we can use l to represent the set of possibilities in W where you did the laundry. We can do logical operations on these—for instance, $\neg l$ is the set of possibilities (in W) where you *didn’t* do the laundry.

On top of an outcome space, a Bayesian model will lay down one or more **probability distributions**, generically denoted with lowercase Greek letters in this book, like π , δ , or η . Think of a probability distribution like an (uncertain) picture of the world that someone might have at any given time; a graded variant of a set of beliefs. It says both what possibilities the person leaves open—the worlds π assigns positive probability to—and also how likely those various possibilities are. Mathematically, it’s a function that assigns numbers between 0 and 1 to propositions (subsets of W), subject to some constraints. Since we will mostly focus on finite outcome spaces, we can think of a probability distribution π simply as an assignment of non-negative numbers to the different worlds in W , that together sum to 1.

On certain occasions, we’ll need to fully represent some probability functions over a (finite, small) outcome space. One way to do this is with plots like the one below, where we lay out the

3.1. THE BAYES NECESSITIES

possible worlds on the x -axis and plot a bar above it, with its height indicating how much probability goes to that world. A more-compact way to do this is with **vectors**, i.e. ordered lists of numbers. For example, suppose you are unsure both whether you did the laundry last month (whether l or $\neg l$), and whether your friend Thomas—who you sometimes call while doing chores—remembers you doing the laundry (r or $\neg r$). Let's suppose Thomas never falsely remembers, so there's no possibility in which r is true but l is false. Then we can use l_r to label the possibility where you did the laundry and he remembers, $l_{\bar{r}}$ to represent the one where you did it but he doesn't remember, and $\bar{l}_{\bar{r}}$ to represent where you didn't do it and he doesn't remember. (Sometimes I'll use overlines line \bar{l} for negations like $\neg l$.) Ordering them in that way, the following probability function might represent your beliefs about the combined question of whether you did the laundry and whether Thomas will remember:



The first row are just labels to remind us of the ordering of the worlds. The number below a world is the probability that π assigns to that world. For instance, since π assigns 0.50 to the possibility where you didn't do the laundry ($\bar{l}_{\bar{r}}$), and a total of $0.33 + 0.17 = 0.50$ to the possibilities where you did (l_r and $l_{\bar{r}}$), π is 50-50 on whether you did the laundry or not. Meanwhile, notice that *amongst* the possibilities where you did the laundry (l_r and $l_{\bar{r}}$), $\frac{2}{3}$ of the probability mass (0.33 of $0.33 + 0.17 = .50$) goes to the world where Thomas remembers (l_r). That means that π 's *conditional* probability for Thomas remembering, given that you did the laundry, is $\frac{2}{3}$. We can write that fact by saying that $\pi(r|l) = \frac{2}{3}$.¹

Suppose we want to represent your uncertainty about a quantity, like *the number of times you did the laundry in the last month*. Let's suppose you know that you did it 1 time in the last 3 weeks, so the only thing you're unsure about is whether you did it a month ago: if you did, you did the laundry 2 times in the last month; if you didn't you did it 1 time. Let's further suppose, unbeknownst to you, that you *did* do the laundry a month ago: l is true, so the number of times you did the laundry is 2. It follows that, in fact, '*the number of times you did the laundry*' and '*2*' pick out the same number. But we obviously can't represent your uncertainty about how many chores you did by saying you're unsure whether '*2 = 2*'—you're not uncertain about arithmetic.

What do we do? Probability theorists introduce what's called a **random variable**, often denoted \mathbf{X} , \mathbf{Y} , or \mathbf{Z} . This is often just called a 'variable', but don't confuse it for the sort of variables (vs. constants) that you learned about in Algebra 1. A (random) variable is a mathematical *function* that takes a state of the world, like w , and outputs a number. The number that X outputs when fed w as an input is written $X(w)$, or sometimes \mathbf{X}_w for compactness. In the case at hand, if we let X be a random variable corresponding to *the number of times you did the laundry last month*, then X takes value 2 at worlds in which l is true and 1 at worlds where l is false. Written

¹I'll write π by rounding to the nearest hundredth, but I mean π to be the exact distribution $(\frac{1}{3}, \frac{1}{6}, \frac{1}{2})$.

explicitly: $X_{l_r} = 2$ and $X_{l_{\bar{r}}} = 2$, while $X_{\bar{l}_{\bar{r}}} = 1$.

Whereas X is a variable—it picks out different numbers at different worlds—numerals like ‘1’ and ‘2’ are *constants* that always pick out the same number. Think of X like a *description* of a number, such as ‘the number of people in this room’, which can vary across states of the world. Think of ‘1’ like a *name* for a number. In probability-land, uncertainty about a quantity is always uncertainty about the value of a variable—never about the value of a constant.

Another way to put it is that when we treat bit of our language (like ‘1’) as a constant, and don’t let it vary across the worlds in the outcome space, then we are implicitly treating the agent modeled by the outcome space as *certain* of the value of the constant. After all: the different possibilities in the outcome space are supposed to capture all the distinctions we want to model uncertainty about; if something (like the number picked out by ‘1’) doesn’t vary across the outcome space, then no probability distribution over that outcome space will be uncertain about it.

Once we introduce a variable, we can use it to pick out sets of worlds (propositions), and see what probability π assigns to them. To flag that we’re using it this way, when unembedded I’ll use angle brackets: $\langle \mathbf{X} = \mathbf{2} \rangle$ is the set of worlds w at which $X_w = 2$. In our model, that is the set of worlds where you did the laundry a month ago, i.e. $\{l_r, l_{\bar{r}}\}$. As we know, π assigns $0.33 + 0.17 = 0.50$ -probability to that set of worlds, which implies that it assigns 0.50-probability to X having the value of 2. We can write that by saying that $\pi(X = 2) = 0.50$ (omitting angle brackets for readability).

I’ll sometimes want to talk about the proposition that X equals a generic value. This is standardly done by using *lowercase* letters like x , y , and z to be names for numbers. So, for example $\langle \mathbf{X} = \mathbf{x} \rangle$ is the set of worlds w where the random variable X outputs the number x , i.e. where $X_w = x$. Here think of x just like a generic stand-in for a numeral, like ‘1’ or ‘2’.

3.1.1 Standard Bayesianism

‘Standard Bayesianism’ is my term for the standard constraints imposed on prior probabilities and on how they are to be updated in response to new evidence. These constraints are imposed (often implicitly) in 99% of all applications of Bayesian models—including most every model used throughout the social and behavioral sciences.

Standard-Bayesian models represent your ‘prior’ probabilities with an outcome space W and a probability distribution π defined over that outcome space. New evidence comes in as learning the true answer to a question like ‘Does Thomas remember me doing the laundry?’ or ‘How many people RSVP’d for the party?’ or ‘How tall does that tree seem to be?’. Such a question is modeled with a **partition**, often denoted \mathcal{Q} in this book—a way of dividing the outcome space into mutually exclusive ‘cells’ that collectively cover the whole space. For example, the question ‘Does Thomas remember me doing the laundry?’ in our above model divides the outcome space into those where he does remember, $r = \{l_r\}$, and those where he doesn’t, $\neg r = \{l_{\bar{r}}, \bar{l}_{\bar{r}}\}$.

Standard-Bayesian updating consists in ‘conditioning’ the prior π on the true answer to the question \mathcal{Q} at each world. This involves going through each world and zeroing out the worlds inconsistent with the true answer to \mathcal{Q} at that world, and then ‘renormalizing’ (dividing by the remaining probability) so that updated probabilities still sum to 1. The details don’t matter. What *does* matter is that, from the perspective of your prior, your posterior is uncertain: if Thomas

remembers, you'll condition on r ; if he doesn't remember, you'll condition on $\neg r$. Since your prior isn't sure whether r is true, it isn't sure what your posterior will be—*your posterior probability that you did the laundry, after asking whether Thomas remembers* is an uncertain quantity.

Recall: in probability-land, uncertainty about a quantity is always uncertainty about the value of a variable. Thus *your posterior probability in l* is a random variable, $P^+(l)$. Likewise, *your posterior probability in r* is a random variable, $P^+(r)$. Instead of treating your posterior probabilities for each proposition separately, we can define them all at once by introducing a **variable probability function** P^+ to represent the entire posterior probability function you'll end up with in the various worlds. Like a random variable, this is a function from worlds; but unlike a random variable, it outputs not a number but a *probability distribution*. Think of P^+ like a *description*: 'The posterior probability distribution you'll end up with—whatever that turns out to be.' Whereas P^+ is the variable for your posterior, P_w^+ is the particular probability distribution you'll end up with at world w . (Think of P_w^+ like a name.) P_w^+ is the same sort of mathematical object as π , and so can be represented with a vector and we can write things like $P_w^+(l)$ to represent what your posterior probability that you did the laundry will be at w .

In the case at hand, if Thomas remembers (i.e. r is true), then your posterior will be $P^+ = \begin{pmatrix} l_r & l_{\bar{r}} & \bar{l}_{\bar{r}} \\ 1 & 0 & 0 \end{pmatrix}$, since you will then be certain both that Thomas remembers and that you did the laundry. Meanwhile, if Thomas doesn't remember (i.e. $\neg r$ is true), then your posterior will be $P^+ = \begin{pmatrix} l_r & l_{\bar{r}} & \bar{l}_{\bar{r}} \\ 0 & .25 & .75 \end{pmatrix}$, since you will then be certain that $\neg r$ and so lower your probability (in this case, to 0.25) that you did the laundry.

It'll be convenient to summarize all this information—saying at which world you wind up with which probability distribution—with a 'stochastic matrix'. This is a square matrix in which each row is a probability vector that states what your posterior is at a given world. In our case, adding labels, that matrix looks like this:

$$P^+ = \left(\begin{array}{c|ccc} & l_r & l_{\bar{r}} & \bar{l}_{\bar{r}} \\ \hline l_r & 1 & 0 & 0 \\ l_{\bar{r}} & 0 & 0.25 & 0.75 \\ \bar{l}_{\bar{r}} & 0 & 0.25 & 0.75 \end{array} \right)$$

This says that if you're at world l_r , your posterior will be $(1, 0, 0)$, the result of conditioning π on r . When we want to focus on the whole distribution, we can write it with a dot in place of a proposition: $\pi(\cdot|r) = (1, 0, 0)$. Thus we can describe the first row of the matrix by saying that if you're at world l_r , then $P^+ = \pi(\cdot|r)$, i.e. $P^+ = (1, 0, 0)$. We can capture the second two rows by saying that if you're at either of worlds $l_{\bar{r}}$ or $\bar{l}_{\bar{r}}$, your posterior will be $P^+ = \pi(\cdot|\neg r)$, i.e. $P^+ = (0, 0.25, 0.75)$.²

Since P^+ is a variable, we can use it (like X) to define propositions about your posterior as sets of worlds in the outcome space. For example, $\langle P^+(l) = 1 \rangle$ is the proposition that your posterior

²We make P^+ a function of the *world* w , rather than just the partition-cell learned, because your posterior also depends on what your prior was (and whether you updated properly). Notice that your posterior *at* a world w is not necessarily your posterior *conditional* on that world. For example, at $l_{\bar{r}}$, your posterior is $\pi(\cdot|\neg r) = (0, .25, .75)$, not $\pi(\cdot\{l_{\bar{r}}\}) = (0, 1, 0)$; the latter would be certain that you did the laundry, misrepresenting your opinions.

will be certain that you did the laundry, and is the set of worlds w where $P_w^+(l) = 1$, which turns out to be $\{l_r\}$. Since this is a subset of the outcome space, your prior π assigns it a probability—in this case, since $\pi(l_r) = 0.33$, that means that π assigns a 33%-probability to the hypothesis that you'll wind up certain that you did the laundry: $\pi(P^+(l) = 1) = 0.33$.

