

Chapter 5

Ambiguous Bayesianism

Abstract

Under clarity, Standard Bayesianism is inevitable. But clarity is not inevitable. This chapter shows how to use probability frames to model ambiguity as genuine higher-order uncertainty. It isolates the structures implicit in such frames, shows how to avoid common pitfalls in reasoning about them, and formulates ‘Ambiguous Bayesianism’ as a weakening of Standard Bayesianism. When evidence is partitional, even under ambiguity, conditioning maximizes expected accuracy.

- citations
- Ellsberg
- proofs
- trim notational niceties
& probabilistic reasoning

5.1 Plan

Standard-Bayesian models presuppose clarity. But our opinions are often ambiguous—we are uncertain what we think. And we are not irrational for that.

This chapter shows how to introduce ambiguity into Bayesian models in a maximally conservative way. We’ll first see why we can’t model ambiguity simply as uncertainty about another, more-ideal probability function (§5.2). I’ll then show how we can use probability frames to write down simple, tractable models of genuine higher-order uncertainty about your own credences (§5.3), showing that it generates features of ambiguity like noise, self-doubt, and biases (§5.3.1). The rest of the chapter will show how to use (and avoid the pitfalls of) ambiguous models. §5.4 shows why natural ‘Reflection’ principles fail under ambiguity, but why defensible variants of them helpfully constrain the models—leading to a formulation of ‘Ambiguous Bayesianism’ as a simple variant of Standard Bayesianism. §5.5 will show that even when priors are ambiguous, when evidence is partitional, conditioning maximizes expected accuracy. §5.6 will use the structures we’ve identified to respond to the common worry that higher-order uncertainty is somehow unstable.

5.2 Probabilistic Uncertainty

I said I’d introduce ambiguity to Bayesianism in the most-conservative way possible. But to get there, we need to see the problem with a common, even *more* conservative approach—one that does

not amount to a model of ambiguity at all.

Imagine a strapping young probabilist, setting out to model ambiguity. She has some examples in mind. Some are cases where people form opinions in low-information settings—*Does Kevin own a dozen spoons?* or *Will the marble drawn from this urn of unknown composition be red?* Others are cases of conflicting information—*Quinn said Kevin owns lots of spoons, but Elizabeth said he owns very few. How many do I think he has?* Both involve a sense of being uncertain what probabilities to assign to the target claim.

Our probabilist knows her stuff, so he’s aware of two features of probabilistic talk and modeling.

First feature. Ambiguity is uncertainty about what to think. What you think is modeled with probabilities. So ambiguity is uncertainty about probabilities. But *which* probabilities? ‘Probable’, ‘likely’, and so on are notoriously context-sensitive. ‘It’s at least 60%-likely to rain’ could express my high credence, or state a fact about objective chances, the strength of the evidence, or what a rational version of myself would think. So our probabilist—who wants to model uncertainty about probabilities—has some room to maneuver.

Second feature. It is straightforward and commonplace to model some types of uncertainty about probabilities—namely, uncertainty about your *future, more-informed* probabilities. That is what we did last chapter, when we modeled your priors P as uncertain about your posteriors P^+ , which were the result of informing your priors about the true answer to a question.

Our two features make it extremely likely that our probabilist will think, ‘Aha! To model ambiguity, I just need to introduce *another* probability function, \mathcal{P} , that your priors P are uncertain about. The values of \mathcal{P} can vary across the possibilities you leave open, just like the values of your posterior P^+ do.’ Thus in addition to your prior P , there will be another probability function, \mathcal{P} —often interpreted as *what the evidence supports*, or *what the objective probabilities are*, or *what a more-ideal version of yourself would think*. Your uncertainty about what to think is modeled as your prior P being uncertain about the ideal probability function \mathcal{P} .

The analogy between the ideal function \mathcal{P} and your posteriors P^+ might give our probabilist pause. After all, there are plenty of cases where you’re unsure what your posteriors will be—any case in which you might learn something nontrivial—in which your prior opinions aren’t ambiguous. If you know that this coin is fair and that you’re about to observe the outcome, then your prior of 0.5 in *heads* is clear, despite being unsure whether your posterior will be 1 or 0. Might uncertainty about \mathcal{P} simply model situations like this?

Three other features will likely stiffen the spine of our strapping probabilist. First, she has probably been taught to model your prior P with a rigidly-specified probability distribution π . And she knows from the first time she accidentally wrote ‘ $P(x = 7) = 0.4$ ’ that there’s an important distinction between constants (like ‘ x ’) and random variables (like ‘ X ’). She realizes that π is a constant, so she thinks (correctly) that *it* can’t literally be an object of uncertainty. She infers (incorrectly) that your prior can’t literally be an object of uncertainty.

Second, she will ask herself what relationship needs to hold between your prior P and the ideal function \mathcal{P} in order for you to count as deferring to it, and wanting to conform to it. The obvious answer that will emerge—familiar from other modeling scenarios—is a version of the Reflection principle we discussed in §4.4.2. You’re unsure what \mathcal{P} is; but *conditional* on the ideal credence function assigning 0.6 to *dozen spoons*, you should assign 0.6 to it. Formally: $P(q|\mathcal{P}(q) = x) = x$. If

our probabilist is especially adventurous, she may experiment with plugging P itself in for \mathcal{P} in this principle, yielding $P(q|P(q) = x) = x$. But she'll check what the models which satisfy this principle look like. Conforming her suspicions, she will find that this version of the constraint trivializes the uncertainty: if P is to satisfy it, it must be certain of the value of P (see §5.4.1).

Third our strapping probabilist may present her ideas at conferences. There she may run into well-known researchers, who have seen plenty of young probabilists—not yet fully understanding the mathematics—write down principles of higher-order probability that are mathematically incoherent. Those researchers will refer our strapping probabilist to classic works by Savage and de Finetti arguing that higher-order probabilities are either incoherent or inert.

Her spine stiffened, our probabilist will likely return to her original idea: you have ambiguity when your prior P knows its own values, but is uncertain about a more-ideal probability function \mathcal{P} . She will diagnose the difference between the clear case of the coin flip—where you don't know what your future, more-informed self will think—and her ambiguous cases as interpretive, not mathematical. In the coin case, you don't know what your *future* self P^+ will think, but you do know what *an ideal person with your current evidence*, \mathcal{P} , would think. In contrast, in the spoons case, you *don't* know what an ideal person with your evidence would think.

Our strapping probabilist is perfectly reasonable. But she is wrong. Although she has come up with a sensible model of a form of probabilistic uncertainty, she has not come up with a model of genuine *ambiguity*—at least not one up to the task of explaining the failure of Reasonable Convergence and the biases that lead to it. The model she has generated is Standard Bayesian. By imposing Reflection, it implies that \mathcal{P} is clear—the ideal credence is certain of its own values. And unless she has deliberately made other choices, your priors P will be implicitly be higher-order certain. As a result, all the limitative results mentioned above (§4.5) and below (Part III) apply.

This is not a true story. But I suspect it has echos of truth in the etiology of a variety of extant models that claim to model ambiguity. The models vary in how they interpret the ideal function \mathcal{P} . Sometimes it is what increasingly-expert versions of yourself would think (e.g. Gaifman 1988). Sometimes it is what you would think if you knew the true value of an unknown variable, as in the Hierarchical Bayesian Models used in cognitive science (e.g. Pearl 1988; Henderson et al. 2010; Perfors et al. 2011; Ullman and Tenenbaum 2020). Sometimes \mathcal{P} is some sort of objective probability like the chances or propensities (e.g. Klibanoff et al. 2005; Roush 2009, 2016). Sometimes it is an ideally-rational version of yourself (e.g. Frisch and Baron 1988; Christensen 2010b; Schoenfeld 2012; Hedden 2019). Sometimes it is an implicit probability function that you only have noisy access to (e.g. Enke and Graeber 2023). Sometimes something else. All of these are interesting models. But none of them are models of genuine ambiguity, understood as (rational) uncertainty about *your own* degrees of uncertainty.

Consider how we would use this to model your uncertainty about whether I own a dozen spoons. You're unsure whether I do (d) or don't (\bar{d}) own a dozen spoons. You're also unsure what the ideal function thinks—let's say you leave open that it is either 0.4- or 0.6-confident of d . Combined, there are four possibilities: $d_4, \bar{d}_4, d_6, \bar{d}_6$. In our stochastic matrix notation, we can write up the possible

?Add or swap out (Enke) point that they all have certainty about their levels of cognitive uncertainty?

values the ideal function \mathcal{P} as follows:

$$\mathcal{P} = \left(\begin{array}{c|cccc} & d_4 & \bar{d}_4 & d_6 & \bar{d}_6 \\ \hline d_4 & 0.4 & 0.6 & 0 & 0 \\ \bar{d}_4 & 0.4 & 0.6 & 0 & 0 \\ d_6 & 0 & 0 & 0.6 & 0.4 \\ \bar{d}_6 & 0 & 0 & 0.6 & 0.4 \end{array} \right)$$

What are your opinions P about these four possibilities? To satisfy our version of the Reflection principle, we must have that $P(d|\mathcal{P}(d) = 0.4) = 0.4$ and $P(d|\mathcal{P}(d) = 0.6) = 0.6$, but that doesn't settle how confident you are *that* the ideal function is 0.4 (vs. 0.6) in d . Since your prior P is a probability function, there must be some number x such that $P(\mathcal{P}(d) = 0.4) = x$. Suppose $x = 0.5$. (The reasoning generalizes.) Then your prior is a 50-50 average of the two possible values for the ideal function \mathcal{P} , i.e. $P = (0.2, 0.3, 0.3, 0.2)$.

Crucial question: what does your prior P know about itself? In other words: *in* those four possibilities it leaves open, what value does P take? If our strapping probabilist uses a constant π to model your prior—as her real-world counterparts usually do—it will have the same value at each of those possibilities. Thus, implicitly, although your prior P is uncertain what the ideal function \mathcal{P} is, it is certain of *exactly how uncertain it is*—certain, for example, that $P(\mathcal{P}(d) = 0.4) = 0.5$. We can write this out explicitly by saying that the ‘variable’ probability function P describing your credences takes the same value (π) at each of the four possibilities:

$$P = \left(\begin{array}{c|cccc} & d_4 & \bar{d}_4 & d_6 & \bar{d}_6 \\ \hline d_4 & 0.2 & 0.3 & 0.3 & 0.2 \\ \bar{d}_4 & 0.2 & 0.3 & 0.3 & 0.2 \\ d_6 & 0.2 & 0.3 & 0.3 & 0.2 \\ \bar{d}_6 & 0.2 & 0.3 & 0.3 & 0.2 \end{array} \right)$$

Thus if we set \mathcal{P} to P^+ in a probability frame, we have written down a Standard-Bayesian model.

Call this **probabilistic uncertainty**—uncertainty about a contextually-relevant, in-some-way-ideal probability function like \mathcal{P} . Such probabilistic uncertainty is mathematically identical to the sort of uncertainty that you have about your *future* opinions in Standard-Bayesian models. Nothing qualitatively different will emerge.¹

¹This fact is often obscured by theorists' focus on much more complicated models. For instance, it is widespread to use Beta models to capture your uncertainty about some probabilistic variable. A beta distribution is just a mathematically-convenient ways of representing your uncertainty about the objective chances of a repeated binary process (success vs. fail; etc.) when you are certain that the trials are independent, but don't know what the probability of each outcome is. Think of a coin of unknown bias. Beta(1,1) and Beta(10,10) both have a mean of 0.5—i.e. their expectation for the objective probability of heads is 0.5—but the former is a flat distribution between 0 and 1, while the latter is peaked around 0.5. Thus the Beta(10,10) represents the opinions of someone who is confident that the bias of the coin is near 50%, while the Beta(1,1) represents someone who's uniformly unsure what the bias might be, between 0 and 1. Neither explicitly encodes what the agent's higher-order opinions about *their own* uncertainty. But standardly the beta distribution isn't itself treated as an object of uncertainty, and thus the agent is implicitly higher-order *certain* that they have a Beta(1,1) (or a Beta(10,10)) distribution over the bias of the coin. It is one thing to be unsure what the bias of a coin is; it is quite another to be uncertain about how much uncertainty you have about the bias of the coin.

If we use a model like this, and we want to explain why people behave differently under ambiguity than clarity, we'll need to do something drastic. Some theorists change the decision rule, saying that instead of maximizing expected value our agents should calculate the expectation of a nonlinear function ϕ of \mathcal{P} , and maximize the expectation of *that* (e.g. Klibanoff et al. 2005). Others say that rather than using our expectation, we should perform some other operation using the possible expectations that \mathcal{P} assigns—for instance, paying special attention to the most-pessimistic expectation the ideal function might have (Ellsberg 1961). This strategy is often paired with a denial that there even *is* a precise probability function P that represents your priors—perhaps your opinions are simply to be represented with the *set* of possible ideal functions \mathcal{P} .² Other times it is said that your priors have more structure than that, but less structure than a full probability function.³

As a model of ambiguity, these moves are desperate. The results usually violate the basic motivations of Bayesian theories of rationality, including the ‘value of rationality’ that we’ll formalize in Chapter 6. They also struggle to delineate their scope of application. Probabilistic uncertainty is pervasive. *Too* pervasive. For here’s a contextually-relevant, in-some-way-ideal probability function: the omniscient probability function $\mathbb{1}$. This function assigns 1 to all truths and 0 to all falsehoods. Whenever you are uncertain about q , you’re uncertain about the values of $\mathbb{1}(q)$. More generally, we are pretty much always uncertain about what our future opinions will be, what a more-ideal (and perhaps more-informed) version of ourselves would think, what the objective chances are, and so on. For example, we’re uncertain what an ideal version of ourselves would think about whether the 99th digit of pi is between 1–6. But our opinion about this isn’t ambiguous—we know that the reasonable credence to have, given our limitations, is 0.6, despite knowing that the ideal credence is either 0 or 1. It would be madness to treat this case in the way the above theories predict. And it would be hopeless to try to find a bright line between this case and the theorist’s preferred cases where we *should* deviate from Standard-Bayesian dictates.