Notice that once we've represented your posterior as a *variable* probability function P^+ , it's odd to continue to represent your prior as a *constant* probability distribution π . After all, your posterior at one time will be your prior at a later time.³ Moreover, your posterior at a given world depends on your prior at that world, and if we are to represent *uncertainty* about your prior, we need to represent your prior as a variable. I'll generally use ' \mathbf{P} ' to denote the variable probability function capturing your prior (at a given time)—again, think of this as a *description*: 'Your prior distribution, whatever it might be'. Thus we can fully represent your prior and posterior at each possible world with two stochastic matrices:

$$P = \left(\begin{array}{c|ccc} & l_r & l_{\bar{r}} & \bar{l}_{\bar{r}} \\ \hline l_r & .33 & .17 & .50 \\ l_{\bar{r}} & .33 & .17 & .50 \\ \bar{l}_{\bar{r}} & .33 & .17 & .50 \end{array} \right) \quad P^+ = \left(\begin{array}{c|ccc} & l_r & l_{\bar{r}} & \bar{l}_{\bar{r}} \\ \hline l_r & 1 & 0 & 0 \\ l_{\bar{r}} & 0 & .25 & .75 \\ \bar{l}_{\bar{r}} & 0 & .25 & .75 \end{array} \right)$$

As things stand, the first matrix is highly redundant—it just asserts that you have the same prior (π) at each of the three worlds. But, as we'll see, it's important to make this explicit when it's true.

In this case, I've bolded $l_{\bar{r}}$ to indicate that it's (unbeknownst to P and P^+) the actual world—i.e. in fact you did the laundry (l is true), and in fact Thomas didn't remember it ($\neg r$ is true). I'll generically use \mathbf{a} as a name for (what we the theorists stipulate is, unbeknownst to P and P^+) the actual world. So in this model, $\mathbf{a} = l_{\bar{r}}$. For any model I write down, $\mathbf{P}_{\mathbf{a}}$ and $\mathbf{P}_{\mathbf{a}}^+$ will be constants specifying the prior and posterior probabilities you have at the actual world. In this model, for example, $\mathbf{P}_{\mathbf{a}}^+ = P_{l_{\bar{r}}}^+$, which equals $(0, 0.25, 0.75)$.

In short, to represent an (arbitrary) Bayesian update from a prior to posterior, what we need is a **probability frame** (W, \mathbf{a}, P, P^+) consisting of an outcome space W , an actual world $\mathbf{a} \in W$, and two variable probability functions P and P^+ describing what your priors and posteriors are at each world.

Standard Bayesianism is a constraint on probability frames, guaranteeing that updates look like the above example. Essentially, a frame (W, \mathbf{a}, P, P^+) is **Standard Bayesian** if and only if two conditions holds. First, P is constant across the possibilities you leave open; and second, P^+ is the result of conditioning P at each world on the true answer to some (partitional) question \mathcal{Q} .⁴

There are two features of Standard Bayesianism which it's important to highlight.

First, Standard-Bayesian updates are always clear. As defined in Chapter 1, your subjective probability function P is clear if it is certain of what your subjective probability function P

³Sometimes 'prior' gets used to mean 'ur-prior'—the prior distribution you started your life with, or some such. I always use 'prior' just to mean your earlier probability distribution (and 'posterior' to mean the later one) in a context where we're considering how your probabilities evolve over time.

⁴Where \mathcal{Q}_w is the cell of \mathcal{Q} that's true at w , that is to say: for all w , $P_w^+ = P_w(\cdot|\mathcal{Q}_w)$. There are slight generalizations discussed in Chapter 4—for instance, P doesn't have to be constant across the whole frame, but we have to be able to partition W into subsets such that (1) P is constant within each subset and (2) P is always certain of which subset it's in. Usually we include only the subset of the frame that contains the actual world.

is. That is: for any proposition q and number x , if $P_w(q) = x$, then $P_w(P(q) = x) = 1$. This is true for both your prior P and posterior P^+ , at every world, in every Standard-Bayesian model.

You can see this in your prior by noting that it doesn't vary, and recalling that in probability-land, uncertainty about a quantity (like $P(q)$) is always modeled as uncertainty about the value of a variable. If the values of $P(q)$ don't vary, there can't be any uncertainty about them.

You can see this in your posterior by noting that, although it *does* vary, your posterior probabilities perfectly track those variations. In worlds where r is true, $P^+ = (1, 0, 0)$; in worlds where it's false, $P^+ = (0, .25, .75)$. Notice that the former assigns probability 1 to the world (l_r) where r is true, and the latter assigns probability 1 to the worlds ($l_{\bar{r}}$ and $\bar{l}_{\bar{r}}$) where r is false. Either way, you are certain of what your posteriors are.

This generalizes. In Standard-Bayesian models you are always (1) certain of what you prior was (since it was clear), (2) certain of what you learned (since evidence was partitional), and (3) certain that you updated on what you learned by conditioning (since that happens at all worlds). It follows that your posterior is always certain of what your posterior is.

Second, Standard-Bayesian updates validate the 'Martingale' principle. This principle says that if you estimate that your posterior will end up at some value, your prior should match that value. Let's unpack this.

Often the value of a variable is uncertain—for example, P_a is unsure whether you did the laundry once or twice in the last month, i.e. whether $\langle X = 1 \rangle$ or $\langle X = 2 \rangle$. A probability function like π can form an **expectation** for the value of a variable X , denoted $\mathbf{E}_\pi(\mathbf{X})$ —when π is your actual probability function P_a , we can write this $\mathbf{E}_{P_a}(\mathbf{X})$. This is not necessarily what P_a 'expects' in the colloquial sense, as it might be a number that P_a knows X won't be. Rather, $\mathbb{E}_{P_a}(X)$ is a weighted average of the various possible values of X , with weights determined by how likely P_a thinks they are. Since P_a is 50-50 between X being 1 or 2, it's expected value is $\mathbb{E}_{P_a}(X) = 0.5(1) + 0.5(2) = 1.5$.

It turns out that expectations tells us many important things about a probability function—for example (by the 'law of large numbers'), what it is confident would happen, on average, if there were a bunch of independent copies of a variable. Since $\mathbb{E}_{P_a}(X) = 1.5$, that means P_a is confident that the average of a bunch of independent copies of X would be close to 1.5.

Expectations are usually introduced to describe how a Bayesian decides what to do. Suppose I offer you two different bets: B^1 pays out \$1 if you did the laundry last month, and pays out nothing otherwise. B^2 pays out \$2 if you did the laundry and Thomas remembers, and pays out nothing otherwise. Suppose you care about money 'linearly' in the sense that you care just as much about each dollar you might gain or lose. Then we can represent the '**utility**' (personal value) you'd get from these bets in various worlds with random variables: $B_{l_r}^1 = B_{l_{\bar{r}}}^1 = 1$ and $B_{\bar{l}_{\bar{r}}}^1 = 0$; while $B_{l_r}^2 = 2$ and $B_{\bar{l}_{\bar{r}}}^2 = 0$.

Which should you take? The standard Bayesian decision rule is to **maximize expected value (utility)**, which means taking the bet that has the highest expectation, according to your prior P_a . In this case $\mathbb{E}_{P_a}(B^1) = 0.5(1) + 0.5(0) = 0.50$, while $\mathbb{E}_{P_a}(B^2) = 0.33(2) + 0.66(0) = 0.66$. Since $0.50 < 0.66$, a Bayesian would take B^2 . One justification for this is the 'law of large numbers' we just saw: if the Bayesian were to face the same choice a bunch of times (for independent copies of B^1 and B^2 , defined on independent copies of propositions like l and r), they're confident that B^1 would net them around +\$0.50 per bet, while B^2 would net them around +\$0.66 per bet.

But expectations are important for many things, not just decisions. A prior distribution P_a has expectations for any random variable defined over W . Recall that when a Bayesian updates their beliefs, their prior is uncertain what their posterior probabilities will be. Yet, like any other variable, they can form an *expectation* for them. $\mathbb{E}_{P_a}(P^+(l))$ is your prior expectation for your posterior probability that you did the laundry. Although you think maybe your probability will go up to 1 (if Thomas remembers), and maybe it will go down to 0.25 (if he doesn't), it turns out that *your prior's expectation* for your posterior matches your prior: $\mathbb{E}_{P_a}(P^+(l)) = .33(1) + .67(.25) = .50 = P_a(l)$.⁵

That's no accident. This principle—sometimes called the 'Martingale' principle, and other times 'Reflection'—is a fundamental property of Standard-Bayesian updating:

Martingale: Your prior expectation for your posterior probability in q must equal your prior probability for q . (Formally, for any proposition q and world w : $\mathbb{E}_{P_w}(P^+(q)) = P_w(q)$.)

Martingale holds at every world in—it is **valid** on—every Standard-Bayesian frame.

It's easy to misunderstand Martingale. It doesn't say that you expect no change in your beliefs—only that, on average, you expect no directional push. We'll discuss it much more in Chapter 8. But it's easy to see that violations of it could exhibit a form of *confirmation bias*: if $\mathbb{E}_{P_a}(P^+(l)) > P_a(l)$, then it looks like you're investigating in a way that *predictably* (on average) will increase your probability that you did the laundry. This turns out to be the crucial feature of confirmation bias: as we'll see, Martingale is essential to Bayesian convergence results, and failures of it can be leveraged into cases where updating will predictably polarize you away from the truth.

3.1.2 Ambiguous Bayesianism

Turn to ambiguity. How can we model a prior probability function P that is uncertain of its own values? Easy. We've already made P a variable: a function from worlds to probability distributions. To model ambiguity, we just need to let it vary: allows P to take different values across the worlds that P_a leaves open. That is, we first let P_w and P_v differ on some claim—say, P_w assigns 0.33 to Thomas remembering ($P_w(r) = 0.33$), while P_v assigns 0.17 to it ($P_v(r) = 0.17$). Then we make sure that P_a assigns positive probability to both w and v : $P_a(w) > 0$ and $P_a(v) > 0$. We can use P to define propositions about probabilities the way we do for any other variable: $\langle P(r) = 0.33 \rangle$ is the set of worlds u in the outcome space at which $P_u(r) = 0.33$, and similarly for $\langle P(r) = 0.17 \rangle$. It then follows that $P_a(\langle P(r) = 0.33 \rangle) > 0$ and $P_a(\langle P(r) = 0.17 \rangle) > 0$. P_a has genuine higher-order uncertainty, i.e. 'ambiguity' in my sense.

P_a will itself assign some probability to r ; as the case may be, perhaps $P_a(r) = 0.33$. Then although you in fact assign 0.33-probability to Thomas remembering, you aren't sure of that: since you leave open the possibility that $\langle P(r) = 0.17 \rangle$, it follows that $P_a(P(r) \neq 0.33) > 0$ —another way to write $P_a(\neg\langle P(r) = 0.33 \rangle) > 0$ —and hence $P_a(P(r) = 0.33) < 1$. Your prior P is uncertain of its own values. That's all there is to it.

Of course, subtleties emerge. To ensure that your prior satisfies minimal rationality conditions, it needs to satisfy certain structural features. The most important one is what I call 'factorability': we

⁵Of course, $0.33 + 0.67(0.25) = 0.4975$; but I've been rounding. Using P_a exactly, $\mathbb{E}_{P_a}(P^+(l)) = \frac{1}{3}(1) + \frac{2}{3}(\frac{1}{4}) = \frac{1}{3} + \frac{2}{12} = \frac{2}{6} + \frac{1}{6} = \frac{1}{2}$.

need to be able to factor your overall uncertainty into (1) the ‘informed probabilities’ you’d have were your higher-order uncertainty about your own beliefs to be removed, and (2) your higher-order distribution over what those informed probabilities might be. Let’s unpack this.

To have higher-order uncertainty is to have a given probability function, but to be less-than-certain that you have that probability function. In other words, it requires a claim like $\langle P = \pi \rangle$ to be *true* at a given world w —your prior P in fact matches a given distribution, π —but at the same time for you to be less than certain that that claim is true. Mathematically: $\langle P = \pi \rangle$ is true at a world w (i.e. $P_w = \pi$), but at the same time $P_w(P = \pi) < 1$.