And we needn’t. There’s a better way to model ambiguity: by modeling uncertainty about your own opinions in the same way that we model uncertainty about everything else.

5.3 Higher-Order Uncertainty

How do we model you as uncertain about the value of a variable X ? By having you leave open two worlds w and v (i.e. $P(w) > 0$ and $P(v) > 0$) where X takes different values ($X_w \neq X_v$). When X is your future credence function P^+ , this involves your prior P leaving open w and v such that $P_w^+ \neq P_v^+$.

What about when X is your *current* credence function? We should model it the exact same way: your prior P must leave open two worlds w and v where your prior differs, i.e. such that $P_w \neq P_v$. Once we’ve started using probability frames (W, \mathbf{a}, P, P^+) —where your prior and posterior are explicitly modeled as variable probability functions—this is easy. We simply need to allow your prior P to vary between two worlds, and have those worlds assign each other positive probability.

²Ellsberg e.g. 1961; Levi e.g. 1974; Seidenfeld and Wasserman e.g. 1993; Joyce e.g. 2010; Schoenfield e.g. 2012; Trautmann and van de Kuilen e.g. 2015; Moss e.g. 2018.

³Dempster e.g. 1967; Shafer e.g. 1976; Moss e.g. 2018; Staffel e.g. 2020; Campbell-Moore e.g. 2021.

Here’s an example. You’re unsure whether I own a dozen spoons (d) or not (\bar{d}). Your opinion is ambiguous—you’re unsure exactly what your credence is. Suppose you’re certain that it’s either 0.45 (*low*) or 0.55 (*high*), but you’re unsure which. Then we need one set of worlds in which $\langle P(d) = 0.45 \rangle$ another in which $\langle P(d) = 0.55 \rangle$, and the probability function associated with each needs to leave open both sets of worlds. Here’s a simple example of a frame (or the prior part of a frame, (W, \mathbf{a}, P) ; forget P^+ for now) which does this, with the actual world $\mathbf{a} = d_l$ in bold :

$$P = \left(\begin{array}{c|cccc} & d_l & \bar{d}_l & d_h & \bar{d}_h \\ \hline \mathbf{d}_l & 0.15 & 0.45 & 0.3 & 0.1 \\ \bar{d}_l & 0.15 & 0.45 & 0.3 & 0.1 \\ d_h & 0.1 & 0.3 & 0.45 & 0.15 \\ \bar{d}_h & 0.1 & 0.3 & 0.45 & 0.15 \end{array} \right)$$

Oops, flipped this from ch_NU example. FIX

Figure 5.1: A simple model of ambiguous opinions about whether I own a dozen spoons (d).

Let’s talk through this slowly.

There are two possible distributions that might be your prior, $\pi_l = (0.15, 0.45, 0.3, 0.1)$ —which you have at d_l and \bar{d}_l —and $\pi_h = (0.1, 0.3, 0.45, 0.15)$, which you have at d_h and \bar{d}_h . The first assigns low credence $0.15 + 0.3 = 0.45$ to d , while the second assign high credence $0.1 + 0.45 = 0.55$ to d . You have low credence at worlds d_l and \bar{d}_l , so $\langle P(d) = 0.45 \rangle$ is true there, while you have high credence at d_h and \bar{d}_h , so $\langle P(d) = 0.55 \rangle$ is true there.

How likely do you think each of these two hypotheses about your credences are? That depends—obviously—on what your credences are. If you have π_l , then you think it’s $0.15 + 0.45 = 0.6$ -likely that you have credence 0.45 in d (i.e. that $\langle P(d) = 0.45 \rangle$), and you think it’s 0.4 -likely that that you have credence 0.55 in d (i.e. that $\langle P(d) = 0.55 \rangle$).⁴ Meanwhile, if you have π_h , then you have credence 0.4 that you have credence 0.45 in d and credence 0.6 that you have credence 0.55 in d .

In this frame, the actual world is $\mathbf{a} = d_l$, so you *in fact* have credence function π_l . But you have a different credence function at other worlds. $\langle P = \pi_l \rangle$ is true at the first two worlds, d_l and \bar{d}_l , while $\langle P = \pi_h \rangle$ is true at the second two, d_h and \bar{d}_h . Since you are in fact at d_l , you *in fact* are 0.45-confident of d (your credence in d , $P(d)$, in fact equals $\pi_l(d)$). But you aren’t certain of that—instead you’re only 60%-confident of it. For if you instead had credence function π_h —which you do at d_h and \bar{d}_h —then you would be 0.55-confident of d (your credence in d , $P(d)$, would equal $\pi_h(d)$). Thus you only assign 60%-credence to the (in fact, true) hypothesis that your credence function is π_l , and you assign 40%-credence to the (in fact, false) hypothesis that your credence function is π_h . Your actual credence function (π_l) is not certain that your actual credence function is π_l (not certain that $\langle P = \pi_l \rangle$). We can write this using $P_{\mathbf{a}}$ as a rigid designator for your actual credence function: although $P_{\mathbf{a}} = \pi_l$, it is also true that $P_{\mathbf{a}}(P = \pi_l) = 0.6 < 1$.

It might seem puzzling that we could model higher-order probabilities with so simple a structure. As soon as you consider claims of higher-order probability, it’s natural to think that you must assign probabilities to infinitely many different claims: the claim that d (d), the claim that *you’re 0.45-confident of d* ($\langle P(d) = 0.45 \rangle$), the claim that *you’re 0.6-confident that you’re 0.45-confident of d* ($\langle P(P(d) = 0.45) = 0.6 \rangle$), and so on. But we only have 4 worlds, and so (since propositions are sets of worlds) $2^4 = 16$ different propositions. Yet don’t we need *infinitely* many different ones? So—a

⁴Formally, $\pi_l(P(d) = 0.45) = \pi_l(\{d_l, \bar{d}_l\}) = 0.15 + 0.45 = 0.6$, while $\pi_l(P(d) = 0.55) = \pi_l(d_h, \bar{d}_h) = 0.3 + 0.1 = 0.4$.

strapping probabilist might conclude—we must recursively define an infinite structure to model genuine higher-order probabilities. Break out the impenetrable math. Right?

Wrong. Compare the following line of reasoning. ‘As soon as you consider logical operations on claims, you have infinitely-many different claims: the claim that d the claim that $\neg d$, the claim that $\neg\neg d$, and so on. But we only have 4 worlds, so 16 propositions. Yet we need *infinitely* many different propositions, not just 16!’

Both lines of reasoning go wrong in the first step. Just because you can infinitely iterate an operation of propositions (going from q to $\neg q$, or from q to $\langle P(q) = x \rangle$) doesn’t mean that you have infinitely-many *different* propositions. Recall that in any set-of-worlds model of propositions, there are many different names for the same proposition. Obviously, in this model $\neg\neg d$ picks out the same proposition as d —namely, $\{d_l, d_h\}$. Less obviously, in this model $\langle P(P = \pi_l) = 0.6 \rangle$ picks out the same proposition as $\langle P = \pi_l \rangle$ —namely, $\{d_l, \bar{d}_l\}$.

Because of this, what look like infinitely-many different claims are actually just infinitely-many different ways of saying finitely many things. Recall that in our models we define the negation operator using operations on sets of worlds: for any $q \subseteq W$, $\neg q := \{w \in W : w \notin q\}$. For example, $\neg d$ is the claim $\{\bar{d}_l, \bar{d}_h\}$. Similarly for probability operators: for any $q \subseteq W$, $\langle P(q) = x \rangle := \{w \in W : P_w(q) = x\}$. For example, $\langle P(d) = 0.45 \rangle$ is the claim $\{d_l, \bar{d}_l\}$. More generally, for any property ϕ that a probability function might have, $\langle P \text{ has } \phi \rangle := \{w \in W : P_w \text{ has } \phi\}$. For example, $\langle P(d) \geq 0.5 \rangle = \{d_h, \bar{d}_h\}$.

Thus claims about your probabilities are sets of worlds. Since there are only 16 different sets of worlds in this model, that means there are at most 16 different claims about your probabilities in this model.⁵ There are ever-so-many ways to say the same thing—for example, $\langle P(d) \geq 0.5 \rangle$, $\langle P(d) \geq 0.51 \rangle$, $\langle P(d) \geq 0.511 \rangle$, are all ways of saying that your credence function is π_h .

Because of this, we can recursively ‘unpack’ arbitrarily-complex statements about your probabilities of probabilities (of probabilities...) into simple claims about which set of worlds you’re in. For instance, consider the Big Claim that you are 40%-confident that you are 60%-confident that you are 45%-confident that I own a dozen spoons:

$$\langle P(P(P(d) = 0.45) = 0.6) = 0.4 \rangle \quad (\text{Big Claim})$$

To figure out what set of worlds this picks out, start with the inner-most probability claim and translate it to its set-of-worlds equivalent. Since $\langle P(d) = 0.45 \rangle$ is $\{d_l, \bar{d}_l\}$, that yields:

$$\langle P(P(\{d_l, \bar{d}_l\}) = 0.6) = 0.4 \rangle$$

Notice that the next-inner-most probability claim has now become a (seemingly, ‘first-order’) claim about the probability of a set of worlds, so we can directly see where it’s true: $\langle P(\{d_l, \bar{d}_l\}) = 0.6 \rangle$ is true at worlds that assign 0.6 to $\{d_l, \bar{d}_l\}$, i.e. at $\{d_l, \bar{d}_l\}$. Substituting this into the above yields:

$$\langle P(\{d_l, \bar{d}_l\}) = 0.4 \rangle$$

⁵In fact, since there are only 2 different probability functions you might have, there are only $2^2 = 4$ claims about your probabilities: $\{\}$, $\{d_l, \bar{d}_l\}$, $\{d_h, \bar{d}_h\}$, and $\{d_l, \bar{d}_l, d_h, \bar{d}_h\}$. These are the (possible empty) unions of cells of the partition induced by the question, ‘What is your credence function?’, i.e. $\{\langle P = \pi_l \rangle, \langle P = \pi_h \rangle\}$.

So the Big Claim is equivalent to the claim that $\{d_l, \bar{d}_l\}$ has a probability of 0.4. That’s true at $\{d_h, \bar{d}_h\}$ —so the Big Claim is just a complicated way of saying that π_h is your credence function!

This ‘unpacking’ strategy might feel like a trick, but it encodes a philosophical insight. You are part of the world, so facts about you—including facts about what you believe—are facts about the world. This is obvious third-personally: when I’m unsure whether you believe q , I’m unsure what the world is like: ‘Am I in a world in which you believe q , or not?’ It’s also obviously true for your first-person *future* probabilities—that’s what we saw in Chapter 4.

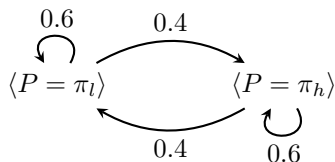
What’s less obvious, but equally true, is that the same holds in the first-person present tense: when *you’re* unsure whether you believe q , you’re unsure what the world is like. Thus when you believe things, you consider various possibilities—and *in those possibilities you consider, there are facts about what you believe*. So long as those possibilities are sufficiently fine-grained to determine what you believe, we can read off your higher-order beliefs from the facts about what you believe in those possibilities. Call this *Hintikka’s Insight*, after Jaakko Hintikka (1962), who I believe was the first to formulate and develop this idea for propositional attitudes.⁶

What exactly it takes to represent your beliefs at each world depends on how we’re modeling beliefs. In Hintikka’s case, he said that what you believe is determined by a set of worlds B_a —the worlds that you leave open. You believe a proposition q iff q is true in all the worlds in B_a , i.e. iff $B_a \subseteq q$. Since each possibility needs to determine what you believe, he modeled your beliefs with a function B from worlds w to the set of worlds B_w that you leave open at w . The proposition that you believe q is then the set of worlds w in which all the worlds you leave open are q -worlds: $\langle Bq \rangle := \{w : B_w \subseteq q\}$. This definition works regardless of whether we plug in a proposition ‘about the world’ (*it’s raining*) or ‘about your beliefs’ (*you believe it’s raining*) for q . Or, put better: claims about your beliefs are claims about the world, and Hintikka merely defined a model where the worlds are sufficiently fine-grained to represent that fact.

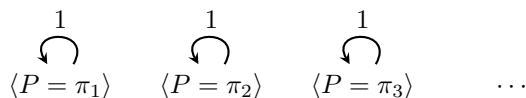
In our case, we say that what you believe (and to which degree) is determined by a probability function π . Since each possibility needs to determine what you believe, we model your beliefs with a function P from worlds w to probability distributions P_w . The proposition that you assign 0.6-credence to q is then the set of worlds w such that the probability function you have at w , P_w , assigns 0.6-credence to q : $\langle P(q) = 0.6 \rangle := \{w : P_w(q) = 0.6\}$. This definition works regardless of whether we plug in a proposition ‘about the world’ (*Kevin owns a dozen spoons*) or ‘about your probabilities’ (*you’re 55%-confident that Kevin owns a dozen spoons*) for q . Or, put better: claims about your probabilities are claims about the world, and we’ve merely defined a model where the worlds are sufficiently fine-grained to represent that fact.