Whenever this happens, we can ask what your probability distribution would become if you were to *learn* that your prior was π —thereby removing your higher-order uncertainty. Call this your **informed probability** at w , and write it as \hat{P}_w .⁶ The difference between your (uninformed) probability P_w and your informed probability \hat{P}_w is crucial for understanding higher-order uncertainty. When your prior is clear at a world w , your informed probability \hat{P}_w equals your ‘uninformed’ probability P_w (i.e. $\hat{P}_w = P_w$)—since under clarity you’re already certain of what P is, so learning that $\langle P = \pi \rangle$ doesn’t provide new information. But whenever your prior is ambiguous at w , your informed probability differs from your uninformed one: $\hat{P}_w \neq P_w$, since \hat{P}_w is certain that $\langle P = \pi \rangle$, while P_w is not. Failing to realize this causes all sorts of confusion (see Ch. 5 and Elga 2013).

Let \hat{P} be a variable probability function capturing your informed probability, whatever it is. This varies across worlds the way that P does. But while your uninformed probabilities P may imperfectly track their own values across worlds (due to higher-order uncertainty), your informed probabilities are by definition certain of what your (uninformed and informed) probabilities are.

In the models used in this book, your uninformed probabilities P are your actual subjective probabilities—for instance, they might be the actual sampling propensities of your generative model. Meanwhile, the informed probabilities are implicit facts about you, which you are uncertain of—they represent the dispositions encoded in your generative model for how it would change if it were to learn more about what generative model you have.

Think of informed probabilities on analogy with the way that your other mental states might contain implicit information that, if revealed, would alter that mental state. Suppose you’re irritable, but are unsure whether this fact is caused by (1) Thomas being insensitive, or (2) you being hungry. Although you are currently irritable, learning about the cause of your irritability might change how irritable you are. If you learned that your irritability is due to Thomas’s insensitivity (that you’re not hungry), that’d make you *more* irritable at Thomas—he must be really acting inappropriately. But if you learned that it’s due to you being hungry (he’s just being his normal self), you’d get *less* irritable at Thomas. Suppose that, unbeknownst to you, your irritability is caused by your hunger. Then although you are in fact irritable at Thomas, if you were to learn more about (the causes of) your irritability, you’d cease being irritable at him. You’re irritable at him only because you don’t know everything there is to know about your irritability.

Likewise for your subjective probabilities P . You might in fact assign 33%-probability to Thomas remembering that you did the laundry. But that probability may be inflected by higher-order uncertainty: perhaps you leave open that you instead assign only 17%-probability to Thomas remembering, and leaving open this possibility ‘pulls down’ your actual probability (to 33%) from what

? change to an example where its about the mental state itself, not its causes?

⁶Mathematically: for any world w and distribution π , if $P_w = \pi$, then $\hat{P}_w := P_w(\cdot | P = \pi)$.

it'd otherwise be. Then if you were to *learn* that you assign 33% (rather than 17%) to Thomas remembering, that provides new information that could *change* your probability that Thomas will remember. Perhaps if it were no longer held down by higher-order doubts, your probability would rise to 50%. If so, then although your actual probability is $P_a(r) = 0.33$, your informed probability would be $\hat{P}_a(r) = 0.50$. This is exactly what happens in the model given in §3.2.

This is the most mind-bending part of higher-order uncertainty. Failing to notice the distinction—which ambiguity forces—between uninformed and informed probabilities has played no small part in the confusion and dissension around higher-order probabilities. If it hasn't clicked yet, don't worry. We'll get more of an intuitive sense for how informed probabilities relate to uninformed probabilities in what follows. See §3.2 for a concrete example and Chapter 5 for an extended discussion.

The core constraint on sensible higher-order uncertainty is that your prior P is **factorable**, meaning that your probability of a claim q always equals your expectation of your *informed* probability for q : $P_a(q) = \mathbb{E}_{P_a}(\hat{P}(q))$. This claim is formally analogous to the 'Martingale' principle discussed above, but instead of deferring to your future self, it involves deferring to a hypothetical, more-informed version of your current self. I hence call it the **Informed Reflection** principle, though it is sometimes called 'New Reflection' (Elga 2013; Stalnaker 2019; Dorst et al. 2021). If P meets this condition, then we can 'factor' it into (1) the various possible informed probabilities you might have, and (2) the various possible (uninformed) higher-order distributions *over* these informed probabilities you might have.

The details aren't crucial. What's important is that factorability both follows from minimal rationality conditions on your priors, and makes models of higher-order probability much more tractable to generate and analyze. Every model I write down in this book will be factorable. (Standard-Bayesian models are trivially factorable, since each world w is certain that $\langle P = \hat{P} \rangle$.)

Suppose, then, that we have a factorable—but perhaps ambiguous—prior P . How do we update it? The same way we update a Standard-Bayesian frame: by introducing a partition \mathcal{Q} representing the possible answer to a question, and having each world w update the prior P_w on the true answer to \mathcal{Q} at w .⁷ An **Ambiguous-Bayesian** update will be a probability frame (W, \mathbf{a}, P, P^+) in which (1) your prior P is factorable,⁸ and (2) there is a fixed question \mathcal{Q} such that at each world your posterior P^+ is the result of updating P on the true answer to \mathcal{Q} . Every model I use in this book will satisfy these constraints.

Mathematically, Ambiguous- and Standard-Bayesian updates look extremely similar. The crucial difference is that in Ambiguous- (but not Standard-)Bayesian updates, since your prior P is uncertain, *learning the true answer to \mathcal{Q} can provide evidence about what prior you had*. For a Bayesian, things are evidence for things that make them likely. So if \mathcal{Q}_w is more likely if your prior in q was high (if $\langle P(q) = 0.8 \rangle$) than if it was low (if $\langle P(q) = 0.2 \rangle$), then \mathcal{Q}_w provides evidence that your prior in q was high: $P_a(P(q) = 0.8 | \mathcal{Q}_w) > P_a(P(q) = 0.8)$. For a Standard Bayesian, since P was certain, this never happens: always, $P_a(P(q) = 0.8 | \mathcal{Q}_w) = P_a(P(q) = 0.8)$. This is what I meant when I said in Chapter 1 that Standard Bayesianism has a 'fixed and certain standard' against which to judge the force of the incoming evidence, whereas Ambiguous Bayesianism does not.

Although much of the literature on higher-order probability—including my own past work—

⁷So for all w , $P_w^+ = P_w(\cdot | \mathcal{Q}_w)$, where \mathcal{Q}_w is the true answer to \mathcal{Q} at w .

⁸And 'self-aware': P never rules out having the value it actually has, i.e. if $P_w = \pi$, then $P_w(P = \pi) > 0$.

? change earlier phrase to 'fixed and certain'?

centers on debates about whether updating is necessarily partitional (see Ch. 5), I now think that this is a red herring. Every update in this book will be partitional; and nevertheless, they will generate biases.

One reason to use partitional updates is that it makes it easier to see how you can reliably update your priors. Under partitional updating, you always are certain of what you learned. If you're unsure what your prior is, how can you reliably condition it on what you learned? In the same way that you can change your heart rate or blood pressure without being sure of what they are currently. If your heart rate is already at 100 beats per minute, doing a couple jumping jacks will raise it only moderately; but if it's at 60 bpm, doing so will raise it quite a bit. You can raise your heart rate (to differing amounts) depending on what it was by doing jumping jacks, even if you don't know what it was. Likewise: if you have a generative model but don't know its sampling propensities, then you can condition it on q by, for example, simply adding a line of code that re-runs the model if the sample it generates isn't one in which q is true. (See §2.3 and §5.3.1.)

So my models will be partitional updates of a factorable prior. Indeed, many of them will satisfy stronger constraints: as discussed in Chapter 6, we can impose further constraints on P to guarantee that both P and any partitional update of it (to P^+) satisfy versions of the 'value of rationality'. This says that conforming to P and acting on its basis is expected to lead to more accurate beliefs and more valuable decisions (Blackwell 1951; Good 1967). These are the 'iron-clad normative credentials' I've referred to.

Even under such assumptions, partitional updating under ambiguity will often lead to Bayesians to violate the 'Martingale' principle discussed above. This happens, for example, if we modify our above example to make your priors about Thomas's memory ambiguous. That is: suppose we make your prior unsure whether you are $\frac{2}{3}$ - or $\frac{1}{3}$ -confident that if you did the laundry, Thomas will remember, so $P_a(P(r|l) = \frac{2}{3}) > 0$ and $P_a(P(r|l) = \frac{1}{3}) > 0$. (See §3.2 for the full example.) Then learning that he doesn't remember will induce *hindsight bias*, lowering your estimate for how likely you thought he was to remember. This will in turn induce a Martingale failure: although your prior that you did the laundry is 0.5, your prior expectation for your posterior is 0.55. That is: $P_a(l) = 0.5$, but $\mathbb{E}_{P_a}(P^+(l)) = 0.55$, exhibiting a form of *confirmation bias*.

That's the ballgame. Martingale—which, recall, always holds under clarity—is fundamental to Bayesian convergence results. Any updating process that violates it can sometimes predictably fail to converge to the truth. We will spool out why at length, over Chapters 7–9. But here, in brief, is how it works.

Suppose there are a bunch of independent copies of the same question (*Did I do the laundry on week i , will Thomas remember it, and how confident I am that he will if I did?*) for bunch of different weeks. Suppose you do exactly analogous (independent) updates on Thomas's memories for each week. Since your prior in each hypothesis l_i (that you did the laundry on the i th week) is 0.5 and you treat them independently, you are initially confident that you did the laundry on roughly 50% of the weeks, and very confident that you did it on less than 52.5% of them.

Let's suppose your prior is well-calibrated: in fact, you did the laundry on exactly 50% of the weeks, and the questions are independent. Since (you're calibrated but) violate Martingale, your (accurate) estimate for the effect of doing all the updates is to shift you, on average, to a 55% posterior probability that you did the laundry on a given week. Thus you are very confident that

if you go through this process for each week, you'll end up estimating that you did the laundry on roughly 55% of the weeks. Since you'll continue to treat them independently, that in turn means you'll become very confident that you did it on *more* than 52.5% of them. This despite the fact that—as you initially were confident—you in fact did the laundry on 50% of the weeks. Under ambiguity, a series of Bayesian updates can predictably radicalize you.

How could this be? The trick is that under ambiguity—but never under clarity—the value of rationality fails to *agglomerate* across independent questions (§6.5.2). Although you expect each update to make you more accurate about the subject-matter it addresses—whether you did the laundry on *that* week—you expect doing *all* of the updates to collectively distort your beliefs about the overall proportion of times you did the laundry. That is: reasonable updates that improve local accuracy can lead, predictably, to going off the rails in your global estimates.

Meanwhile, suppose your partner goes through the inverse update, asking *their* friend Taylor whether she remembers *them* (rather than you) doing the laundry for each month, with the same starting uncertainty. By parallel reasoning, although their prior is 0.5 that you did the laundry on each week, their expectation for their posterior is *0.45*, and they will predictably become very confident that you did the laundry on less than 47.5% of the weeks. In other words: although you and your partner start in agreement, both have priors that are calibrated to the true rate at which you do the laundry and the reliability of the evidence you'll receive, the ambiguity in your priors will lead you to predictably polarize. If you each get evidence like this, you will become quite confident—and she will become quite doubtful—that you do the majority of the laundry.

This is the theoretical observation that forms the backbone of this book. Although it's an extremely stylized case, it turns out that the dynamics generalize. Under clarity, we should expect reasonable updating to lead to convergence. But under ambiguity, we shouldn't. That's why Reasonable Convergence is false.

You now have all the concepts and notation you need to understand the theories and arguments to come. If you want a better understanding of how the Ambiguous-Bayesian models work, read the next section for an in-depth example. If you've had enough of the math, skip to §3.3 to see how we can use such models to generate empirical predictions.

3.2 The Ambiguous Details[†]

In this section, I'll talk through the mathematical details of the Ambiguous-Bayesian model described informally in the last section.

Notice one way in which our Standard-Bayesian model of your laundry was unrealistic. In every possibility you left open, your prior assigned $\frac{2}{3}$ -probability to Thomas remembering, if you did the laundry: at all worlds w , $P_w(r|l) = \frac{2}{3}$. Thus we treated your prior as *clear* about this fact: you were (implicitly) certain that you thought Thomas was exactly $\frac{2}{3}$ -likely to remember, if you did the laundry: $P_a(P(r|l) = \frac{2}{3}) = 1$. But cases like this are paradigmatic cases of *ambiguity*, not clarity: if I ask you how likely you think it is that Thomas will remember, if you did the laundry, then any precise number you give me will feel arbitrary and noisy. How could we model you as uncertain what you think about how likely Thomas is to remember?

3.2. THE AMBIGUOUS DETAILS[†]

We need to distinguish possibilities where you're confident (*c*) versus doubtful (*d*) that Thomas will remember. This'll require splitting up our possibilities further, which can be done in many different ways. Recall your prior from before: $\pi = \begin{pmatrix} l_r & l_{\bar{r}} & \bar{l}_{\bar{r}} \\ .33 & .17 & .50 \end{pmatrix}$. Let's say that $\frac{2}{3}$ of the rightmost $\bar{l}_{\bar{r}}$ -possibility is one where you're confident ($\bar{l}_{\bar{r}}^c$), and $\frac{1}{3}$ is where you're doubtful ($\bar{l}_{\bar{r}}^d$). Thus the .50 probability gets divided between the two, yielding: $\begin{pmatrix} l_r & l_{\bar{r}} & \bar{l}_{\bar{r}}^c & \bar{l}_{\bar{r}}^d \\ .33 & .17 & .33 & 0.17 \end{pmatrix}$

We also need to say whether you're confident or doubtful at the other two possibilities. Let's assume that if you did the laundry, you being confident in Thomas's memory is correlated with him remembering. For simplicity—so we don't need to multiply worlds any more—let's make the correlation perfect. That is: that if you're at l_r , you're confident; and if you're at $l_{\bar{r}}$, you're doubtful. Let's mark this with *c* and *d* superscripts: $\begin{pmatrix} l_r^c & l_{\bar{r}}^d & \bar{l}_{\bar{r}}^c & \bar{l}_{\bar{r}}^d \\ .33 & .17 & .33 & 0.17 \end{pmatrix}$.

Finally, let's re-order the possibilities by whether you're confident or not: $\begin{pmatrix} l_r^c & \bar{l}_{\bar{r}}^c & l_{\bar{r}}^d & \bar{l}_{\bar{r}}^d \\ .33 & .33 & .17 & 0.17 \end{pmatrix}$. This prior assigns $\frac{2}{3}$ -probability to Thomas remembering, if you did the laundry, since $P_a(r|l) \approx \frac{0.33}{0.33+0.17} \approx \frac{2}{3}$. So let's call this prior π_c to indicate that it is confident in Thomas's memory.⁹

To have ambiguity, you need to leave open that you are *not* confident in his memory. We can create such a hypothetical prior by shifting probability toward the doubtful (*d*-)possibilities, and away from the confident ones; for example, with $\begin{pmatrix} l_r^c & \bar{l}_{\bar{r}}^c & l_{\bar{r}}^d & \bar{l}_{\bar{r}}^d \\ .17 & .17 & .33 & 0.33 \end{pmatrix}$. Call this prior π_d , since it is doubtful that Thomas will remember—it assigns $\frac{1}{3}$ probability to Thomas remembering, if you did the laundry, since $\pi_d(r|l) \approx \frac{0.17}{0.17+0.33} \approx \frac{1}{3}$.

To represent your uncertainty about your prior, we use a probability frame that tells us at which world you have which prior:

$$P = \begin{pmatrix} & l_r^c & \bar{l}_{\bar{r}}^c & l_{\bar{r}}^d & \bar{l}_{\bar{r}}^d \\ l_r^c & .33 & .33 & .17 & .17 \\ \bar{l}_{\bar{r}}^c & .33 & .33 & .17 & .17 \\ l_{\bar{r}}^d & .17 & .17 & .33 & .33 \\ \bar{l}_{\bar{r}}^d & .17 & .17 & .33 & .33 \end{pmatrix}$$

This says: at l_r^c and $\bar{l}_{\bar{r}}^c$, you have π_c and are confident in Thomas's memory, while at $l_{\bar{r}}^d$ and $\bar{l}_{\bar{r}}^d$, you have π_d and are doubtful in it. I've bolded the actual world to be $\mathbf{a} = \bar{l}_{\bar{r}}^c$, the possibility where you're confident he would remember you did the laundry if you did (*c*), but in fact you didn't (\bar{l}) so he doesn't remember (\bar{r}).

In this model, if you're confident (top two rows), you assign $0.33 + 0.33 = \frac{2}{3}$ -probability that you're confident, and $0.17 + 0.17 = \frac{1}{3}$ -probability that you're doubtful. Meanwhile, if you're doubtful (bottom two rows), you assign $0.33 + 0.33 = \frac{2}{3}$ -probability to being doubtful, and $0.17 + 0.17 = \frac{1}{3}$ -probability to being confident.

Thus you have genuine higher-order uncertainty. For example, the claim that you're 33%-

⁹Strictly, π_c is the distribution $(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6})$ but I'll round to $(0.33, 0.33, 0.17, 0.17)$. Similarly for π_d below.

3.2. THE AMBIGUOUS DETAILS[†]

confident that he'll remember, $\langle P(r) = 0.33 \rangle$, is true at only the first two worlds, $\{l_r^c, \bar{l}_r^c\}$, including the actual world. So although you actually assign 33%-probability to Thomas remembering, you're only $\frac{2}{3}$ -confident of that: $P_a(r) = 0.33$, but $P_a(P(r) = 0.33) = \frac{2}{3}$. Thus you're $\frac{1}{3}$ -confident that you *don't* assign 0.33 to r : $P_a(P(r) \neq 0.33) = \frac{1}{3}$, despite the fact that you do: $P_a(r) = 0.33$.

Likewise, you have higher-order uncertainty about your *conditional* probabilities. The claim that you are $\frac{2}{3}$ -confident that if you did the laundry Thomas will remember, $\langle P(r|l) = \frac{2}{3} \rangle$, is true at only the first two possibilities, $\{l_r^c, \bar{l}_r^c\}$. Meanwhile, the claim that you are $\frac{1}{3}$ -confident of this (i.e. doubtful), $\langle P(r|l) = \frac{1}{3} \rangle$, is true at the second two possibilities, $\{l_r^d, \bar{l}_r^d\}$. So although in fact you are confident, i.e. $P_a(r|l) = \frac{2}{3}$, you leave open that you are doubtful: $P_a(P(r|l) = \frac{1}{3}) = \frac{1}{3}$. Thus you have ambiguity about how likely you think Thomas is to remember if you did the laundry: you're unsure whether your (conditional) probability is $\frac{2}{3}$ or $\frac{1}{3}$, i.e. whether $\langle P(r|l) = \frac{2}{3} \rangle$ or $\langle P(r|l) = \frac{1}{3} \rangle$.

There are many questions to ask about the structure and stability of models like this—for detailed answers, see Chapter 5. Here I just want to do three things: (1) show how we can ‘factor’ such frames into first- and higher-order components, (2) explain how to think of sampling from them, and (3) in show how to update them.

(1) Factoring frames. There is a structure underlying sensible higher-order uncertainty. Since each possible probability function in the frame (π_c and π_d) is uncertain what P is, it follows that if they were to *learn* what P is, that would change their opinions. Recall that $\langle P = \pi_c \rangle$ is true at the first two worlds, $\{l_r^c, \bar{l}_r^c\}$, while $\langle P = \pi_d \rangle$ is true at the second two, $\{l_r^d, \bar{l}_r^d\}$.

To have higher-order uncertainty is to have a given probability function, but to be less-than-certain that you have that probability function. In other words, it requires a claim like $\langle P = \pi_c \rangle$ being *true* at a world without being certain at that world. This holds at the actual world, $\mathbf{a} = l_r^c$: $P_a = \pi_c$ (so $\langle P = \pi_c \rangle$ is true) and yet $P_a(P = \pi_c) = \frac{2}{3}$, less than 1.

What would happen if we took each world one by one, and *informed* it of what P is at that world, removing P 's higher-order uncertainty? This would involve learning the true answer to the question, ‘What is P ?’ at each world. In our case, that involves conditioning on the true cell of the partition, $\{\langle P = \pi_c \rangle, \langle P = \pi_d \rangle\}$, i.e. $\{\{l_r^c, \bar{l}_r^c\}, \{l_r^d, \bar{l}_r^d\}\}$. Doing that yields the following frame, representing the **informed probability function**, \hat{P} :

$$\hat{P} = \left(\begin{array}{c|cccc} & l_r^c & \bar{l}_r^c & l_r^d & \bar{l}_r^d \\ \hline l_r^c & .50 & .50 & 0 & 0 \\ \bar{l}_r^c & .50 & .50 & 0 & 0 \\ \hline l_r^d & 0 & 0 & .50 & .50 \\ \bar{l}_r^d & 0 & 0 & .50 & .50 \end{array} \right)$$

The informed probability function \hat{P} is by definition higher-order certain, i.e. clear. Let's call the two possible informed probability functions $\hat{\pi}_c = (.50, .50, 0, 0)$ and $\hat{\pi}_d = (0, 0, .50, .50)$, respectively. $\hat{\pi}_c$ is 50-50 between you doing and not-doing the laundry (l vs. \bar{l}), and also 50-50 between Thomas remembering or not (r vs. \bar{r}). It is also certain of c , that your *uninformed* probability was confident in Thomas's memory, i.e. that $\langle P(r|l) = \frac{2}{3} \rangle$ —since, recall that is the event $\{l_r^c, \bar{l}_r^c\}$. Meanwhile, $\hat{\pi}_d$ is 50-50 between you doing and not-doing the laundry, certain that Thomas won't remember

3.2. THE AMBIGUOUS DETAILS†

regardless, and certain that your uninformed probabilities were doubtful in Thomas’s memory, i.e. that $\langle P(r|l) = \frac{1}{3} \rangle$ —since, recall, that is the event $\{l_r^d, \bar{l}_r^d\}$.

Notice that in the uninformed frame, the possible probability functions π_c and π_d were each weighted averages of the possible *informed* probability functions, $\hat{\pi}_c$ and $\hat{\pi}_d$. In particular, π_c is a $\frac{2}{3}$ - $\frac{1}{3}$ average of the two. We can take averages of vectors by averaging their components:

$$\frac{2}{3}(.50, .50, 0, 0) + \frac{1}{3}(0, 0, .50, .50) \approx (.33, .33, 0, 0) + (0, 0, .17, .17) = (.33, .33, .17, .17)$$

And $(.33, .33, .17, .17) = \pi_c$. Similarly, π_d is a $\frac{1}{3}$ - $\frac{2}{3}$ -average of $\hat{\pi}_c$ and $\hat{\pi}_d$.

As mentioned above, this is no accident. All the Ambiguous-Bayesian models used in this book validate a principle I call ‘**Informed Reflection**’, sometimes called ‘New Reflection’ (Elga 2013). This principle ensures that the uninformed probability at a given world, P_w , is always a weighted average of the possible informed probabilities \hat{P} it leaves open. And not just any weighted average: P_w always matches P_w ’s *expectation* of \hat{P} :

Informed Reflection: Your probabilities should always match your expectation of your *informed* probabilities. (For any claim q and world w , $P_w(q) = \mathbb{E}_{P_w}(\hat{P}(q))$.)

Thus we can always ‘factor’ an ambiguous frame into its informed components and its higher-order distribution over those components. For example, we can represent the above frame P as having the informed distributions on the left, and the higher-order distributions *over* them on the right:

$$\left(\begin{array}{c|cccc} & l_r^c & \bar{l}_r^c & l_r^d & \bar{l}_r^d \\ \hline \hat{\pi}_c & .50 & .50 & 0 & 0 \\ \hat{\pi}_d & 0 & 0 & .50 & .50 \end{array} \right) \qquad \left(\begin{array}{c|cc} & P = \pi_c & P = \pi_d \\ \hline P = \pi_c & 2/3 & 1/3 \\ P = \pi_d & 1/3 & 2/3 \end{array} \right)$$

This says that the two possible informed distributions are $\hat{\pi}_c$ and $\hat{\pi}_d$, while the two possible uninformed distributions are a $\frac{2}{3}$ - $\frac{1}{3}$ -average of the two (π_c), and a $\frac{1}{3}$ - $\frac{2}{3}$ -average of the two (π_d). The right matrix captures the higher-order component, saying that if you have π_c , you’re $\frac{2}{3}$ -confident that you have π_c and $\frac{1}{3}$ -confident that you have π_d ; and if you have π_d , you’re $\frac{1}{3}$ -confident you have π_c and $\frac{2}{3}$ -confident that you have π_d .