⁶Many authors in many fields have used variants of Hintikka’s insight. It was pioneered by Kripke (1963) and Hintikka (1962) for modal and epistemic logic, and then picked up by metaphysicians, linguists, and epistemologists—see e.g. Stalnaker 1968, 1975; Lewis 1970, 1973; Kratzer 1977, 1986; Williamson 2000, 2013 for classic examples. In the probabilistic case, it was pioneered by Harsanyi and has mainly been used by epistemic game theorists to study the interpersonal case (e.g. Aumann 1976; Aumann and Brandenburger 1995; Aumann 1999; Geanakoplos 1992; Lederman 2015, 2018; van Ditmarsch et al. 2015). Some have studied the single-agent case (Gärdenfors 1975; Skyrms 1980; Gaifman 1988); these models have increasingly been used in recent years to study the case of (rational) people who are unsure what their own (rational) beliefs are—for example, Geanakoplos 1989; Hild 1998; Samet 1999, 2000; Williamson 2000, 2008, 2014, 2019; Schervish et al. 2004; Lasonen-Aarnio 2013, 2015; Campbell-Moore 2016; Ahmed and Salow 2018; Salow 2018, 2019; Dorst 2019, 2020a, 2023b; Dorst et al. 2021; Levinstein 2023; Das 2022, 2023; Zhang and Meehan 2025.

Returning to our frame from Figure 5.1, it may be easier to see what is going on by coarsening the picture, and focusing on the different probability functions you might have—noticing that, by Hintikka’s insight, this will determine what probabilities you assign to having those probabilities. You might have π_l or π_h . If you have π_l , you’re 60%-confident that you have π_l and 40%-confident that you have π_h . Meanwhile, if you have π_h , you’re 40%-confident that you have π_l and 60%-confident that you have π_h . We can represent these facts with a ‘Markov-chain-style’ diagram⁷:



Drawn this way, it’s easy to see the mathematical line between clarity and ambiguity. Every probability frame can be partitioned into classes that agree on what your prior is—the possible answers to the question, ‘What is your prior?’: $\{\langle P = \pi_1 \rangle, \langle P = \pi_2 \rangle, \dots\}$. You have *clarity* when each cell of this partition assigns probability 1 to itself: whatever your probability function is, you are certain that that’s what it is:



You have *ambiguity* when some of these cells assign positive probability to each other.

5.3.1 Features of Ambiguity

Could it really be so simple?

In a way, no. Though conceptually simple, ambiguous frames quickly get philosophically and mathematically subtle. The rest of Part II shows how to construct and constrain them.

But in a way, yes. That change—representing your prior P as a *variable* probability function, and allowing it to assign non-maximal credence to having the value it actually does—generates a model of ambiguity with all the features we’re after. Let’s illustrate by showing how the simple spoon model gives rise to cognitive noise, self-doubt, and hindsight bias.

First, **cognitive noise**. How likely do you think it is that I own a dozen spoons? In trying to answer truthfully, your answer will inevitably exhibit noise or stochasticity. As we’ve seen (Chapter 2 and §4.5), Standard Bayesians would do no such thing. Since they have clarity, they are certain of what their probabilities are. So long as Certainty Precludes Chance, they should reliably report their credence—for example, by *estimate*-sampling from their distribution over $P(d)$, and reporting the result.

But *Ambiguous* Bayesians—like you and me—inevitably exhibit noise when sampling from their opinions. To illustrate, suppose your opinions are accurately modeled by Figure 5.1. If you outcome-sample, the result is (as usual) noisy. For example, suppose you draw 10 samples and report the

⁷Named after the standard way of diagramming Markov chains, which consist of a set of states and transition probabilities between states. They are formally equivalent to probability frames, but have a different interpretation.

proportion of them in which d is true. Then—since $P_a(d) = 0.45$ —your response propensities will follow a Binomial(10,0.45) distribution as shown in Figure 5.2. But that’s not surprising—the same goes if *Standard* Bayesians outcome-sample. Moreover, outcome-sampling generates noise in ways that violate Certainty Precludes Chance. In this example, you are certain that your credence in d is between 0.45 and 0.55, yet outcome-sampling yields a nontrivial chance of reporting a credence as low as 0.2 or as high as 0.8.

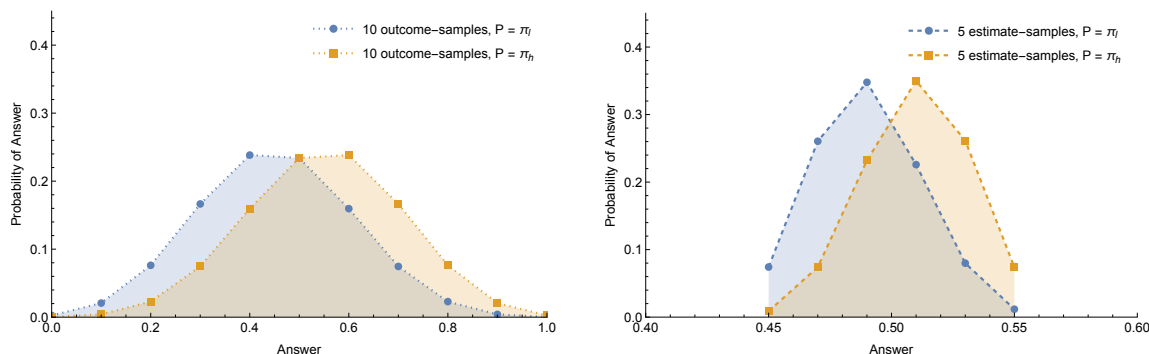


Figure 5.2: Under ambiguity, even estimate-sampling is noisy. Plots show your answering propensities for $P(d)$ given the model in Figure 5.1 given outcome-sampling (left) and estimate-sampling (right).

As we’ve seen, the sensible way around this is to do *estimate*-sampling: sample from your distribution over *what your credence in d is* (i.e. over $P(d)$) and report the mean. Under clarity there would be no noise in such samples: since you are certain of the value of $P(d)$, every sample you draw will yield that value. But *under ambiguity*, noise remains. After all, you are in fact 60%-confident that $\langle P(d) = 0.45 \rangle$ and 40%-confident that $\langle P(d) = 0.55 \rangle$. Thus if you draw a single sample of $P(d)$ and report it, you are 60%-likely to report ‘0.45’ and 40%-likely to report ‘0.55’. Meanwhile, if your credence function were instead π_h , you’d be 40%-likely to report ‘0.45’ and 60%-likely to report ‘0.55’. Either way, there’s noise. If you draw more samples, then your propensities to give various answers blend between 0.45 and 0.55—respecting Certainty Precludes Chance—but remain noisy. For example the right of Figure 5.2 shows your propensities if you report the mean of 5 estimate-samples of $P(d)$. Upshot: like you and me, Bayesians with ambiguous opinions inevitably have noisy elicitation.

Second, **self-doubt**. Suppose I offer you a choice between \$5 for sure, or a bet X that pays \$10 if I own a dozen spoons, \$0 otherwise. Suppose you know that the rational thing for you to do is maximize expected value by the lights of your (rational) credences P . What should you do? Presumably you’re unsure. By that I don’t just mean you’re unsure which option will yield more money; I mean you’re unsure which option is the *rational* one for you to take, by your own lights. You’re unsure which one *you expect* to yield a higher value, on average. More precisely: if we imagine repeatedly taking bets on independent claims q_i that you’re equally-confident in as d^8 , you’re unsure whether you expect the average payout of such bets to be greater or less than \$5.

As discussed in §4.5, Standard Bayesians would do no such thing. Since they have clarity, they’re always certain of which option maximizes expected value by their own lights—there is a number x

⁸That is: for all i , $P(q_i) = P(d)$, and P treats the q_i as mutually independent.

such that $P_a(\mathbb{E}_P(X) = x) = 1$. So if $x > 5$, they know that they should take X ; if $x < 5$, they know that they should take the sure \$5; and if $x = 5$, they know they should be indifferent.

But *Ambiguous* Bayesians—like you and me—are often unsure which option maximizes expected value by their own lights. Again, suppose your opinions are accurately modeled by Figure 5.1. Then you are unsure whether your credence function is π_l or π_h . If the former, then your expected value of X is $\mathbb{E}_P(X) = 0.45(10) + 0.55(0) = \4.50 . If the latter, it's $\mathbb{E}_P(X) = 0.55(10) + 0.45(0) = \5.50 . Thus you're 60%-confident that the rational thing for you to do is to take the sure \$5, and you're 40%-confident that the rational thing for you to do is to take the bet. (And if you instead had π_h , you'd be 40%-confident that the rational thing is to take the sure \$5, and 60%-confident that it's to take the bet.) Whichever option you take, you'll be uncertain whether it's the option that maximizes expected value.⁹ Upshot: like you and me, Ambiguous Bayesians often exhibit self-doubt.

Finally, **hindsight bias**. Guess what? I *do* own a dozen spoons. Now let me ask: how likely did you think that was, before I told you? You might find yourself tempted to think, 'Eh, pretty likely. I've noticed he's got a spoon-theme going in this book, so I sort of expected that.' Indeed, you might find yourself *more* tempted to think this than you would have if I had informed you that I *don't* own a dozen spoons.

If so, you exhibit hindsight bias. This is the finding wherein people increase their estimate for how likely they thought something was after they learn that it happened ('I knew it all along!'). We'll discuss it far more in Chapter 7; but for now, I want to sketch why Standard Bayesians won't exhibit it, while Ambiguous Bayesians will. Along the way, we can illustrate how to update an ambiguous frame like Figure 5.1.

Hindsight bias can be formalized as follows (see Chapter 7, and Hedden 2019). You have some prior estimate for what your prior in d is, $\mathbb{E}_P(P(d))$.¹⁰ Then you learn whether d is true, updating to P^+ . You then have an updated estimate for what your prior in d was, $\mathbb{E}_{P^+}(P(d))$. You exhibit hindsight bias if, upon learning that d is true, the latter estimate is higher than the former: if d , $\mathbb{E}_{P^+}(P(d)) > \mathbb{E}_P(P(d))$.

Standard Bayesians will never exhibit this. For at the prior time, they are certain of what their prior is: there is an x such that $P_a(P(d) = x) = 1$. And when they learn whether d , they update by conditioning: if d , $P^+(\cdot) = P(\cdot|d)$. Conditioning preserves certainties: if $P(q) = 1$, then $P(q|d) = 1$. So when learn that d , they're estimate for their prior in d remains the same: $\mathbb{E}_{P_a^+}(P(d)) = \mathbb{E}_{P_a}(P(d)|d) = \mathbb{E}_{P_a}(P(d)) = x$.

But Ambiguous Bayesians will often exhibit hindsight bias. Suppose your prior is modeled by Figure 5.1, and you are going to be told whether or not d is true. In situations like this, where the evidence you learn is clear (you are always certain whether you've learned d or $\neg d$), the sensible way to update is by conditioning on what you've learned. That is, for each world w , your posterior should be the result of conditioning your prior at w on what you learned at w : if $w \in d$, then

⁹Can our Bayesian *learn* about what they prefer by forcing themselves to (perhaps hypothetically) make a choice? Yes, their actions provide evidence about what they expect to be best. But since elicitation are noisy, they can't conclude with certainty that they've acted in the way they prefer—perhaps, due to cognitive noise, they chose an option with lower expected value. Moreover, note that this doesn't break the analogy. Real people *also* learn about what they prefer by (perhaps hypothetically) making choices. Recall the common advice about hard choices: 'Imagine the outcome of a coin flip commits you to a course of action; watch how it lands; and see how that makes you feel.'

¹⁰How does your estimate of your prior $\mathbb{E}_P(P(d))$ relate to your actual prior $P(d)$? Hold that thought.

$P_w^+(\cdot) = P_w(\cdot|d)$, and if $w \notin d$, then $P_w^+(\cdot) = P_w(\cdot|\neg d)$. The result of doing this update (rounded to 2 decimal places) in each world is displayed below:

$$P = \left(\begin{array}{c|cccc} & d_l & \bar{d}_l & d_h & \bar{d}_h \\ \hline \mathbf{d}_l & 0.15 & 0.45 & 0.3 & 0.1 \\ \bar{\mathbf{d}}_l & 0.15 & 0.45 & 0.3 & 0.1 \\ d_h & 0.1 & 0.3 & 0.45 & 0.15 \\ \bar{d}_h & 0.1 & 0.3 & 0.45 & 0.15 \end{array} \right) \quad P^+ \approx \left(\begin{array}{c|cccc} & d_l & \bar{d}_l & d_h & \bar{d}_h \\ \hline \mathbf{d}_l & 0.33 & 0 & 0.67 & 0 \\ \bar{\mathbf{d}}_l & 0 & 0.82 & 0 & 0.18 \\ d_h & 0.18 & 0 & 0.82 & 0 \\ \bar{d}_h & 0 & 0.67 & 0 & 0.33 \end{array} \right)$$

Now notice what conditioning on d does to your estimate of your prior. Initially, you are unsure whether your prior in d is 0.45 or 0.55. In fact you have π_l , so your credence is 60-40 between these two possibilities. Thus your estimate for your prior is $\mathbb{E}_{P_a}(P(d)) = 0.6(0.45) + 0.4(0.55) = 0.49$. (Meanwhile, if you had π_h , your estimate would be $0.4(0.45) + 0.6(0.55) = 0.51$.)

Then you learn d , and condition on it. This shifts your distribution to $\pi_l(\cdot|d) \approx (0.33, 0, 0.67, 0)$ —that is, your credence that your prior in d was 0.45 has dropped to 33%, while your credence that your prior in d was 0.55 has jumped to 0.67. Thus your posterior estimate for your prior is $\mathbb{E}_{P_a^+}(P(d)) \approx 0.33(0.45) + 0.67(0.55) = 0.517$. This is higher than your prior estimate of 0.49, exhibiting hindsight bias.¹¹

Why is this happening? Because you're unsure what your prior is, but you *trust* it in the sense that you expect it to be correlated with the truth. Thus learning what the truth is (that I *do* own a dozen spoons) provides evidence in favor of the hypothesis that your prior was high (that you had π_h) and against the hypothesis that your prior was low (that you had π_l). Since *in fact* your prior was low (in fact you had π_l), such evidence is misleading—it's 'misleading higher-order evidence', in the terminology of epistemologists.¹² But since you don't know what your prior is, you (obviously) don't know that. There's much more to discuss about hindsight bias, but we'll hold off until Chapter 7. Upshot: like you and me, Ambiguous Bayesians often exhibit hindsight bias.

This example of an ambiguous update might raise a question. In the above model, I took a higher-order-uncertain prior P and updated it on the true cell of a partition $\{d, \neg d\}$ by going through each world w , and conditioning your prior at w on the true cell of the partition at w . Doing so presupposes that, at each world, you will successfully condition your prior on the evidence.

Question: How can you reliably do this? Since your evidence is partitional, you always know what evidence you received (d or $\neg d$, as the case may be), so you know what to condition on. But the result of conditioning on d depends on your prior. How can you successfully condition your prior on d if you don't know what your prior is? To be concrete: in d_l , how can you successfully move from $\pi_l = (0.15, 0.45, 0.3, 0.1)$ to $\pi_l(\cdot|d) \approx (0.33, 0, 0.67, 0)$, given that your prior (π_l) is unsure whether your prior is π_l or π_h ?

Answer: The same way you can raise your blood pressure without knowing what it is, or alter

¹¹The same is true if your prior had instead been π_h , and/or you had instead learned $\neg d$. If you had π_h and learned d , your prior estimate of 0.51 would jump to a posterior estimate of $0.18(0.45) + 0.82(0.55) = 0.532$. If you had prior π_l and learned $\neg d$, your prior estimate for your prior *in* d would've dropped from 0.49 to $0.82(0.45) + 0.18(0.55) = 0.468$.

¹²E.g. Christensen 2010a; Lasonen-Aarnio 2013, 2014, 2019; Horowitz 2014; Greco 2014, 2019; Littlejohn 2018; Worsnip 2018; Neta 2019.

the sampling propensities of a computer program without knowing what they are (recall §2.3). For example, consider the following program, `randFxn`. It generates two real numbers, and then outputs an integer between 1 and 4 depending on what their product is.

```
randFxn := (x1 = RandomReal[10, 20];
            x2 = RandomReal[5, 10];
            If[x1*x2 < 80, result = 1];
            If[80 < x1*x2 < 118, result = 2];
            If[118 < x1*x2 < 157, result = 3];
            If[157 < x1*x2, result = 4];
            Return[result])
```

What propensity does it have to return a 1? A 3? Obviously you're unsure. But suppose you want to *condition* the sampling propensities so that it only ever generates 1s and 3s, and maintains the same ratio of 1s to 3s. Easy: simply re-run the function if its output is *not* a 1 or 3:¹³

```
randFxn2 := (randFxn;
             While[result != 1 && result != 3, randFxn];
             Return[result])
```

Whatever `randFxn`'s sampling propensity for 1 and for 3 is—write those $R(1)$ and $R(3)$ —`randFxn2` has `randFxn`'s *conditional* sampling propensities, given that it generates a 1 or a 3: $R(1|1 \vee 3) = \frac{R(1)}{R(1 \vee 3)}$. Indeed—as perhaps you've guessed—`randFxn` has sampling propensities that match π_l : it has a 15%-chance to generate a 1, a 45%-chance to generate a 2, a 30%-chance to generate a 3, and a 10%-chance to generate a 4. As a result, `randFxn2` has a 33%-chance to generate a 1 and a 67%-chance to generate a 3, matching $\pi_l(\cdot|d)$.

In short, there's nothing more mysterious about conditioning a prior whose values are uncertain than there is about updating a computer program whose propensities are uncertain. The insistence that we must *know* (exactly) what our beliefs are in order to use or revise them is a holdover from an overly-zealous Cartesian epistemology—one that has doubtful philosophical motivation (Williamson 2000) and little cognitive-scientific credibility (Carruthers 2011). We just aren't like that.

5.4 Constructing Ambiguous Models

Here's what we've seen so far. Though natural, (mere) probabilistic uncertainty is not a good model of ambiguity (§5.2). And once we model your prior with a *variable* probability function, it's mathematically straightforward to model genuine higher-order uncertainty, i.e. probabilistic uncertainty about *your own* probabilities (§5.3). Doing so generates a Bayesian model that has the features—noise, self-doubt, and biases—that we want a model of ambiguity to explain (§5.3.1). So: contrary to widespread belief, higher-order probabilities are neither mathematically incoherent nor behaviorally inert. What we need now is a clearer picture of their mathematical and philosophical structure. This section will explain the structures implicit in Ambiguous-Bayesian models, and

¹³'!=' is code for '≠', and `While[blah, bleh]` repeatedly checks whether `blah` holds and runs `bleh` if it does.

how to use them. §5.5 will justify updating by conditioning in the context of ambiguity. §5.6 will answer conceptual questions about higher-order uncertainty’s stability.

5.4.1 Why Reflection Fails

What is the structure of rational higher-order uncertainty? To begin, it’s natural to think that rational credences should defer to *themselves* in familiar ways. Recall the Reflection principle from Chapter 4: conditional on your future-self having credence x in q , you should (conditionally) have credence x in q : $P_a(q|P^+(q) = x) = x$. This principle is valid on any Standard-Bayesian model, and principles like it pervade probabilistic modeling.¹⁴ They encode a natural way of saying that P defers to P^+ , expecting P^+ ’s estimates to be accurate. It is thus extremely tempting to assume that, when P is uncertain about *itself*, it should obey an analogous principle:¹⁵

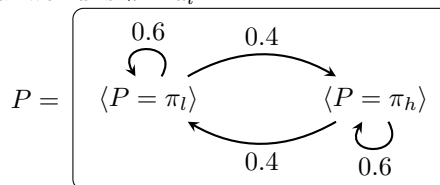
Self-Reflection: Conditional on your current credence in q being x , adopt credence x in q .

$$P_w(q|P(q) = x) = x \quad (\text{Whenever the conditional probability is well-defined})^{16}$$

Welcome to higher-order probability club. Rule number one? Under ambiguity, Self-Reflection fails. The spirit of it is on the right track—in order for credences to *rationally* be higher-order uncertain, they must defer to themselves in some sense. But the implementation is wrong—Self-Reflection is not the proper way to state that higher-order-uncertain credences defer to themselves. To think sensibly about higher-order uncertainty, we need to understand how and why.

Return to our simple spoon example, where the actual world is $\mathbf{a} = d_l$:

$$P = \begin{pmatrix} & d_l & \bar{d}_l & d_h & \bar{d}_h \\ d_l & 0.15 & 0.45 & 0.3 & 0.1 \\ \bar{d}_l & 0.15 & 0.45 & 0.3 & 0.1 \\ d_h & 0.1 & 0.3 & 0.45 & 0.15 \\ \bar{d}_h & 0.1 & 0.3 & 0.45 & 0.15 \end{pmatrix}$$



Self-Reflection fails in this model. The claim that you assign 0.45-credence to d is true at $\{d_l, \bar{d}_l\}$. What is your credence in d , *conditional* on that claim? It’s $P_a(d|P(d) = 0.45)$, i.e. $P_a(d|\{d_l, \bar{d}_l\})$. That equals 0.25.¹⁷ So conditional on you having credence 0.45 in d , you adopt (conditional) credence 0.25 in d : $P_a(d|P(d) = 0.45) = 0.25$. In other words: if you were to learn what your credence was, that would *change* your credence.

When you first see this, it feels like it must be a mistake. But, in fact, this failure of Self-Reflection is as it should be. Once you understand the state you’d be in if P were your credence function, you’ll see that Self-Reflection *should* fail here.

To see why, let’s think through a different instance of Reflection. In this model, you might have credence function π_l , or you might instead have π_h . If you have π_l , you are 60%-confident that

¹⁴E.g. Miller 1966; Lewis 1980; Skyrms 1980; van Fraassen 1984; Gaifman 1988; Samet 1999; Christensen 2010b; Roush 2009, 2016; Mahtani 2017.

¹⁵E.g. Samet 2000; Christensen 2007b; Roush 2016; Rescorla 2023; van Fraassen 2023

¹⁶When stating constraints on conditional probabilities, I’ll assume they’re trivially satisfied when undefined. This generically just makes things simpler to state.

¹⁷By the ratio formula, $P_a(d|\{d_l, \bar{d}_l\}) = \frac{P_a(d_l)}{P_a(\{d_l, \bar{d}_l\})} = \frac{0.15}{0.15+0.45} = 0.25$. By parallel reasoning, conditional on you having credence 0.55 in d , you adopt credence $P_a(d|P(d) = 0.55) = \frac{0.3}{0.3+0.1} = 0.75$.

you have π_l : $\pi_l(P = \pi_l) = 0.6$. Meanwhile, if you have π_h , you have 40%-credence that you have π_l : $\pi_h(P = \pi_l) = 0.4$. You are certain of those facts. Now ask yourself: conditional on you having 0.6-credence that you have π_l , how confident should you be that you have π_l ? According to Self-Reflection, you should be 0.6-confident of that: $P_a(P = \pi_l | P(P = \pi_l) = 0.6) = 0.6$. But that's wrong. You're sure that if you *do* have π_l , you're 60%-confident that you do; and you're sure that if you *don't* have π_l , you're 40%-confident that you do. In other words: you're sure that *you have π_l if and only if you're 60%-confident that you have π_l* .¹⁸ So if you were to learn that you're 60%-confident that you have π_l , you could infer with certainty that you *did* have π_l : $P_a(P = \pi_l | P(P = \pi_l) = 0.6) = 1$.

What's going on? If in fact your credence in $\langle P = \pi_l \rangle$ is 0.6, then *you don't know that*. Indeed, since you know that your credence is 0.6 if and only if you *do* have π_l , your credence is 0.6 *only because* you aren't sure that your credence is 0.6—if you were sure of that fact, that would erase your higher-order uncertainty. Thus if you were to *learn* that you have credence 0.6, that provides you with new information. And in general: new information can change your credences. In this case: learning that you're 0.6 in $\langle P = \pi_l \rangle$ would shift you to the new credence function $P_a(\cdot | P(P = \pi_l) = 0.6) = P_a(\cdot | \{d_l, \bar{d}_l\}) = (0.25, 0.75, 0, 0)$. That's why Self-Reflection fails.

The same reasoning applies to your credence in d . If you have credence 0.45 in d , that's only because you don't realize that you do—because you have 40%-credence that you instead have credence 0.55. Indeed, you know that you have credence 0.45 in d iff you have credence function π_l .¹⁹ So if you learn the former, you can infer the latter. And as we've just seen, learning that you have credence function π_l would change your credence to the new function $(0.25, 0.75, 0, 0)$, which assigns 0.25 credence to d . Hence $P_a(d | P(d) = 0.45) = P_a(d | \{d_l, \bar{d}_l\}) = 0.25$.

The reasoning generalizes completely. This failure of Self-Reflection is not an accident. In *any* model with higher-order uncertainty, Self-Reflection fails:

Fact 5.4.1. In any finite, self-aware frame, if P_w is ambiguous, then there is a q and x such that $P_w(q | P(q) = x) \neq x$.²⁰

The issue is that when P_w is higher-order uncertain, learning of its own values is news. Thus we can always find some claim q such that learning its credence in *that* claim changes its credence in that claim. Put the other way: the only models that validate Self-Reflection are those in which P is clear. In that case, Self-Reflection is trivially satisfied.²¹ This underlies Fact 4.4.4 from last chapter, which said that you can only reflect a probability function P^+ if you are certain that it is clear.

So, under ambiguity, Self-Reflection fails. What replaces it?

5.4.2 Factoring Frames

When you learn what credence you had in q , the reason that you shouldn't simply adopt that credence is that—under ambiguity—this may well have provided new, relevant information. In

¹⁸Formally: $\langle P = \pi_l \rangle \leftrightarrow \langle P(P = \pi_l) = 0.6 \rangle$ is true at each world in the frame, so $P_a(\langle P = \pi_l \rangle \leftrightarrow \langle P(P = \pi_l) = 0.6 \rangle) = 1$.

¹⁹Formally, $\langle P(d) = 0.45 \rangle \leftrightarrow \langle P = \pi_l \rangle$ is true at each world.

²⁰See Problem [TODO] for proof. See Gaifman 1988; Samet 2000; Dorst 2020a for similar results.

²¹If P_w is clear and self-aware, then if $P_w(P(q) = x) > 0$, $P_x(P(q) = x) = 1$. Thus the only value of x for which $P_w(\cdot | P(q) = x)$ is well-defined is the one it is already certain it has.

other words: the reason you don't completely defer to your unconditional credences P is that they might be distorted by higher-order doubts.

Consider the spoon case. If in fact your credence in d is 0.45, that's in part because you leave open that it might be higher (0.55). This higher-order uncertainty 'pulls' your credence in d up, to 0.45. Once you learn that you *do* have credence 0.45, that removes the tempering effects of your higher-order doubts, making you more confident in your hunch that I don't own a dozen spoons: your credence in d drops to 0.25.