(2) **Eliciting probabilities.** Recall the frame representing your uninformed probabilities:

$$P = \left(\begin{array}{c|cccc} & l_r^c & \bar{l}_r^c & l_r^d & \bar{l}_r^d \\ \hline l_r^c & .33 & .33 & .17 & .17 \\ \bar{l}_r^c & .33 & .33 & .17 & .17 \\ l_r^d & .17 & .17 & .33 & .33 \\ \bar{l}_r^d & .17 & .17 & .33 & .33 \end{array} \right)$$

If P represents your subjective probabilities at a given time, how do they influence your action?

Our primary interpretation of subjective probabilities is as the sampling propensities of your generative model (Chapter 2). This model says that your actual sampling propensities over the subject matter of ‘Did I do the laundry, will Thomas remember, and am I confident in his memory?’

3.2. THE AMBIGUOUS DETAILS†

is $P_{\mathbf{a}} = (.33, .33, .17, .17)$. That means that in any simulation-run of your model, you are 33%-likely to generate a sample where you did the laundry, are confident, and Thomas will remember (l_r^c), 33%-likely to generate one in which you didn't do the laundry, are confident, and Thomas won't remember (\bar{l}_r^c), 17%-likely to generate one in which you did the laundry, are doubtful, and Thomas won't remember ($l_{\bar{r}}^d$), and 17%-likely to generate one in which you didn't do the laundry, are doubtful, and Thomas won't remember ($\bar{l}_{\bar{r}}^d$).

Suppose you generate 10 samples to form your elicitation, ϵ —a probability vector that tracks how many samples of each possibility you drew. The contents will be noisy, but you might well draw $\epsilon = \left(\frac{l_r^c}{5/10} \quad \frac{\bar{l}_r^c}{3/10} \quad \frac{l_{\bar{r}}^d}{1/10} \quad \frac{\bar{l}_{\bar{r}}^d}{1/10} \right)$. How can you use this to determine your action?

You might maximize expected utility according to it. This is equivalent to ‘outcome-sampling’ as described in Chapter 2. For example, if you are asked whether you'd prefer a sure \$0 or to bet that you did the laundry at moderate odds—winning +\$3 if yes and losing −\$4 if no, then outcome-sampling using ϵ would warrant taking the bet, since $\epsilon(l) = \frac{5}{10} + \frac{1}{10} = 0.6$, and $0.6(-3) + 0.4(4) = +\$0.20$, greater than \$0.

Alternatively, you might use ϵ to form an expectation for what P is, and then maximize expected utility according to *that*. The expectation of a probability function is always itself a probability function, since it's a weighted average of different possible distributions. In this case, $\mathbb{E}_{\epsilon}(P) = 0.8(\pi_c) + 0.2(\pi_d)$, which equals $(0.3, 0.3, 0.2, 0.2)$. If you then maximize expected value according to *this* probability function—your elicited estimate of P —that's equivalent to ‘estimate-sampling’ as described in Chapter 2. In this case, you would decline the bet, since $\mathbb{E}_{\epsilon}(P(l)) = 0.3 + 0.2 = 0.5$. Notice that your prior is certain that your prior in l is 0.5— $P(l)$ is clear, since both $\pi_c(l) = 0.5$ and $\pi_d(l) = 0.5$ —so estimate-sampling will always assign a probability of 0.5 to l . (Averages of probability functions that all assign between x and y to q will always assign between x and y to q .)

Finally, you might *update* on your samples to learn more about what you think. That would involve conditioning on some fact about your elicitation—say, that there were 8 samples in which $\langle P = \pi_c \rangle$ and 2 in which $\langle P = \pi_d \rangle$, which would provide (inconclusive) evidence that you had π_c . Modeling this precisely involves expanding the outcome space W to include facts about what samples you draw—see Chapter 5 for how it can be done.

(3) Updating ambiguous priors. Suppose you start out with an ambiguous prior P , and are going to update it on whether or not Thomas remembers (r or \bar{r}) you doing the laundry. How does this work?

We update the probability function at each world by conditioning it on the true cell of the partition $\{r, \bar{r}\}$, i.e. $\{\{l_r^c\}, \{\bar{l}_r^c, l_{\bar{r}}^d, \bar{l}_{\bar{r}}^d\}\}$ just as we did for a Standard-Bayesian update. That yields the following:

$$P = \left(\begin{array}{c|cccc} & l_r^c & \bar{l}_r^c & l_{\bar{r}}^d & \bar{l}_{\bar{r}}^d \\ \hline l_r^c & .33 & .33 & .17 & .17 \\ \bar{l}_r^c & .33 & .33 & .17 & .17 \\ l_{\bar{r}}^d & .17 & .17 & .33 & .33 \\ \bar{l}_{\bar{r}}^d & .17 & .17 & .33 & .33 \end{array} \right) \quad P^+ = \left(\begin{array}{c|cccc} & l_r^c & \bar{l}_r^c & l_{\bar{r}}^d & \bar{l}_{\bar{r}}^d \\ \hline l_r^c & 1 & 0 & 0 & 0 \\ \bar{l}_r^c & 0 & .50 & .25 & .25 \\ l_{\bar{r}}^d & 0 & .20 & .40 & .40 \\ \bar{l}_{\bar{r}}^d & 0 & .20 & .40 & .40 \end{array} \right)$$

For example, $P_a^+ = (0, .50, .25, .25)$ since $\frac{.33}{.33+.17+.17}$ is .50, while $\frac{.17}{.33+.17+.17}$ is .25. As you can see, the effect of conditioning on $\neg r$ depends on what your prior was—if you were confident Thomas would remember (second row), it lowers your probability that you did the laundry to .25; but if you were doubtful that he’d remember (third and fourth rows), it lowers your probability only to .40.

At the same time, we can see that the update has induced a form of *hindsight bias*. Your initial estimate for your prior that Thomas would remember is $\mathbb{E}_{P_a}(P(r)) = \frac{2}{3}(0.33) + \frac{1}{3}(0.17) = 0.28$. But upon learning that he didn’t remember, your posterior estimate for *your prior* that Thomas would remember is $\mathbb{E}_{P_a^+}(P(s)) = 0.5(0.33) + 0.5(0.17) = 0.25$. Learning that he didn’t remember lowers your estimate for how likely your prior thought he was to remember (from 0.28 to 0.25). Chapter 7 will show that this dynamic can explain many of the empirical findings around hindsight bias.

Because of this, the update induces a Martingale failure. Your prior probability that you did the laundry is 50%: $P_a(l) = 0.5$. But your prior expectation for your posterior is $\mathbb{E}_{P_a}(P^+(l)) = 0.33(1) + 0.33(.25) + .17(.40) + .17(.40) = 0.55$. When your priors about Thomas’s memory are ambiguous, asking him whether he remembers you doing the laundry leads you, in expectation, to *raise* your probability that you did the laundry. Chapter 8 will show that this dynamic can explain many of the empirical findings around confirmation bias.

3.3 Ambiguity Predicts Probability-Weighting

I’ll now show how we can simulate Ambiguous-Bayesian models to generate empirical predictions, focusing on the patterns of behavior encoded by *prospect theory* (Kahneman and Tversky 1979; Tversky and Kahneman 1992). Doing so will highlight the importance of the difference between your probabilities (P) and your *informed* probabilities (\hat{P}) under ambiguity.

People (in)famously exhibit a four-fold pattern of risk sensitivity:

Unlikely gains: With low probabilities of gains, people are *risk seeking*—for example, buying lottery tickets despite this losing money on average.

Unlikely losses: With low probabilities of losses, people are *risk-averse*—for example, buying fire insurance despite this losing money on average.

Likely gains: With high probabilities of gains, people are *risk-averse*—preferring (i) a sure \$90 to (ii) a bet that’s 95%-likely to yield \$100 and 5% likely to yield \$0.

Likely losses: With high probabilities of losses, people are *risk-seeking*—preferring (i) a 95%-chance to lose \$100 and 5%-chance to lose nothing to (ii) a sure loss of \$90.

The classic explanation of these findings is *prospect theory*—a descriptive (non-normative) decision theory that hypothesizes that when people make decisions, they over-weight low probabilities above 0, under-weight high probabilities below 1, and are relatively insensitive to changes in probability in the middle of the scale. If we plot the objective probability on the x -axis and decision weight (or subjective probability, understood as what actually guides their actions) on the y axis, this shows up as an inverse-S curve—see the left of Figure 3.1. Philosophers will be more familiar with Buchak’s (2013) risk-sensitive decision theory, which builds on and normatively defends

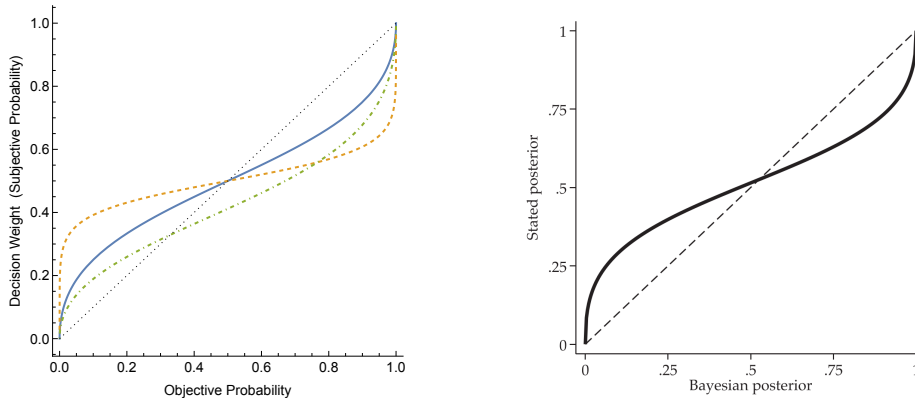


Figure 3.1: *Left:* Classic prospect-theory weighting functions that vary in both how insensitive people are to shifts in middling probabilities (blue vs. dashed-orange lines), and whether there isn’t or is average deflation (or inflation) in probabilities (blue vs. dot-dashed green lines). *Right:* Enke and Graeber 2023’s model fit of their empirical findings, with x -axis as the ideal (Standard-)Bayesian posterior and y -axis subjects’ stated posteriors.

Quiggin’s (1982) ‘rank-dependent utility’ (RDU) theory. These theories account for probability-weighting using a similar mechanism of decision weights.¹⁰

Focus on the solid blue curve. When objective probabilities are low, this is *above* the dotted diagonal, so objectively-low probabilities are subjectively inflated. This explains the first two patterns. If a gain is unlikely (like winning the lottery), people over-weight its probability and hope for the best—taking a chance on the lottery ticket. If a loss is unlikely (like a house burning down), people over-weight its probability and fear for the worst—being safe by buying insurance.

When objective probabilities are high, the blue curve is *below* the dotted diagonal, so objectively-high probabilities are subjectively deflated. This explains the second two patterns. If a gain is likely (like a 95%-chance to win \$100), people under-weight its probability and fear for the worst—being safe by taking the sure \$90. If a loss is likely (like a 95%-chance of losing \$100), people under-weight its probability and hope for the best—taking the chance to avoid any loss at all.

Notice a few things. First, there is high sensitivity at the end-points of the probability scale: the difference between a 1% chance and a 0% chance (or 99% and 100%) looms quite large.

Second, across different subject-matters these probability-weighting curves can vary in both (1) how insensitive to middling probabilities they are (blue vs. dashed-orange line), and (2) whether there is any overall deflation (or inflation) in probabilities (blue vs. dot-dashed green line). (1) controls how steep the curve is, while (2) controls its elevation and therefore the cross-over point where over-weighting of ‘low’ probabilities becomes under-weighting of ‘high’ probabilities. Empirical meta-analyses suggest (1) substantial insensitivity and (2) little if any *overall* inflation or deflation, across topics (Imai et al. 2025). But those overall rates hide heterogeneity: for some types of questions people exhibit average inflation, on others deflation.