This suggests that if we were to hypothetically *remove* P 's higher-order doubts, creating a (higher-order) *informed* version of your credence function, then you would completely defer to that informed credence function (Elga 2013; Stalnaker 2019; Dorst et al. 2021).

How can we do that? Consider the question, 'What credence function do you have?' The possible answers induce a partition—in the spoon case, $\{\langle P = \pi_l \rangle, \langle P = \pi_h \rangle\}$. Now imagine we took your credence function P and informed it of the true answer to that question—at each world w , conditioned P_w on the true claim of the form $\langle P = \pi_i \rangle$. This would create a new credence function, \hat{P} , which is the result of removing P 's higher-order doubts. Precisely let your **informed credence function**, \hat{P} , be the variable probability function induced by performing this operation. Notice that since $\langle P = \pi_i \rangle$ is true at w iff $\pi_i = P_w$, we can define this by saying that for every world w , $\hat{P}_w(\cdot) := P_w(\cdot | P = P_w)$.²²

I know, I know—that looks like nonsense. But if you keep track of the difference between the constant P_w and the variable P , it makes perfect sense. Consider what it means in the spoon case. We go through each world, and condition the prior P on the true cell of $\{\langle P = \pi_l \rangle, \langle P = \pi_h \rangle\}$. At the first two worlds, we condition π_l on $\langle P = \pi_l \rangle$, i.e. on $\{d_l, \bar{d}_l\}$; and at the second two, we condition π_h on $\langle P = \pi_h \rangle$, i.e. on $\{d_h, \bar{d}_h\}$. Thus:

$$\hat{P} = \left(\begin{array}{c|cccc} & d_l & \bar{d}_l & d_h & \bar{d}_h \\ \hline d_l & 0.25 & 0.75 & 0 & 0 \\ \bar{d}_l & 0.25 & 0.75 & 0 & 0 \\ d_h & 0 & 0 & 0.75 & 0.25 \\ \bar{d}_h & 0 & 0 & 0.75 & 0.25 \end{array} \right) \quad \hat{P} = \boxed{\begin{array}{cc} \begin{array}{c} 1 \\ \curvearrowright \\ \langle P = \pi_l \rangle \end{array} & \langle P = \pi_h \rangle \\ & \begin{array}{c} \langle P = \pi_h \rangle \\ \curvearrowright \\ 1 \end{array} \end{array}}$$

As you can see from the diagram on the right, this 'cuts' the higher-order-uncertainty arrows between the different cells of the partition answering the question, 'What credence function do you have?'—making the informed credence function \hat{P} look like a Standard Bayesian model.

Indeed—as we hoped— \hat{P} reflects P . We can check this by hand for each proposition; for instance, $P_a(d | \hat{P}(d) = 0.25) = P_a(d | \{d_l, \bar{d}_l\}) = 0.25$, and $P_a(d | \hat{P}(d) = 0.75) = P_a(d | \{d_h, \bar{d}_h\}) = 0.75$. We can also check it by seeing that when P_w conditions on the *entire* \hat{P} function being a certain value, P_w always adopts that value. For example, $P_a(\cdot | \hat{P} = (0, 0, 0.75, 0.25)) = P_a(\cdot | \{d_h, \bar{d}_h\}) = (0, 0, 0.75, 0.25)$. Finally, we can check it by seeing that P_w 's *expectation* of \hat{P} always equals P_w 's own credence function. For example, $\mathbb{E}_{P_a}(\hat{P}(d)) = P_a(\hat{P}(d) = 0.25)(0.25) + P_a(\hat{P}(d) = 0.75)(0.75) = 0.6(0.25) + 0.4(0.75) = 0.45$.

We can formalize this constraint as follows²³:

²²Note that in any self-aware probability frame, \hat{P} is well-defined at all worlds.

²³Informed Reflection is equivalent to what Elga 2013 calls 'New (Rational) Reflection', on analogy with the

Informed Reflection: Conditional on having *informed* credence x in q , adopt credence x in q .

Formally, $P_w(q|\hat{P}(q) = x) = x$.

Equivalently: $P_w(\cdot|\hat{P} = \delta) = \delta$.

Equivalently: $\mathbb{E}_{P_w}(\hat{P}(q)) = P_w(q)$.²⁴

Informed Reflection is not trivial—there are models of higher-order uncertainty that violate it (Lasonen-Aarnio 2015). The constraint it imposes on a frame is that even if two worlds w and v have different probabilities, $P_w \neq P_v$, they agree *conditional on* the informed credence function having any given value $\langle \hat{P} = \delta_i \rangle$. Thus any differences in credences between P_w and P_v are parasitic on differences in how likely they think the various informed-probability-classes $\langle \hat{P} = \delta_i \rangle$ are.

Moreover, it turns out that these informed-probability classes are *the same* as the *uninformed*-probability classes—the two questions, ‘What is your credence function?’ and ‘What is your informed credence function?’ induce the same partition. We can see this in the spoon model: the two uninformed classes are $\langle P = \pi_l \rangle$ and $\langle P = \pi_h \rangle$, i.e. $\{d_l, \bar{d}_l\}$ and $\{d_h, \bar{d}_h\}$. Meanwhile, the two informed classes are $\langle \hat{P} = \pi_l(\cdot|P = \pi_l) \rangle$ and $\langle \hat{P} = \pi_h(\cdot|P = \pi_h) \rangle$, i.e. $\{d_l, \bar{d}_l\}$ and $\{d_h, \bar{d}_h\}$. Once again, these are different names for the same propositions. Likewise more generally. After all, if you know that the informed credence function is \hat{P}_w , you can figure out what the uninformed credence function is by seeing which one it assigns probability 1 to. And if you know that the uninformed credence function is P_w , you can figure out what the informed credence function—namely, $P_w(\cdot|P = P_w)$.²⁵

As a result, Informed Reflection is also equivalent to the constraint that if two worlds w and v have different probabilities, $P_w \neq P_v$, then still agree *conditional on the uninformed* credence function having any given value $\langle P = \pi \rangle$. Thus any differences in credences between P_w and P_v are parasitic on differences in how likely they think the *uninformed*-probability-classes $\langle P = \pi_i \rangle$ are.

This is a very useful fact. It means that any probability frame that validates Informed Reflection can be factored into a ‘first-order’ and a ‘higher-order’ component. The ‘first-order’ component says what your opinions would be without any higher-order doubts—your informed probabilities or, as I sometimes call them, your ‘hunches’—while the ‘higher-order’ component says what those higher-order doubts are, for each possible hunch. Together, these components generate your ‘all-doubts-considered’ opinions by taking your higher-order expectation of your informed probabilities.

Let’s see how this works in the spoon frame. The two possible informed distributions are $\delta_l = (0.25, 0.75, 0, 0)$ and $\delta_h = (0, 0, 0.75, 0.25)$. Let’s order these (δ_l, δ_h) . The two possible higher-order distributions over these informed credences are $(0.6, 0.4)$ —i.e. the function that assigns 0.6 to $\langle P = \delta_l \rangle$ and 0.4 to $\langle P = \delta_h \rangle$ —and $(0.4, 0.6)$.

We can represent this with a more-rigorous version of our Markov diagrams. To represent the first-order component, we divide possibilities into partition-cells that share the same informed

‘New Principal Principle’ from Hall 1994; Lewis 1994. New Reflection says that when you defer to an expert P^+ , upon learning what their credences are, you should adopt the credences they *would* adopt, upon learning what you learned: $P_a(\cdot|P^+ = \pi) = \pi(\cdot|P^+ = \pi)$. When the expert’s opinions are clear, this is equivalent to Reflection; but unlike Reflection, it allows ambiguity. See Stalnaker 2019 and Dorst et al. 2021 for the equivalence.

²⁴These three constraints are not *in general* equivalent, if we substitute another function P' in for \hat{P} . But they are equivalent whenever P' is clear—which \hat{P} is by definition. [\[add proof?\]](#)

²⁵Formally in any finite, self-aware probability frame, $\langle P = \pi \rangle = \langle \hat{P} = \pi(\cdot|P = \pi) \rangle$. *Proof:* If $w \in \langle P = \pi \rangle$, then $P_w = \pi$ and so by definition $\hat{P}_w = \pi(\cdot|P = \pi)$ (which is well-defined, by self-awareness), i.e. $w \in \langle \hat{P} = \pi(\cdot|P = \pi) \rangle$. And if $w \in \langle \hat{P} = \pi(\cdot|P = \pi) \rangle$, then $\hat{P}_w = \pi(\cdot|P = \pi)$, hence by definition of \hat{P}_w , we know $P_w = \pi$.

probabilities (and, therefore, share the same uninformed probabilities), writing those probabilities as truncated vectors that omit the zeros. To represent the higher-order component, we draw labeled-arrows *between* cells to represent the higher-order distributions η_i associated with each cell. Here it is for our spoon frame:

$$0.6 \curvearrowleft \left(\begin{array}{cc} d_l & \bar{d}_l \\ 0.25 & 0.75 \end{array} \right) \begin{array}{c} \xrightarrow{0.4} \\ \xleftarrow{0.4} \end{array} \left(\begin{array}{cc} d_h & \bar{d}_h \\ 0.75 & 0.25 \end{array} \right) \curvearrowright 0.6$$

We then recover the *uninformed* probability P_w at each world w by taking the higher-order η_w associated with w 's partition-cell, and taking it's expectation of \hat{P} . We can do this all at once, averaging the vectors of each informed distribution. Thus the probability function associated with the first cell is $0.6(0.25, 0.75, 0, 0) + 0.4(0, 0, 0.75, 0.25) = (0.15, 0.45, 0, 0) + (0, 0, 0.3, 0.1) = (0.15, 0.45, 0.4, 0.1)$ —which is π_l , as we wanted.

(If you don't know linear algebra, ignore this paragraph. If you do, it's often more convenient to represent factorable frames with *two* matrices: one encoding its higher-order distributions (left), and the other encoding its informed distributions (right):

$$\left(\begin{array}{c|cc} & P = \pi_l & P = \pi_h \\ \hline P = \pi_l & 0.6 & 0.4 \\ P = \pi_h & 0.4 & 0.6 \end{array} \right) \quad \left(\begin{array}{c|cccc} & d_l & \bar{d}_l & d_h & \bar{d}_h \\ \hline \hat{\pi}_l & 0.25 & 0.75 & 0 & 0 \\ \hat{\pi}_h & 0 & 0 & 0.75 & 0.25 \end{array} \right)$$

If we duplicate the left matrix's rows to match the number of worlds in each $\langle P = \pi_i \rangle$ -class (what the left-multiple matrix of 1s and 0s does), and then multiply it by the right matrix, we recover our original frame:

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} \begin{pmatrix} 0.25 & 0.75 & 0 & 0 \\ 0 & 0 & 0.75 & 0.25 \end{pmatrix} = \begin{pmatrix} 0.15 & 0.45 & 0.3 & 0.1 \\ 0.15 & 0.45 & 0.3 & 0.1 \\ 0.1 & 0.3 & 0.45 & 0.15 \\ 0.1 & 0.3 & 0.45 & 0.15 \end{pmatrix} = P$$

Of course, that is monstrous to look at; but it's useful for manipulating these frames in a computer.)

This 'factorization' strategy will work on any frame that validates Informed Reflection. For simplicity, I'll state this for a **synchronic frame** that contains just a prior P (no posterior P^+), and restrict to frames with a finite number of **candidates** δ_i for the informed credence function:

Theorem 5.4.2 (Factorization Theorem). A self-aware synchronic frame (W, \mathbf{a}, P) with finitely many candidates validates Informed Reflection iff it can be factored into two components:

1. A set of disjoint probability spaces $\{(W_1, \delta_1), \dots, (W_n, \delta_n)\}$ where $W = \bigcup_i W_i$, and for any $w_i \in W_i$, we define $\hat{P}_{w_i} := \delta_i$; and
2. For each space, a (unique²⁶) 'higher-order' distribution η_i defined over the partition $\{W_1, \dots, W_n\}$;

²⁶It's important (and easy to forget) that the higher-order distributions must be distinct. If $\eta_i = \eta_j$, then they'll give rise to the same overall distribution P , meaning that learning what P is *won't* determine what \hat{P} is, as it must.

Such that for any world $w_j \in W_j$ and $q \subseteq W$: $P_{w_j}(q) = \mathbb{E}_{\eta_j}(\hat{P}(q))$, i.e. equals $\sum_{W_i} \eta_j(W_i) \cdot \delta_i(q)$.

Thus I will say a probability frame is **factorable** if P validates Informed Reflection.

This theorem makes ambiguous frames much easier to work with. The left-to-right chops down the degrees of freedom and ensures that we can represent them with Markov diagrams, as above. The right-to-left direction gives us a recipe for generating frames that validate Informed Reflection:

1. Begin with a set of clear distributions $\{\delta_1, \dots, \delta_n\}$ over their own (disjoint) sample spaces W_i . (These can be any familiar distribution you like—Binomial, Gaussian, Beta, etc.)
2. Generate unique higher-order distribution η_i for each δ_i , defined over the partition induced by the question ‘Which (informed) distribution do you have?’, i.e. $\{W_1, \dots, W_n\}$.
3. Set your *uninformed* credence P_{w_j} at each $w_j \in W_j$ to the η_j -weighted average of the δ_i .