Third, it’s common to treat decision-weights as an independent component of the decision pro-

¹⁰Prospect theory further introduces reference-dependence and differing treatments of gains and losses—see Wakker 2010. I omit these further complications, but it’d be worth investigating how (or whether) a theory like mine can account for the data that motivate them.

cess, assuming that people’s subjective probabilities match the objective probabilities (on average). But recent work suggests that people’s *subjective probabilities* are often genuinely distorted. For instance, the right of Figure 3.1 shows the empirical results from a belief-updating task in Enke and Graeber 2023. In these experiments, people got evidence about which of two bags (with differing proportions of blue vs. red marbles) had been selected. They reported their subjective probabilities in the form of an estimate about what the ‘optimal guess’ (Standard-Bayesian posterior) was, given their evidence. Even though these guesses were incentivized for accuracy, they exhibited the same inverse-S shape—a common finding in studies of the calibration (see Ch. 10).

3.3.1 Cognitive Uncertainty

Enke and Graeber 2023 show that we can explain probability-weighting in terms of ‘*cognitive uncertainty*’, defined as ‘people’s subjective uncertainty over their ex ante utility-maximizing decision’ (p. 2021). ‘Ex ante utility-maximizing decision’ is econ-speak for ‘what’s rational to do’. So people exhibit cognitive uncertainty when they have *self-doubt* in the sense from Chapter 1: uncertainty about whether the action they’ve chosen (or the estimate they’ve made) corresponds to the reasonable action (estimate), given their evidence and preferences.

Thus I’m going to argue that we can understand cognitive uncertainty in terms of *ambiguity* in my sense: cognitive uncertainty is *higher-order* uncertainty about what your own (and, therefore, the informed) subjective probabilities are. Doing so will illustrate the empirical methods I’ll use in the rest of this book to measure ambiguity and quantify its significance.

Enke and Graeber argue that cognitive uncertainty is not only explanatorily powerful, but easy to measure. Their methodology is to elicit estimates of uncertain quantities, and then ask subjects how certain they are in those estimates. For example, if someone states that \$8 is their true valuation (‘certainty equivalent’) for a lottery with uncertain outcomes, they are then asked, ‘How certain are you that you actually value this lottery somewhere between \$7.50 and \$8.50?’ (§III.3). In the probabilistic context, they introduce subjects to the notion of an ‘optimal guess, given their information’—understood as the posterior of a Standard Bayesian who started with a specified prior and conditioned on the evidence the subject had received about a bag of unknown composition. Then, if the subject reports a 70% subjective probability for the bag being mostly-red, they are asked, ‘How certain are you that the optimal guess is somewhere between 69.0% and 71.0%?’ The less certain subjects are, the more ‘cognitive uncertainty’ they have. Enke and Graeber go on to show the importance of cognitive uncertainty, measured in this way.

They are fairly neutral on the interpretation of cognitive uncertainty, but model it using noisy sampling.¹¹ Their model assumes that subjects are unsure what the objective probability $\mathcal{P}(q)$ of the given claim q is, but deliberation about it has the effect of eliciting samples from it. (*Outcome-samples*, in the terminology of Chapter 2—this will be important in a moment.) So if in fact $\mathcal{P}(q) = 0.8$, the subject can get a noisy signal about $\mathcal{P}(q)$ in the form of the n samples that each (unbeknownst to them) have an 80%-probability of being ones in which q is true. In other words, if we let S be a random variable for *the number of (n) samples in which q was true*, they update on the true answer to the question, $\mathcal{Q} =$ ‘What is S ?’ If $\mathcal{P}(q) = 0.8$, then S follows the objective

¹¹See Hartmann 2025 for a different sampling model of probability-weighting, which I also take inspiration from.

probability distribution $\text{Binomial}(n, 0.8)$ —which is likely to give values near $0.8n$. Varying n varies how much cognitive uncertainty subjects will end up with about $\mathcal{P}(q)$: learning that q was true in 8 of 10 samples provides weak evidence that $\mathcal{P}(q)$ is around 0.8; learning that it was true in 80 of 100 samples provides strong evidence that $\mathcal{P}(q)$ is around 0.8.

So they model deliberation (‘thinking’) as receiving signal S . What do people do with this signal? Enke and Graeber assume that they start with a ‘default’ prior distribution δ over the possible values of $\mathcal{P}(q)$ (a beta distribution—a standard distribution for modeling uncertainty about probabilities). The mean of this distribution is their prior expectation of $\mathcal{P}(q)$, i.e. $\mathbb{E}_\delta(\mathcal{P}(q))$. This number is also the subject’s prior probability for q before deliberating—we assume that subjects defer to \mathcal{P} in the sense that their expectation of $\mathcal{P}(q)$ sets their probability for q . (Or, perhaps better: sets their *elicitation* for how likely q is, in the terminology of Chapter 2. It’s at times unclear whether to understand δ or \mathcal{P} as the subject’s probability; but δ governs their reports.) They assume, reasonably enough, that this default expectation is intermediate: if $\mathcal{P}(q)$ is high, then $\mathbb{E}_\delta(\mathcal{P}(q)) < \mathcal{P}(q)$, while if $\mathcal{P}(q)$ is low, then $\mathbb{E}_\delta(\mathcal{P}(q)) > \mathcal{P}(q)$.

What does this predict? Suppose that the default prior’s mean is 0.5, that the objective probability is in fact $\mathcal{P}(q) = 0.8$, and that the subject draws 10 samples. Then it’s decently likely that q will be true in around 8 of the samples. If $\langle S = 8 \text{ of } 10 \rangle$ is true, learning that will shift their default distribution δ some *but not all* of the way toward thinking that $\mathcal{P}(q)$ is near 0.8—perhaps their estimate shifts to 0.7, i.e. $\mathbb{E}_\delta(\mathcal{P}(q) | S = 8 \text{ of } 10) = 0.7$. (Why not all of the way? Because that would be ignoring the prior δ . If you draw 10 samples and all 10 are ones in which q is true, you shouldn’t yet be certain that $\mathcal{P}(q) = 1$, and so should have an estimate below $\frac{10}{10}$. Similarly for less extreme cases.) Thus elicited responses will be regressive (flattened) toward the intermediate default prior, as we vary objective probabilities $\mathcal{P}(q)$. So long as subjects tend to get stronger signals if $\mathcal{P}(q)$ is close to the extremes of 0 or 1—either because n varies, or simply because Binomial distributions with extreme probabilities have less variance than those with intermediate probabilities—then their estimates will be *less* regressive as $\mathcal{P}(q)$ approaches extreme values. That’ll generate our inverse-S shaped probability weightings.

This is a brilliant proposal—especially the general hypothesis and empirical methods. But there are a couple reasons to worry about Enke and Graeber’s model of it. First, they assume that your expectation of $\mathcal{P}(q)$ equals your probability for q , i.e. that $\mathbb{E}_\delta(\mathcal{P}(q)) = \delta(q)$. Although standard, we’ll see later that this only makes sense if $\mathcal{P}(q)$ is *clear*, i.e. is itself higher-order certain. We’ve seen a glimmer of this already in the fact that under ambiguity, Martingale often fails: often your expectation for posterior probability of q doesn’t equal your current probability of q . And we’ll see in Chapters 5 and 8 that fully general theorems show that if a probability function \mathcal{P} is higher-order *uncertain* (ambiguous), then it’s impossible to always be such that your expectation of $\mathcal{P}(q)$ equals your probability for q .

Why would it be a problem to assume that \mathcal{P} is clear? Because then you should never *outcome*-sample from it, as Enke and Graeber’s model does. If \mathcal{P} is clear and $\mathcal{P}(q) = 0.8$, then $\mathcal{P}(\mathcal{P}(q) = 0.8) = 1$. If \mathcal{P} is certain of what its probability for q is in this way, why not just *ask* it what it’s probability for q is? As we saw in Chapter 2: under clarity, a single *estimate*-sample from \mathcal{P} —eliciting a sample w from \mathcal{P} , and checking what $\mathcal{P}(q)$ is in that sample, i.e. what $\mathcal{P}_w(q)$ is—will always perfectly reveal $\mathcal{P}(q)$. So if \mathcal{P} is clear and genuinely captures your implicit generative model

of the subject-matter, then reasonable elicitation of it will be *noiseless*. As argued in Chapter 2, a reasonable agent will have cognitive noise only if they have higher-order uncertainty. If they can sample from their generative models, their elicitation will be noisy only if those models are themselves higher-order uncertain.

Enke and Graeber may protest that this uses an overly-literal interpretation of their sampling model. Maybe so. But if so, it's worth asking what we'd have to do to make the model more realistic. If I'm right, we'd have to introduce higher-order uncertainty—and all the ramifications of doing so that are traced throughout this book, including failures of Reasonable Convergence.

I take this to be a friendly suggestion. For here's my second, more-important point: Enke and Graeber's model is a Standard-Bayesian one. A subject's elicited probabilities are modeled starting with a *constant* default prior distribution δ —not itself an object of uncertainty—which is updated on a partitional signal ('What is S ?'). Thus their elicitation can be modeled with a probability frame $(W, \mathfrak{a}, P, P^+)$ in which $P_w = \delta$ at all worlds w and $P_w^+(\cdot) = P_w(\cdot | S = \mathcal{Q}_w)$, where \mathcal{Q}_w is the true cell of the 'What is S ?' partition at w . If we like, we can explicitly add \mathcal{P} to the frame, so that we have $(W, \mathfrak{a}, P, P^+, \mathcal{P})$ —making explicit that what we've formalized is (mere) *probabilistic* uncertainty about more-informed probability functions (P^+ and \mathcal{P}), rather than genuine higher-order uncertainty: none of P , P^+ , or \mathcal{P} are uncertain about *their own* values. It thereby follows that this model satisfies Reasonable Convergence, as formalized in Bayesian convergence theorems. As with all Standard-Bayesian models, it will never exhibit confirmation bias in the form of Martingale failures (Ch. 8); and given enough evidence, it'll converge to the truth and agreement (Ch. 9) and become calibrated (Ch. 10). There are sharp limits on the biases it can explain.¹²

I'm offering an alternative. We can understand 'cognitive uncertainty' as genuine higher-order uncertainty, arising in cases where your subjective probabilities P are uncertain about *themselves*. Although conceptually nearby, this proposal is mathematically and empirically very different.

We know from Chapter 2 that higher-order uncertainty explains why reasonable elicitation would be noisy. And we saw above that it guarantees uncertainty about the *informed* probabilities, \hat{P} . These informed opinions are those you'd have if you had no ambiguity; P defers to them in the sense that (by factorability) P 's probability for q always matches its expectation of $\hat{P}(q)$. We can thus use \hat{P} as a stand-in for the objective probabilities $\mathcal{P}(q)$ —the probabilities our subject would have, if they had no ambiguity (no cognitive uncertainty)—and ask how variations in it correlate with variations in their subjective probabilities P . I'll show that this will lead to both probability-weighting, and the links Enke and Graeber demonstrate between it and cognitive uncertainty.

3.3.2 Cognitive Uncertainty as Higher-Order Uncertainty

Here's how the simulations will work. I will generate Ambiguous-Bayesian frames with random informed probabilities \hat{P} , random higher-order distributions (to generate a factorable P via their expectations of \hat{P}), and a random partition \mathcal{Q} to capture the subject's evidence and generate their posterior P^+ (by conditioning P on \mathcal{Q}). These models will contain a target proposition q which is (higher-order) uncertain. I will select a random world w , and record (1) q 's informed probability,

¹²At least, unless we distort the model—say, assuming that people return to their default prior after some time passes. Doing so would be epistemically irrational in obvious ways; if our explanation of biases makes essential use of it, we are arguably moving from rational psychology into mechanistic psychology (Chapter 1).