This recipe will allow us to start with familiar distributions as the possible informed (and so clear) distributions, and then add ambiguity on top. For those who want to see the Factorization Theorem in action, the Appendix (§5.8.2) shows how to use it to construct a toy model of ambiguity in Ellsberg’s (1961) famous case of urns of unknown composition. Though it’s not my focus, there I discuss the prospects for using Ambiguous Bayesianism to explain ‘ambiguity aversion’ in the sense economists have studied widely.²⁷

Moreover, we’ll see in Chapter 6 that Informed Reflection is one component of plausible rationality constraints on frames—they simply further constrain way the higher-order distributions η_i must relate to each other. Thus we can also use this recipe to generate frames that meet more restrictive conditions. So it would be both natural and convenient to assume that (rational) higher-order uncertain probabilities obey Informed Reflection. Would it also be well-motivated?

Yes. Informed Reflection follows from a variant of the No Foregone Conclusions principle discussed in Chapter 4, which bears affinity to famous Sure-Thing Principle (Savage 1954). Precisely:

No Foregone Questions: If you are certain that learning the true answer to a question would lead you to have a high estimate for X , you should already have a high estimate for X .

Formally: if there is a partition \mathcal{Q} such that $P_w(\mathbb{E}_{P^Q}(X) \geq x) = 1$, then $\mathbb{E}_{P_w}(X) \geq x$.²⁸

This principle is not undeniable. But it is well-motivated, and is a minimal component of the ‘value of rationality’ constraint we’ll discuss in Chapter 6. It implies Informed Reflection:

Fact 5.4.3. If a synchronic frame (W, \mathbf{a}, P) validates No Foregone Questions, then P is factorable.

Thus, for the sake of this book, I will restrict attention to probability frames in which P is self-aware and factorable. When I talk of ‘Ambiguous Bayesian’ models, I mean those which modify Standard Bayesianism by swapping out clarity for factorability:

Ambiguous Bayesianism: A frame (W, \mathbf{a}, P, P^+) is *Ambiguously-Bayesian* iff:

- i) P is self-aware and factorable; and
- ii) There is a partition \mathcal{Q} such that your posterior is the result of conditioning on the true cell of \mathcal{Q} : for all w and q , $P_w^+(q) = P_w(q|\mathcal{Q}_w)$.

²⁷E.g. Einhorn and Hogarth 1985; Heath et al. 1991; Camerer and Weber 1992; Klibanoff et al. 2005; Halevy 2007; Al-Najjar and Weinstein 2009; Etner et al. 2012; Baliga et al. 2013; Machina and Siniscalchi 2014; ?; Bradley and Steele 2016; Li et al. 2018. [BuchakXX]

²⁸Where P^Q is the result of conditioning P on the true cell of \mathcal{Q} : $P_w^Q := P_w(\cdot|\mathcal{Q}_w)$.

Oopsy. H.6–7 show that the converse fails. Weaken?

Does this imply that P^+ is factorable? H.8c suggests yes, but proof eludes me

In Chapter 6 we will add further rationality constraints. But any update that satisfies Ambiguous Bayesianism is already ‘sensible’ in a minimal sense, and obviously retains many of the explanatory virtues of Standard Bayesian models.

5.5 Conditioning Maximizes Expected Accuracy

Ambiguous Bayesianism is a natural, conservative refinement of Standard Bayesianism to allow for ambiguity. We have talked through why self-awareness and factorability are plausible minimal-rationality constraints on the prior P .

But what about the constraint that P^+ is the result of conditioning P on the true cell of a partition? We saw in Chapter 4 that *under clarity*, minimally-rational updating is partitional updating. This was for two reasons: (1) knowing your posterior allows you to be certain of what information you received; and (2) under such conditions, No Foregone Conclusions forces you to be sure that you’ll update by conditioning. But under ambiguity, you do not know your posterior, so the above reasoning fails. So why should we assume that, under ambiguity, the rational updates from P to P^+ are still the result of conditioning on a partition? Granted, it is a tractable and conservative assumption to make. But it is motivated?

There are two steps to answering this question. (i) First, under ambiguity is there any reason why new evidence must come in as a partition? (ii) Second, given that new evidence has come in as a partition, why is the rational way to update on it to *condition* on it, as opposed to do something else? Formal epistemologists and some economists have been debating these questions, under various guises, for the past couple decades.²⁹

The main focus of these discussions has been on (i) whether new evidence must be partitional—and if not, how to rationally respond to it. I believe this is because it’s been assumed that (ii) is settled in the affirmative, due to a result from Greaves and Wallace 2006 that when evidence is partitional, conditioning maximizes expected accuracy on any sensible (‘strictly proper’) way of measuring the accuracy of a probability function. But that result assumes that you start with a *clear* prior—the prior is not an object of uncertainty, so there is no possibility of uncertainty about *which* prior you will condition on the true answer to the question. So both (i) and (ii) are open questions.

As mentioned, I won’t wade into the deep waters surrounding (i) whether evidence can be non-partitional under ambiguity. I’ve personally been convinced by Williamson 2000 that it can be; but it won’t matter if you disagree. All updates in this book will be partitional. This is for three reasons. First, we do not have agreed-upon theories for when and how rational people should update on non-partitional evidence, let alone how we could induce such updates in the lab. Partitional updates are much easier: simply tell people the answer to a clear question. Second, there are genuine philosophical and cognitive-scientific questions of how people *could* reliably update on non-partitional evidence; as we’ve seen above, the answer is much simpler for partitional evidence.

²⁹See e.g. Gaifman 1988; Geanakoplos 1989; Samet 1999; Hild 1998; Williamson 2000, 2014, 2019; Weisberg 2007; Briggs 2009b; Christensen 2010b; Elga 2013; Bronfman 2014, 2015; Schoenfield 2017b; Salow 2018; Ahmed and Salow 2019; Gallow 2019b, 2021; Dorst 2020a, 2023b; Dorst et al. 2021; Das 2022, 2023; Isaacs and Russell 2022; Zendejas Medina 2022; Schultheis 2023; Zhang and Meehan 2025.

And third, I don't want any of the results of this book to hinge on controversial choice points about how to handle non-partitional updates. This is an important point. Many who argue that we should respond to non-partitional evidence in non-standard ways are motivated by the goal of recovering principles similar to Reflection, Martingale, or the avoidance of bias. By showing that these principles inevitably fail even under *partitional* updating once *priors* are ambiguous, I'll show that this debate is missing an important step. If you want to defend those principles, you cannot simply address how to respond to non-partitional evidence; you must refute the arguments (e.g. in Ch. 2) that we ever rationally fail to know what our priors are.

But what about (ii)? Suppose your priors P are ambiguous and new partitional evidence \mathcal{Q} comes in through the eyeballs. Why is the rational thing to update by *conditioning* P on the true answer to \mathcal{Q} ? Because that is the result of making and implementing an update-plan that maximizes expected accuracy according to your prior. We can generalize Greaves and Wallace's (2006) result to the case where your priors are ambiguous. Here's how.

Suppose we have a frame (W, \mathbf{a}, P, P^+) . Suppose we have a strictly-proper accuracy measure A : given a probability function π and world w , $A(\pi, w)$ is a real number measuring the accuracy of π at w . To say that A is strictly proper is to say that each probability function π expects itself to be more accurate than any other constant distribution δ : for all $\delta \neq \pi$: $\mathbb{E}_\pi(A(\pi)) > \mathbb{E}_\pi(A(\delta))$.³⁰ Assume for simplicity that the frame is *regular* over the partition \mathcal{Q} , in the sense that each world assigns positive probability to each answer to \mathcal{Q} : for all w and $q \in \mathcal{Q}$, $P_w(q) > 0$. Let an **update plan** $u : \mathcal{Q} \rightarrow \Delta(W)$ be a function from cells of the partition $\mathcal{Q}_w \in \mathcal{Q}$ to probability functions $u(\mathcal{Q}_w)$ over W . The P_w -expected accuracy of an update plan u is the expected accuracy of the probability function that results from implementing the plan: $\mathbb{E}_{P_w}(A(u)) = \sum_v P_w(v) \cdot A(u(\mathcal{Q}_v), v)$.

Given P_w , let u_w^c be the update plan to condition P_w on the true cell of \mathcal{Q} , that is, for any $q \in \mathcal{Q}$, $u_w^c(q) = P_w(\cdot|q)$. Then Greaves and Wallace's result is that u_w^c uniquely maximizes expected accuracy according to P_w :

Theorem 5.5.1 (Greaves and Wallace 2006). If P_w is regular over \mathcal{Q} and A is strictly proper, then for all $u \neq u_w^c$, $\mathbb{E}_{P_w}(A(u_w^c)) > \mathbb{E}_{P_w}(A(u))$.

In other words: if you go through each world w and ask the prior P_w at that world how *it* thinks its best to react to the information provided by \mathcal{Q} , the answer is to update P_w itself by conditioning it on the true answer to \mathcal{Q} . Notice that if P_w leaves open v , and $P_v \neq P_w$, then P_w is endorsing an update plan that 'projects itself out' onto v : P_w wants, at world v , to update to $P_w(\cdot|\mathcal{Q}_v)$, *not* to $P_v(\cdot|\mathcal{Q}_v)$. Meanwhile, P_v wants the reverse: it wants, at w , to update to $P_v(\cdot|\mathcal{Q}_w)$. So under ambiguity, the update-plans that look best according to the various priors you might have (P_w, P_v, \dots) can be at cross-purposes.

What will happen if each world w *implements* its preferred update plan? Then at each world w , you will take the update plan preferred at w , u_w^c , and update it on the true answer to \mathcal{Q} at w , \mathcal{Q}_w . Thus each world w will update from P_w to $u_w^c(\cdot|\mathcal{Q}_w)$, i.e. to $P_w(\cdot|\mathcal{Q}_w)$. The result will be to condition your prior *whatever it is* on the true answer to \mathcal{Q} , whatever *it* is.

³⁰It's important that δ be a constant distribution, rather than a variable; no probability function expects itself to be more accurate than the omniscient function $\mathbb{1}$ that assigns 1 to all truths and 0 to all falsehoods. Strict propriety is the standard operating procedure in the accuracy literature. For discussions, see Schervish 1989; Gibbard 2008; Joyce 2009; Predd et al. 2009; Levinstein 2017; Campbell-Moore and Levinstein 2020; Williams and Pettigrew 2023.

Formally, let a *variable* update plan U be a function from worlds w to update plans U_w . Given a strictly-proper accuracy measure A and evidence partition \mathcal{Q} , say that a frame, $(W, \mathfrak{a}, P, P^+)$ is *maximizes expected accuracy with respect to \mathcal{Q}* iff there is a variable update plan U such that (1) for each w , U_w is the update-plan that maximizes expected A -accuracy according to P_w , and (2) for each w , the update from P_w to P_w^+ implements U_w , i.e. $P_w^+ = U_w(\mathcal{Q}_w)$. Then:

Corollary 5.5.2 (Conditioning Maximizes Expected Accuracy). If A is strictly proper, P is regular over \mathcal{Q} , and the frame $(W, \mathfrak{a}, P, P^+)$ maximizes expected accuracy with respect to \mathcal{Q} , then P^+ is the result of conditioning P on \mathcal{Q} : for all w , $P_w^+(\cdot) = P_w(\cdot|\mathcal{Q}_w)$.

In short: even if your prior is ambiguous, when new evidence comes as a partition—i.e. as a complete answer to a clear question—then the rational response to it is to update by conditioning.³¹

5.6 Is Higher-Order Uncertainty Stable?

We’ve seen how to model genuine higher-order uncertainty in a mathematically coherent way, pinpointed its factorable structure, and shown how to update such opinions. But in every discussion of higher-order uncertainty, a question eventually crops up: How is higher-order uncertainty *stable*?

Sometimes this question is framed as a synchronic problem: When I’m uncertain what my own probabilities are, why can’t I do (something like) take my *expectation* of them, and use *that* as my probability distribution? Other times it’s framed as a diachronic one: Why can’t I just force myself to *act*, and then infer what my probabilities are from my action? Let’s take these in turn.

5.6.1 Current Stability

You’re in the club, now. Remember rule number one? Here’s rule number two: under ambiguity, your expectation of your credences does *not*, in general, equal your credences. The principle that states such an equality is a synchronic version of the ‘Martingale’ principle we’ll discuss more later:

Self-Martingale: Your expectation for your credence in q equals your credence in q .

$$\text{Formally: } \mathbb{E}_{P_w}(P(q)) = P_w(q)$$

Self-Martingale follows from Self-Reflection by simply averaging over the possible values of $P(q)$. By definition, $\mathbb{E}_{P_w}(P(q))$ is a weighted average of the possible values of $P(q)$:

$$\mathbb{E}_{P_w}(P(q)) = \sum_x P_w(P(q) = x) \cdot x$$

³¹In an apparent paradox, we’ll see in Ch. 6 that if P has pathological degrees of self-doubt, sometimes conditioning on the true answer to a question \mathcal{Q} will result in a posterior P^+ that is less expectedly-accurate than the prior P . What’s going on? Each P_w is choosing an update-plan u_w^c that would improve on P_w ’s expected accuracy *if* u_w^c were implemented at all worlds. But when P_w and P_v disagree profoundly, they may be working at cross-purposes in a way that leads to an overall decrease in accuracy when each implements its own plan. Structurally, the situation is analogous to a prisoner’s dilemma: strategies that are optimal relative to each P_w can lead to a collective implementation that is worse than doing nothing. We’ll see in Ch. 6 how if P is constrained in the right way (to be ‘Valuable’), this will never happen.