$\widehat{P}_w(q)$; (2) a measure of q 's ambiguity (cognitive uncertainty) at w ; and (3) q 's uninformed probability, $P(q)$. I won't fine-tune the parameters of the simulation, let alone do any model fitting. The point is to show that the qualitative relationships that Enke and Graeber find between objective probability, ambiguity, and subjective probability are to be expected for completely general reasons. See this chapter's Mathematica notebook for the simulation details.¹³

There are different reasonable ways to measure ambiguity so as to (roughly) match what Enke and Graeber are measuring. Here are three different measures of the *clarity* of your opinion in q at world w , C_w^q . Recall that P is your (potentially uncertain) true subjective probability distribution, while ϵ is your (constant, so known) noisy *elicitation* of that distribution (Ch. 2). Then, for some 'tolerance' $t \geq 0$:

Simple clarity: The true probability you assign to assigning the probability to q that you actually assign. If $P_w(q) = x$, then $C_w^q = P_w(x - t \leq P(q) \leq x + t)$.

Elicited clarity: The probability that your *elicitation* assigns to its probability being near your true probability in q . If $\epsilon(q) = x$, then $C_w^q = \epsilon(x - t \leq P(q) \leq x + t)$.

Elicited informed clarity: The probability that your elicitation assigns to its probability being near your *informed* probability in q . If $\epsilon(q) = x$, then $C_w^q = \epsilon(x - t \leq \widehat{P}(q) \leq x + t)$.

Whichever measure of clarity we use, our measure of *ambiguity* will be its inverse: $A_w^q := 1 - C_w^q$. Simple clarity is the simplest, and ignores the noise in people's actual elicitations—it tracks the distortions in underlying probability due to ambiguity alone. Elicited clarity is the closest to Enke and Graeber's measure of lottery valuation, since they elicit a valuation and then ask how certain you are that it matches your true valuation, determined by your actual subjective probabilities. Elicited *informed* clarity is the closest to Enke and Graeber's measure for probability estimation, where they elicit an estimate and then ask how confident you are that it's close to the 'optimal guess'. (Recall that your informed probability \widehat{P} is equivalent to the opinions of a Bayesian who started with clarity and observed the evidence you observed—Enke and Graeber's 'optimal guess'.)

As shown in the Mathematica notebook, all three measures generate qualitatively-similar results. I'll display those using **simple ambiguity** (i.e. 1-simple clarity). Since it ignores noise, this highlights that my explanation of probability-weighting works through a different mechanism than Enke and Graeber's, as we'll see.

Begin with the overall probability-weighting curves, displayed in Figure 3.2. These curves are generated by generating 50,000 frames and randomly selecting 5 worlds from each, for 250,000 datapoints. I've binned datapoints by rounding informed-probabilities $\widehat{P}(q)$ to the nearest tenth (0, 0.1, 0.2, ...), and then plotted the median subjective probability $P(q)$ in each bin.¹⁴ As you can see, a jagged version of the inverse-S shape emerges: subjective probabilities are regressive toward a default point in the middle of the scale, but jump to extremes at the endpoints. Why?

In short: the middle of the (informed) probability scale has substantial ambiguity, making uninformed probabilities regressive; but the endpoints tend to induce clarity, aligning uninformed with informed probabilities.

¹³Link: [TODO]

¹⁴Technically, I'm tracking the *posteriors* in the frame, \widehat{P}^+ and P^+ ; but since these can be seen as priors for a future update, I omit the '+' superscript.

3.3. AMBIGUITY PREDICTS PROBABILITY-WEIGHTING

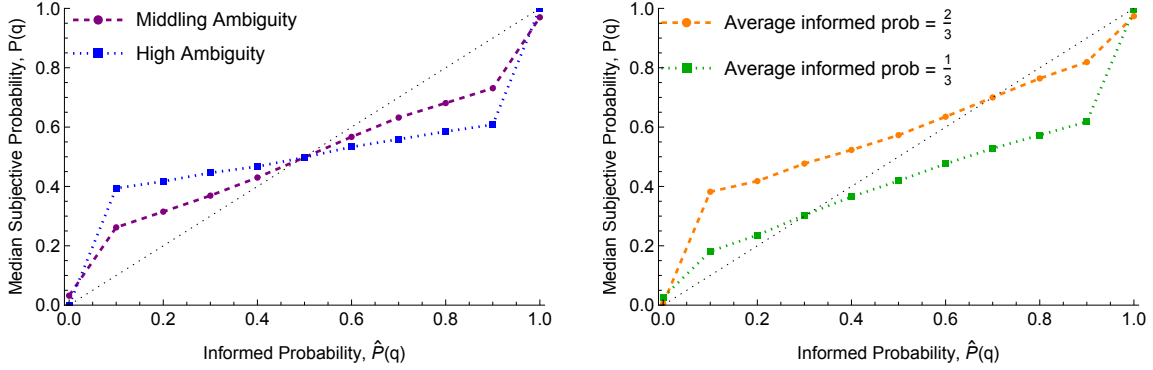


Figure 3.2: *Ambiguity generates probability-weighting curves.* Informed probabilities (x -axis) are plotted against uninformed probabilities (y -axis) in random (factorable) ambiguous frames. **Left:** the degree of ambiguity controls how regressive P is. **Right:** the average informed probability controls which default-point P is regressive toward.

Focus on the purple (‘Middling ambiguity’) dashed line on the left plot. Why is it regressive toward 0.5? For example, when the informed probability is 0.8 the median uninformed probability is 0.68. Why? In this simulation informed probabilities are generated uniformly at random between 0–1, meaning that *on average* they are 0.5. Recall that uninformed probabilities are expectations (weighted averages) of informed probabilities: $P_w(q) = \mathbb{E}_{P_w}(\hat{P}(q))$. Since in fact $\hat{P}_w(q)$ is 0.8, the probability that P_w assigns to having the the actual value it does (its simple clarity) goes to the hypothesis that $\langle \hat{P}(q) = 0.8 \rangle$. This pulls P_w some of the way toward 0.8, but the probability it assigns to the *other* possible informed probabilities distorts it. Why does that distortion pull it *toward* 0.5, on average? Because the *other* informed probabilities in the frame are on average 0.5, since they were pulled uniformly from $[0, 1]$. So when the informed probability is 0.8, although in any given situation $P(q)$ might be pulled up or down from 0.8, *on median* (and average), it gets pulled down, toward 0.5.

That’s why the median subjective probability is regressive in the middle of the probability scale. Why is it *sensitive* at the endpoints, jumping to near 0 and 1 when the informed probability reaches near those points? Because extremely strong evidence about q tends to also be *clear* evidence, leading P to track \hat{P} more closely. The most common way to generate an extreme (0 or 1) informed probability for q is to learn something (an answer to the random question \mathcal{Q}) that *settles* whether q is true or not, i.e. either q is true throughout \mathcal{Q}_w or $\neg q$ is true throughout \mathcal{Q}_w . Entailments like that are always clear: if you learn something that implies that q is true, then although your posterior may still have ambiguity about other claims, it won’t have any ambiguity about your posterior in q . Since this is the most common way to wind up with an extreme informed probability for q , ambiguity falls sharply at those extreme points, making P track \hat{P} more closely.

We can explore these dynamics by seeing how changing the simulations changes the result. If we increase ambiguity (by introducing more possible informed-probabilities, keeping higher-order distributions random), subjective probabilities become *more regressive*, since $P_w(q)$ is less pulled toward $\hat{P}_w(q)$ —see the blue dotted curve in the left plot. Meanwhile, if we increase or decrease the average informed probability (instead of generating them with a mean of 0.5), that will change the elevation of the curve and thereby the cross-over point between over-weighting and under-weighting.

This is displayed in the right plot: the orange dashed curve generates informed probabilities from a distribution with an average of $\frac{2}{3}$ (a Beta(2,1) distribution), while the green dotted curve generates them from a distribution with an average of $\frac{1}{3}$ (a Beta(1,2)).

Notice that my model generates these curves for different reasons than Enke and Graeber’s. Noise is essential to both models, but in different ways. In their model, noise in the signal from the objective probability \mathcal{P} is what leads subjective probabilities to be regressive, and the fact that signals are less noisy when $\mathcal{P}(q)$ is extreme is what leads the end-points to snap back toward the diagonal. That’s not what’s happening in my model. Noise is essential in the sense that we saw in Chapter 2: it’s only in the presence of cognitive noise that higher-order uncertainty is stable—if you could noiselessly sample from your subjective probabilities, reasonable people would remove their ambiguity before acting. But although noise is a prerequisite for my model, it’s not what’s driving the above curves. Those curves plot your *true* subjective probabilities ($P_w(q)$), not their noisy elicitation ($\epsilon(q)$). The inverse-S curve emerges because ambiguity distorts your true subjective probabilities away from what they would be with clarity (away from \hat{P}).¹⁵

So my model can recreate the basic probability-weighting finding of Enke and Graeber’s model. But there are many explanations of probability-weighting—the reason theirs is promising is that they find empirical evidence that probability-weighting distortions are robustly associated with cognitive uncertainty. That suggests that cognitive uncertainty is *driving* the effect. In my model, ambiguity is driving the effect. Does it do so in a way that predicts the associations they find?

Yes. The first effect Enke and Graeber find is that their measures of cognitive uncertainty predict how much subjective probabilities will be distorted from objective probabilities. They find this for valuations of risky bets, belief-updating, and stock-market expectations; I’ll just show belief-updating. The top left of Figure 3.3 shows median elicited posteriors as a function of the true Bayesian posterior, separated by cognitive uncertainty. Red Xs are estimates with high (above median) cognitive uncertainty; blue dots are those with low (below median). Red Xs are more regressive, while blue dots better follow the diagonal line. This effect is not just driven by extreme cognitive uncertainty. The top middle divides judgments into quartiles based on their cognitive uncertainty (0–25th percentile, 26th–50th, etc.), regresses elicited probability against Bayesian posterior, and plots the regression coefficients. The higher it is, the more a change in objective probability leads to a change in elicited probability. Notice that the coefficient decreases monotonically with cognitive uncertainty—more cognitive uncertainty leads to less sensitivity to the Bayesian posterior. Finally, the top right plots cognitive uncertainty against the absolute distance between elicited and Bayesian posteriors, showing that cognitive uncertainty leads to larger distortions.

The bottom of Figure 3.3 performs parallel analyses on the data from my above simulations. The bottom left shows that when we split the data by whether ambiguity is below (blue, dashed) or above (red, dotted) the median, the latter is much more regressive than the former. The bottom middle

¹⁵Here’s another way to make the point. It’s essential for Enke and Graeber’s model—but not for mine—that cognitive noise generates a *statistically-biased* indicator of your underlying generative model. X is a statistically-biased indicator of a probability $\mathcal{P}(q)$ if, when $\mathcal{P}(q)$ is a given value like 0.8, the average value of X will not be 0.8. (Formally, if, for some x , $\mathbb{E}_\pi(X|\mathcal{P}(q) = x) \neq x$, where π is some objective probability distribution.) Since you combine the (unbiased) signal S from \mathcal{P} with the default prior δ , the elicited estimate of $\mathcal{P}(q)$ is regressive toward the default prior, making it a statistically-biased indicator of \mathcal{P} . In contrast: on my model, even if your elicited probabilities are a statistically-unbiased indicator of your true probabilities P (as they will be if you outcome-sample), probability-weighting still emerges relative to the informed probabilities \hat{P} .