Alternative: each P_w only cares about the accuracy of its plan if \mathcal{Q}_w ; Gallow, learning-and-value-change-style. That’ll give the same result, without the apparent naivety of P_w in thinking about what’ll happen if it’s at v with $P_v \neq P_w$. Which is better?

cite/discuss Savage §4.2

But the possible values of $P(q)$ form a partition, so by the law of total probability, $P(q)$ is an average of what it would be conditional on these various values:

$$P_w(q) = \sum_x P_x(P(q) = x) \cdot P_w(q \mid P(q) = x)$$

By Self-Reflection, the last term always equals x , so the two sums are equal.

Less obviously, Self-Martingale itself implies clarity, and hence implies Self-Reflection:

Theorem 5.6.1 (Gaifman 1988; Samet 2000). If a self-aware synchronic frame (W, \mathfrak{a}, P) validates Self-Martingale, P is clear.

In other words: the only case in which your expectations of your credences are always certain to match your credences is the trivial one in which you are certain of what your credences are.³² The proof requires some footwork³³, but the basic reason can be seen intuitively in two ways.

First, notice that Self-Martingale asserts that your estimates about a certain quantity—namely, your credence in q , $P(q)$ —are always accurate. But, in general, if you can be uncertain about a quantity X —the number of spoons I own, the proportion of chores you do—then your estimate about X can sometimes be inaccurate. After all, uncertainty inevitably opens up the possibility of misleading evidence.

For instance, suppose you are uncertain how many spoons I own, X , and you estimate $\mathbb{E}_{P_a}(X) = 15$. Perhaps your initial estimate is inaccurate—maybe I own 21 or 14 spoons. But even if it’s accurate (I own 15), learning new evidence can throw you off. By definition—since your estimate is an average of the various possible values you leave open—if $\mathbb{E}_{P_a}(X) = 15$ and you are uncertain of the value of X , then you must leave open possible values of X that are both higher and lower than 15.³⁴ So suppose you are about to learn the true answer to the question, ‘Does Kevin own at least 15 spoons?’, conditioning on the true cell of $\{\langle X < 15 \rangle, \langle X \geq 15 \rangle\}$. If you learn the former, that’ll lower your estimate below 15, and if the latter, that’ll raise your estimate above 15.³⁵ So even if your estimates begin perfectly accurate, uncertainty guarantees that they might become inaccurate.

Second, Self-Martingale must fail when your credence is on the edge of the range of possible credences you leave open. To see why, consider the spoon frame. You are sure that your credence in d might be either 0.45 or 0.55. Consider the possibility $w = d_l$ where it’s at the lower end of this range: $P_w(d) = 0.45$. At that possibility, you are sure that your credence in d is no lower than 0.45, and you leave open that it might be higher—thus your *expectation* for your credence in d must be higher

³²But wait. Credences are expectations of truth values. And doesn’t the ‘tower rule’ say that your expectation of your expectation always equals your expectation? No—at least not in the sense at issue. The quick way to see this is that the tower rule—i.e. the law of total expectation—is a theorem of probability, and hence is valid on every probability frame. Yet Self-Martingale fails on our spoon frame. The difference is that Self-Martingale uses a variable P for ‘your credence’, while the tower law uses a constant P_w —see §5.8.5 for discussion.

³³See Dorst 2019 for an elementary proof. For those familiar with Markov chains, you can quickly see the idea. Self-Martingale asserts that each probability function π in the frame is a stationary distribution with respect to the whole frame: $\pi P = \pi$. We can partition the frame into communicating classes—where no world in any class assigns positive probability to any other class—and each such class has a unique stationary distribution. Thus the only way to satisfy Self-Martingale is for all the P_w in each class to be the same such stationary distribution; hence clarity.

³⁴If $\max_x (P_a(X = x) > 0) = h$ and $\min_x (P_a(X = x) > 0) = l$, then $l < \mathbb{E}_{P_a}(X) < h$.

³⁵If $P_a(X < x) > 0$ and $P_a(X \geq 0) > 0$, then $\mathbb{E}_{P_a}(X \mid X < x) < \mathbb{E}_{P_a}(X)$ and $\mathbb{E}_{P_a}(X \mid X \geq x) > \mathbb{E}_{P_a}(X)$.

than 0.45. In this case, it's $\mathbb{E}_{P_w}(P(d)) = 0.6(0.45) + 0.4(0.55) = 0.49$. Similarly, if your credence in d is 0.55, then your expectation for your credence is lower: $\mathbb{E}_{P_{a_h}}(P(d)) = 0.4(0.45) + 0.6(0.55) = 0.51$. In both cases, your higher-order uncertainty leads your expectation of your credence to be pulled toward the other possible values you leave open, and hence to deviate from your actual credence.

The point generalizes. You have ambiguity only if you leave open a range of credences you might have—but Self-Martingale would force the possibilities at the extreme edge of that range to be pulled inward. The only stable situation, given Self-Martingale, is to have no ambiguity at all. My hunch is that this largely explains the feeling that higher-order uncertainty should be unstable: people presuppose that Self-Martingale holds, realize that it would destabilize higher-order uncertainty, and conclude that higher-order uncertainty is unstable.

The correct conclusion, instead, is that Self-Martingale fails under ambiguity. *And it should.* The reason Self-Martingale seemed plausible is that encodes the sort of deference to your own opinions captured by Self-Reflection. But, as we've seen, Self-Reflection is a mistake—for it fails to account for the fact that, under ambiguity, *learning* what your rational credence in q was can make it rational to *change* your credence in q : $P_a(q|P(q) = x) \neq x$. Thus when we use total probability to decompose $P_w(q)$ as we did above, many of the terms in the sum do *not* equal x —so there's no reason, in general, for $P_w(q)$ to equal $\mathbb{E}_{P_w}(P(q))$. (Though it may, by accident, for particular propositions.)

Of course, there *are* true principles in the vicinity of Self-Martingale. As we've seen, in any factorable frame, your credence in q will always equal your expectation of your *informed* credence in q : $P_w(q) = \mathbb{E}_{P_w}(\hat{P}(q))$. Why can't you use your expectation of the informed credence to figure out what your credence is?

Consider an analogous question. Probabilities are expectations of truth values: if $\mathbb{1}$ is the omniscience probability function that assigns 1 to all truths and 0 to all falsehoods, then $P_w(q) = \mathbb{E}_{P_w}(\mathbb{1}(q))$. Why can't you use your expectation of the truth-value to figure out what your credence is? That's obvious. By definition, your expectation of the truth value *is* your credence—so if you're unsure what the latter is, you'll thereby be unsure what the former is. Likewise: your expectation of your informed credence *is* your credence—so if you're unsure what the latter is, you'll thereby be unsure what the former is. Expectations are simply facts about your credence function—under ambiguity, they're just as uncertain as facts about your credences.

What about your expectations of *your expectations* of your informed credences? The problem iterates—in general, if your expectations of a variable are uncertain, so too are your expectations of your expectations of that variable. We can see this in our spoon frame, repeated here:

$$\left(\begin{array}{c|cc} & \hat{P}(d) = 0.25 & \hat{P}(d) = 0.75 \\ \hline \hat{P}(d) = 0.25 & 0.6 & 0.4 \\ \hat{P}(d) = 0.75 & 0.4 & 0.6 \end{array} \right)$$

If in fact your informed credence is 0.25, then in fact your expectation of your informed credence is $0.6(0.25) + 0.4(0.75) = 0.45$. But you don't know that. For you leave open that your informed credence is 0.75, in which case your expectation of your informed credence is $0.4(0.25) + 0.6(0.75) = 0.55$. Thus if in fact your informed credence is 0.25, your expectation of *your expectation* of your

informed credence is $0.6(0.45) + 0.4(0.55) = 0.49$. But—again—you don’t know that. For you leave open that your informed credence is 0.75, in which case your expectation of *your expectation* of your informed credence is $0.4(0.45) + 0.6(0.55) = 0.51$. And so on. Under ambiguity, higher-order uncertainty goes ‘all the way up’.³⁶

Indeed, that’s a theorem. In general, if you have second-order uncertainty (you’re unsure what your credence in q is) then you must have third-order uncertainty (you’re unsure what your expected credence in q is), and so on. More precisely, your probabilities equal your expectations of truth values: $P_w(q) = \mathbb{E}_{P_w}(\mathbb{1}_q)$. So let’s say you have *first-order* uncertainty if, for some q , you’re unsure of the value of $\mathbb{1}_q$. You have *second-order* uncertainty if, for some q , you’re unsure of your expectation of $\mathbb{1}_q$, i.e. unsure of $\mathbb{E}_P(\mathbb{1}_q)$, i.e. unsure of your credence in q , $P(q)$. You have *third-order* uncertainty if, for some q , you’re unsure of your expectation of your expectation of $\mathbb{1}_q$, i.e. unsure of $\mathbb{E}_P(\mathbb{E}_P(\mathbb{1}_q))$, i.e. unsure of your expectation of your credence in q , $\mathbb{E}_P(P(q))$. And so on.³⁷ Generally, if you have second-order uncertainty, you must have n th-order uncertainty for all n .

I believe this is true in full generality; but I’ll just prove a special case where we constrain your higher-order uncertainty to have an easy-to-work-with structure. One form of higher-order uncertainty is being uncertain *what you are certain of*. That is: being unsure whether or not you assign credence 1 to a given proposition. In some simple models, your credences are fully determined by what you’re certain of;³⁸ but in our models, they usually aren’t—for example, in the spoon frame you are certain that the only proposition you assign probability 1 to is $W = \{d_l, \bar{d}_l, d_h, \bar{d}_h\}$, and yet you still don’t know what your credence function is. Let’s say that P_w is **certainty-certain** if it is certain of its certainties in this way.³⁹

Let’s constrain P_w to be certainty-certain. Then if it has second-order uncertainty, it has third-order uncertainty, fourth-order uncertainty, and so on. It’s turtles all the way up:

Theorem 5.6.2 (Turtle Theorem). Suppose that P_w is certainty-certain. Then if P_w is uncertain what $P(\cdot)$ is (i.e. what $\mathbb{E}_P^1(\cdot)$ is), then for all n , P_w is uncertain what $\mathbb{E}_P^n(\cdot)$ is.

This result bears on an old debate. It’s sometimes said that ‘If anything is to be probable, something must be certain’ (Lewis 1946). If we interpret this as saying that the facts which

I think just need transitivity here

³⁶You might notice that these expectations are converging. Treating P and \hat{P} as stochastic matrices, expectations become matrix multiplication. The claim that, for all w , $P_w(\cdot) = \mathbb{E}_{P_w}(\hat{P}(\cdot))$ is equivalent to the claim that $P\hat{P} = P$. The failure of Self-Martingale means that, under ambiguity, $P\hat{P} \neq P$. But the Markov chain convergence theorem implies that in any finite, self-aware (hence ‘aperiodic’) frame, the sequence $P, P\hat{P}, P\hat{P}\hat{P}, \dots$ will converge—in particular, each row will converge to the stationary distribution of its communicating class. This means that, as we iterate expectations, you become increasingly certain of your expectation of your expectation of ... of your expectation of X . The flip side is that, since all the P_w you leave open converge to the same infinitely-iterated expectation, you can’t use it to infer what your expectation of X is. See Williamson 2008 for a helpful discussion.

³⁷Formally, let your ‘0-order uncertainty’ be the truth value, $\mathbb{E}_P^0(\cdot) = \mathbb{1}_{(\cdot)}$, and let $\mathbb{E}_P^{n+1}(\cdot) = \mathbb{E}_P(\mathbb{E}_P^n(\cdot))$. Thus $\mathbb{E}_P^1(q) = P(q)$, $\mathbb{E}_P^2(q) = \mathbb{E}_P(P(q))$, $\mathbb{E}_P^3(q) = \mathbb{E}_P(\mathbb{E}_P(P(q)))$, etc. Notably: for all n , $\mathbb{E}_P^n(\cdot)$ is a variable probability distribution, so $\mathbb{E}_{P_i}^n(\cdot)$ can be represented with a stochastic vector, and $\mathbb{E}_P^n(\cdot)$ with a stochastic matrix. You can calculate these by multiplying the matrix P by itself. $\mathbb{E}_{P_i}^2(\cdot)$ is the i th row of PP , $\mathbb{E}_{P_i}^n(\cdot)$ is the i th row of P^n , etc.

³⁸These are sometimes called ‘prior frames’. See e.g. Williamson 2000, 2019; Dorst 2020a,b.

³⁹Formally, P_w is certainty-certain iff for all v, u such that $P_w(v) > 0$ and $P_w(u) > 0$, for all q , $P_v(q) = 1$ iff $P_u(q) = 1$. This is equivalent to saying that P_w is certain that P obeys both **positive access** and **negative access**: if $P(q) = 1$, then $P(P(q) = 1) = 1$ and if $P(q) < 1$, then $P(P(q) < 1) = 1$. Equivalently, w is certain that the binary relation vRu that holds iff v assigns positive probability to u , $P_v(u) > 0$, is Euclidean (and so transitive). Equivalently, the modal operator $\langle P(\cdot) = 1 \rangle$ obeys the logic KD45.

determine those probabilities must themselves be certain, then this is *true under clarity, but false under ambiguity*. Under clarity, there’s always a layer of certainty, a ‘cognitive home’ (Williamson 2000) on which to base your decisions—namely, what your (rational) expectations are. Under ambiguity, there is no such layer. If you knew your higher-order expectations, you’d be able to tell what your first-order expectations are. But you don’t, so you can’t.

Upshot: since your expectations of your own expectations can be both inaccurate and uncertain, you can’t use them to remove your higher-order uncertainty.