3.3. AMBIGUITY PREDICTS PROBABILITY-WEIGHTING

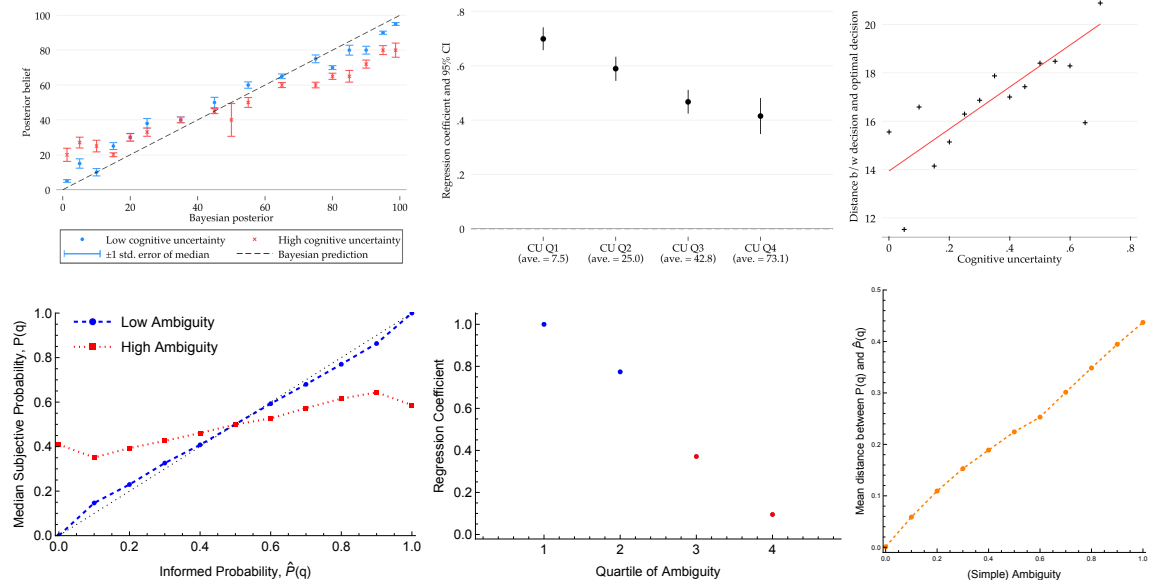


Figure 3.3: *Cognitive uncertainty predicts diminished sensitivity to objective probabilities.* **Top:** Enke and Graeber (2023)’s empirical findings. The left plots median reported posterior against true Bayesian posterior; highly cognitively-uncertain (red) judgments are more regressive. The middle divides judgments into quartiles based on their cognitive uncertainty, and regresses elicited probability against Bayesian posterior: higher uncertainty leads to less sensitivity. The right shows that as cognitive uncertainty goes up, the absolute distance between elicited and Bayesian posterior rises. **Bottom:** Analogous plots generated by parallel analyses of my simulations.

3.3. AMBIGUITY PREDICTS PROBABILITY-WEIGHTING

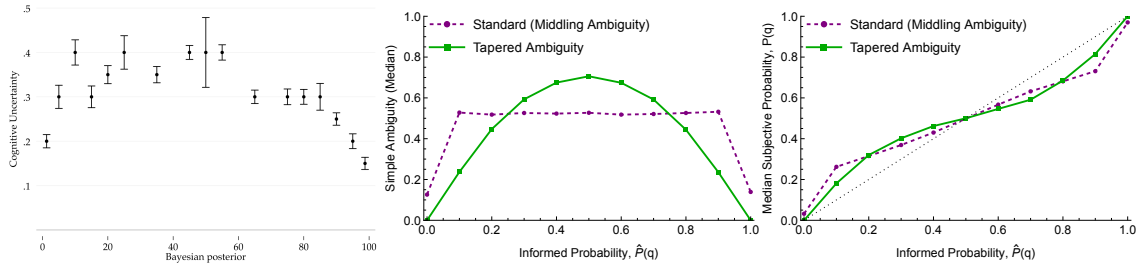


Figure 3.4: *Cognitive uncertainty is lower at extremes.* **Left:** Enke and Graeber’s data for belief-updating, showing that median cognitive uncertainty has an inverse-U shape when plotted against the Bayesian posterior. **Middle:** In my simulations, simple ambiguity is a (jagged or tapered) inverse-U shape in informed probability. **Right:** If ambiguity tapers off smoothly, this smooths out the inverse-S probability weighting curve.

shows that when we split the data into quartiles of ambiguity and regress subjective probability against informed probability, the coefficient decreases monotonically with ambiguity. The bottom right bins judgments by (simple) ambiguity and plots the mean absolute distance between subjective and informed probabilities, showing that ambiguity leads to larger distortions. None of these results have been fine-tuned to Enke and Graeber’s data; but the qualitative trends are robust.

If cognitive uncertainty is what drives divergences from the objective probability, why does that result in an inverse-S shape in objective probability—with low sensitivity in the middle of the scale, but high sensitivity at the endpoints? It turns out that in Enke and Graeber’s data, cognitive uncertainty takes an inverse-U shape in objective probability, as displayed on the left of Figure 3.4: cognitive uncertainty is highest when objective probabilities are middling, and drops off when they are extreme. This makes intuitive sense: it’s difficult to know exactly how confident to be that the bag is 60%- (vs. 40%-)-blue if you’ve seen 5 blue and 3 red draws (70%? Or 80%?); but easy to know when you’ve seen 30 blue and 20 red (around 99%).

Similar results hold in my simulations. The middle of Figure 3.4 plots median ambiguity in my simulations as a function of informed probability. The dashed purple line shows the results from the simulation displayed in the left of Figure 3.2. As you can see, there is a jagged inverse-U shape: median ambiguity is flat in the middle of the scale, but drops sharply as informed probability becomes extreme. This jagged drop-off is what produces the jagged versions of the probability-weighting curves we’ve already seen—reproduced on the right of Figure 3.4 in dashed purple. These curves are jagged because in the (simple) simulations I ran, ambiguity is only correlated with informed probability at the endpoints, when the evidence settles whether q . If we run simulations where ambiguity tapers off as informed probabilities get more extreme (solid green line in the middle plot), the probability-weighting curve smooths out to take on a familiar shape (solid green line in the right plot).¹⁶

In sum: ambiguity—understood as higher-order uncertainty—can account for both classic probability weighting curves, as well as the details of a promising explanation of them in terms of ‘cognitive uncertainty’. I don’t claim that my theory is superior to Enke and Graeber’s in its

¹⁶There are many ways to taper ambiguity in this way. I’ve implemented a simple, brute-force one in which more-extreme informed probabilities are associated with higher levels of simple clarity. It’s worth exploring more-nuanced ways to generate such ambiguity tapering.

explanation of probability-weighting (though it is interestingly different). We can use the simulations to make new predictions—for instance, increasing ambiguity should make the curve more regressive. Some of these predictions may differ from Enke and Graeber’s—for instance, my theory predicts that if we lower the average informed probabilities that people leave open, that’ll lower the cross-over point from over-weighting to under-weighting. We can also make predictions that *connect* my theory’s predictions across biases. For instance, Chapter 7 will show that (in theory, and empirically) hindsight bias increases with ambiguity; so a straightforward prediction is that when people’s probability judgments are more (less) regressive, they’ll be more (less) inclined to exhibit hindsight bias. Such cross-cutting predictions are worth exploring.

But my main takeaway is that Enke and Graeber’s empirical work together with my simulations suggest that something very much like higher-order uncertainty is both easy to measure and psychologically-important for understanding deviations from Standard-Bayesian models. These methods—of measuring ambiguity, and using simulations of Ambiguous-Bayesian models to making predictions—preview what is to come in Chapters 7–10. There I will show that my model of ambiguity (unlike Enke and Graeber’s, since it is Standard-Bayesian) can *also* account for large swaths of empirical trends surrounding hindsight bias, confirmation bias, polarization, and overconfidence—as well as make novel (and correct) predictions about the role of ambiguity in these biases.

3.4 How to think about (our) statistics

When we make new predictions, we’ll need to evaluate them against the experimental data with statistical models. I want to say something about how we should think about the statistical methods I’ll be using, for two reasons. First, I want to show how even if you don’t know statistics, you can still understand the experimental results. Second, experts often have idiosyncratic understandings of statistical methods, so I want to be clear about how *I’m* understanding them.

In the past decade, the ‘replication crisis’ has generated much (warranted) skepticism about large swaths of behavioral science. Many of the surprising, widely-publicized findings of psychology and related fields have failed to replicate: independent attempts find the effect have yielded null effects or much smaller effect sizes.¹⁷ This pattern has a long list of causes, but two important ones are (1) the lack of a theoretically-motivated and unified theory, and (2) the use of questionable statistical practices commonly known as ‘*p*-hacking’—removing outliers to favor your hypothesis, running many tests (or gathering more data) until you get a statistically significant result, etc.

This book is offering a theory of bias—so it is obviously aiming to address (1). But when I venture into experiments, I want to also address (2), so I’ll try to follow the best practices promoted by the open science movement.

That means being transparent. I’ll ‘pre-register’ all studies before collecting data: specifying exactly how many participants will be recruited, exactly what exclusions I’ll apply, and exactly which statistical models I’ll run. And I’ll post versions of the raw data, and full code used to analyze it, on online repositories.¹⁸

¹⁷See e.g. Ioannidis 2005; Simmons et al. 2011; OSF 2015.

¹⁸[TODO: ResearchBox link]

It also means using rigorous statistical methods, such as ‘random-effects’ models—models that are aware of potential correlations in the data, such as those between repeated measures from a given subject. This avoids treating datapoints as independent when they’re not, and so makes the models less likely to be overconfident about the effects.

For example, suppose we want to know whether giving people ice cream makes them more likely to donate to charity. We recruit two people—Joe and Jim—and get many datapoints from each of them. For a year, every day we give Joe ice cream and Jim peanuts, and then ask each of them how much they want to give to a charity. We end up with $365 \times 2 = 730$ datapoints. Suppose Joe gives around \$4 every day, while Jim gives around \$3. A naive statistical test would be extremely confident that ice cream increases charitable giving: we have 365 instances of ice-cream-receivers giving \$4, and 365 instances of peanut-receivers giving \$3. But, of course, those datapoints are correlated: it might just be that *independent* of the ice cream, Joe is inclined to give more money than Jim. Since we only have two participants, we effectively have very *few* datapoints to test the general connection. ‘Random-effects’ models are ways of specifying our statistical models that allow them to be aware of these sorts of correlations amongst our datapoints.

The random-effects models I’ll use are all ‘Bayesian’ statistical models (run using BRMS, in R), using default priors unless otherwise stated. We can think of the statistical model as a (Standard-)Bayesian ‘golem’ that has a very simple model of the world, but is very good at updating that model in response to large datasets (McElreath 2018). Effectively, it’s a Bayesian agent who starts out with very weak (flat) priors over the parameters specified in the model, and is certain that the data it’ll observe will be drawn from the assumed (e.g., normal) distributions determined by those parameters. It then does Bayesian conditioning on the data from our experiment,¹⁹ and outputs a probability distribution over the parameters of interest capturing its posterior beliefs about where they are likely to fall.

When I report statistics from experiments, I will be reporting features of these posterior distributions, like estimates (expectations) and ‘credible intervals’ that say how confident our Bayesian golem is that the parameter is close to the estimate. I use Bayesian models not because I’m a rabid Bayesian—‘frequentist’ models can be just as sophisticated and informative. Rather, I use them because it’s easy to interpret what the model is saying, at least at a high level: We imagine a simple Bayesian agent, feed them data, and see how confident they are of various hypotheses. Such information is useful for telling *us* how to update *our* beliefs. Often we should (roughly) defer to the Bayesian agent’s opinions—after all, it has much more data than we do. So its opinions tell us how to nudge *our* beliefs in light of the data.

I’ll do my best to tuck away the details of statistical models in footnotes or optional empirical (€) subsections. Even if you’ve never done any statistics, you should be able to understand what to make of the results simply by looking at the plots and reading my descriptions.

¹⁹Actually, we run a sampling algorithm that *approximates* what the posterior would be, conditioned on the data. As discussed in Chapter 2, exact Bayesian conditioning of complex models on substantial data is intractable.

3.5 What Next?

You now know the notation I'll be using, the distinction between subjective ('uninformed') probabilities and informed probabilities, how I'll be using simulations and statistics. If you'd like to get to applications, jump to Part III ('Uses'). There I'll show how ambiguity as higher-order uncertainty offers a unified explanation of hindsight bias, confirmation bias, polarization, and overconfidence. This will constitute my case for the hypothesis that we are Reasonably Polarized—that people reasonably approximate Bayesian solutions to the problems they face, and that ambiguity is what leads to bias and polarization. Part IV ('Upshots') will ask what follows, if we accept this hypothesis.

On the other hand, if you want a better theoretical grasp of the workings and foundations of Ambiguous-Bayesian models, read on to Part II ('Theory').