5.6.2 Future Stability

But might there be another way? Here you are, unsure whether you’re 0.45- or 0.55-confident that I used a spoon. But (let’s suppose) you know that you’re a rational person—so you know that you tend to act in the way that you expect to be best. So why can’t you just force yourself to *act*, and then learn what your credence is? For example, you could force yourself to choose between \$5 for sure, and a bet X that pays \$10 if I own a dozen spoons and \$0 if I don’t. You know the rational thing to do is to take X iff $\mathbb{E}_{P_a}(X) > 5$, iff $P_a(d) > 0.5$. So why can’t you just observe what you do—take the bet, or take the \$5—and figure out whether your credence was above or below 50%? And if you can do *that*, why do you need to take the action at all? Can’t you just *hypothetically* take the action? Or *simulate* doing the action, and see what you do in your simulation?

The answer, of course, is cognitive noise. Or rather: if you knew you *didn’t* have cognitive noise, then you could indeed remove your higher-order uncertainty by acting (Zollman and Dorst 2025).⁴⁰ I’ve argued in Chapter 2 that sensible agents have cognitive noise only if they have higher-order uncertainty. This suggests that a near converse is true: if you’re a sensible agent and you’re sure that you *don’t* have cognitive noise, then higher-order uncertainty will be transient.

But for agents like us, the strategy won’t work. If you are unsure whether you will reliably act on your credences—in the sense that multiple different credence functions *might* lead to the same action—then even after you condition your credences on how you acted, you’ll still be unsure what your credences were.

For example, suppose you know that you will elicit your action by drawing 5 outcome-samples from your credence function, and then taking the bet iff at least 3 of them are d -possibilities. Then you know that if $P(d) = 0.55$, it’s around 59%-likely that you’ll take the bet; and if $P(d) = 0.45$, it’s around 41%-likely that you will. Thus taking the bet provides a bit of evidence in favor of your original credence being 0.55, with a Bayes factor of around $\frac{0.59}{0.41} = 1.44$. Conditioning on this would shift your distribution slightly in favor of $\langle P = \pi_h \rangle$. If you started with π_l , your credence that you have π_h would jump from 0.4 to 0.49.⁴¹ And if you started with π_h , your credence that you have π_h would jump from 0.6 to 0.68. Ambiguity remains.

To see the dynamics more clearly, consider what happens when you draw samples from your distribution and update on them. This provides a model of *thinking*: using the results of your mental simulations to get evidence about your opinions, and (thus) about the world.

⁴⁰The quick way? Force yourself to write down the truth-value of q and get paid using a strictly-proper scoring rule; then you know you’ll write down your credence in q .

⁴¹The posterior odds are $\frac{0.4}{0.6} \cdot \frac{0.59}{0.41} = 0.96$, so the posterior is $\frac{0.96}{1+0.96} = 0.49$.

Consider the spoon case. Suppose you draw a single estimate-sample of $P(d)$ from your distribution—yielding either 0.45 or 0.55—and condition on the outcome. To model this update explicitly, we need to fine-grain our outcome space so that each of our original 4 worlds is compatible with you sampling 0.45 or 0.55, i.e. a *low* or *high* sample. Thus our original outcome space W gets fine-grained to $W \times \{l, h\}$, where (w, l) is what’s true at an original world w where you draw a low (0.45) sample from your distribution, and (w, h) is where you draw a high one. Let’s abbreviate these w_l and w_h for readability, so for example d_{ll} is the possibility where you’re at d_l and draw a low sample.

Suppose you’re at d_l . How likely are you draw a low sample, l ? That is: what proportion of the probability assigned to $\{d_{ll}, d_{lh}\}$ should go to d_{ll} ? That’s determined by your credence (at d_l) that $P(d) = 0.45$ —which is 60%: $P_{d_l}(P(d) = 0.45) = 0.6$. Meanwhile, you’re 40%-likely to draw a high sample, h . Thus we use your higher-order distribution at w to determine how to divide the probability assigned to w into that assigned to w_l and w_h .

For example, $\pi_l = \begin{pmatrix} d_l & \bar{d}_l & d_h & \bar{d}_h \\ .15 & .45 & .3 & .1 \end{pmatrix}$ becomes $\begin{pmatrix} d_{ll} & d_{lh} & \bar{d}_{ll} & \bar{d}_{lh} & d_{hl} & d_{hh} & \bar{d}_{hl} & \bar{d}_{hh} \\ .09 & .06 & .27 & .18 & .12 & .18 & .04 & .06 \end{pmatrix}$. For instance, the 0.15 credence that went to d_l is now split into $0.6(0.15) = .09$ at d_{ll} and $0.4(0.15) = .06$ at d_{lh} . And so on.

The matrix representation quickly gets messy, but the factorized form is a bit easier:

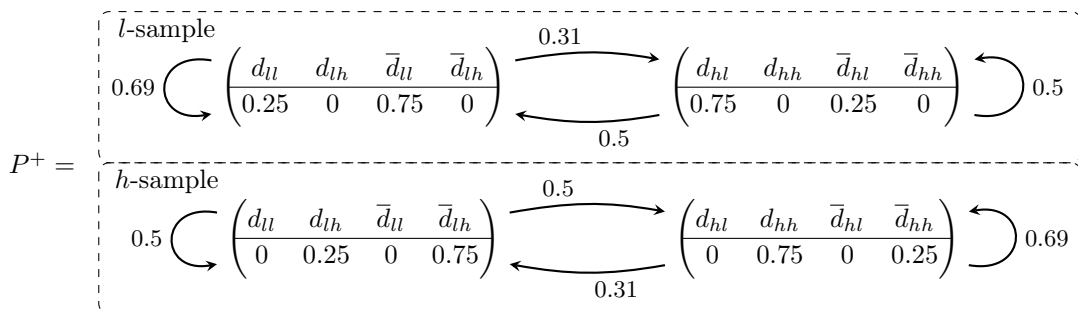
$$P = \begin{matrix} & \begin{matrix} \text{0.6} \end{matrix} & & & & & & & \\ & \begin{matrix} \curvearrowright \\ \curvearrowleft \end{matrix} & \begin{pmatrix} d_{ll} & d_{lh} & \bar{d}_{ll} & \bar{d}_{lh} \\ .15 & .1 & .45 & .3 \end{pmatrix} & \begin{matrix} \xrightarrow{0.4} \\ \xleftarrow{0.4} \end{matrix} & \begin{pmatrix} d_{hl} & d_{hh} & \bar{d}_{hl} & \bar{d}_{hh} \\ .3 & .45 & .1 & .15 \end{pmatrix} & \begin{matrix} \curvearrowright \\ \curvearrowleft \end{matrix} & \begin{matrix} \text{0.6} \end{matrix} & \\ & & & & & & & \end{matrix}$$

Indeed, though the order has changed, you can see that the informed distributions assign the same probabilities over samples that the original (uninformed) probabilities do to worlds. That’s not an accident: the informed probabilities know what P is, and therefore know your sampling propensities.

So, supposing you’re at d_{ll} , your prior is (as before) 60%-confident that you have a low credence in d and $0.6(0.25)+0.4(0.75) = 0.45$ -confident of d , and it (newly) is $0.6(0.15+0.45)+0.4(0.3+0.1) = 0.52$ -confident that you’ll draw a low sample, and 0.48-confident you’ll draw a high sample.

Now what happens when you draw a sample, and condition on whether it is low or high? Doing so shifts your distribution over the informed-credence classes, but doesn’t affect your informed credences about the original possibilities—since sampling only provides evidence about what credence function you had, which your informed credences already know. We *could* calculate all these shifts by hand, for each pair (w, l) zeroing out the pairs (v, h) and renormalizing, and vice versa for (w, h) . But a more convenient way is to multiply the higher-order distributions by the ratios between the likelihoods of the sample you observed and marginal (average) likelihood—that is, how much more likely the sample is given that class than on average. If your prior higher-order distribution is η and you observe an l sample, your credence that $\langle P = \pi_i \rangle$ is multiplied by the ratio $\frac{\pi_i(P(d)=0.45)}{\mathbb{E}_\eta(P(P(d)=0.45))}$. For example, if $\eta = (0.6, 0.4)$ over $(\langle P = \pi_l \rangle, \langle P = \pi_h \rangle)$, then $\mathbb{E}_\eta(P(P(d) = 0.45)) = 0.6(0.6) + 0.4(0.4) = 0.52$, so your posterior higher-order distribution equals

$(0.6(\frac{0.6}{0.52}), 0.4(\frac{0.4}{0.52})) \approx (0.69, 0.31)$.⁴² Repeating this calculation for the different priors and samples yields the following posterior:



Focus on the bottom right, where you started with π_h , elicited a high estimate-sample of 0.55 for your credence in d , and conditioned on that. This boosts your credence that you originally had π_h from 0.6 to 0.69. We expected that: sampling from your distribution gives you evidence about what credence function you had.

More surprisingly, updating the fact that you sampled 0.55 *also* boosts your credence *that I own a dozen spoons* from 0.55 to $0.31(0.25) + 0.69(0.75) = 0.595$. Why? For the same reason that Self-Reflection fails. Under ambiguity, when you get *evidence* about what your credences were, that should *change* your credences. Drawing an *h*-sample is evidence that you had a high credence to begin with, i.e. evidence that your informed credence in d is 0.75, i.e. evidence for d . Thus, under ambiguity, *thinking*—using your beliefs to simulate what might happen, and then updating on how your simulations went—can help you figure out what’s true. When someone asks you, ‘Does Kevin own a dozen spoons?’, and you say, ‘I’m trying to figure out what I think about that’, you might well be speaking literally—and, by figuring out what you think, get a more-accurate answer about my spoon collection.

Upshot: sampling from your own distribution is a form of updating on your own actions which provides evidence—but, crucially, *inconclusive* evidence—about what your opinions were. Ambiguity remains.⁴³

5.7 Upshots

We’ve now seen how to reason about ambiguity rigorously, without falling into the traps which so often confuse discussions of higher-order probability. The result is a simple, easy-to-use class of probabilistic models—Ambiguous-Bayesian ones—that maintain the structure of Standard-Bayesian updates but allow ambiguity. Such models have a rich structure, and are incredibly flexible.

⁴²In the general case we’ll need to define the likelihood $L_w(\sigma)$ at each class $\langle P = P_w \rangle$ for how likely you are to draw a sample with whatever features you’re tracking. (In the case at hand, feature we’re tracking is ‘What is P(d)?’) For example, if you were outcome-sampling, drawing 10 samples and conditioning on (say) the proportion of samples that were d -possibilities, then the likelihood function at w would be $\text{Binomial}(10, P_w(d))$.

⁴³Of course, ideal agents might be able to draw infinitely-many samples from their distribution and condition on them, thereby removing their higher-order uncertainty. But agents like us can’t—indeed, such a process would be computationally infeasible, due to the combinatorial explosion induced by tracking all the possible sets of samples you might draw.

Too flexible? As we'll see, Ambiguous-Bayesian models permit some quite strange—perhaps pathological—behavior. If we want to make the case that people are *rational* under ambiguity, we will need an independent assessment of what rationality amounts to. The next chapter propose that we take the *value of rationality* as axiomatic: being rational, whatever else it is, is meant to help you make better decisions and have more accurate beliefs. I'll show how we can formalize this as an additional constraint on frames—one that, under clarity, subsumes many classic arguments for Standard Bayesianism, but which itself permits rich structures of ambiguity. Such “Valuable Bayesian” models will be the ones we will use throughout the rest of this book.

For those eager to get to those philosophical details, skip ahead. For those who want a better sense of how to work with Ambiguous-Bayesian models, this chapter contains several further theoretical sections: practice problems (§5.8.1), a discussion of Ellsberg urns that illustrates the factorization theorem (§5.8.2), a glossary on notation (§5.8.4), and a list of the standard theorems of probability theory that are valid on probability frames (§5.8.5).

Should I invite people to skip to Part III? ch_RA is really only needed for aficionados

5.8 Appendix[†]

5.8.1 Practice Problems

[Under construction—ignore this]

1. [Construct an example that requires using the factorization theorem]
2. Prove Fact 5.4.1 that in any finite, self-aware frame, if P_w is ambiguous, then there is a q and x such that $P_w(q|P(q) = x) \neq x$.
3. Prove Theorem 5.6.1, that Self-Martingale implies clarity. (Hint: given [problem from Chapter 4], it suffices to show that Self-Martingale is valid only if the ‘ w assigns positive probability to v ’-relation, $P_w(v) > 0$, is an equivalence relation.)

5.8.2 Ellsberg Urns

[Under construction—ignore this]

Let's use this recipe to specify a higher-order-uncertain model. Suppose you are holding a spoon—you get to examine it, note the curve of its sides, its weight etc. But you're not allowed to toss it. This provides you with some evidence about its bias. But let's suppose you're higher-order uncertain: you're not sure what you think. As a result, you don't know what your informed distribution is—what you would think, were your higher-order-uncertainty removed.

Let's suppose you're unsure whether your informed distribution over the possible biases is $\hat{P}_1 = \text{Beta}(6,2)$, $\hat{P}_2 = \text{Beta}(4,4)$, or $\hat{P}_3 = \text{Beta}(2,6)$ —illustrated in Figure 5.3. ($\text{Beta}(1+n,1+m)$ is the distribution you'd have if you'd started out uniform over the biases—at a $\text{Beta}(1,1)$ distribution—and then tossed $n + m$ times, getting n ups and m downs. If you want to brush up on Beta distributions, see §??; but the details shouldn't matter.)⁴⁴

⁴⁴Since Betas are only defined over the possible biases of the spoon (between 0 and 1), we need to enrich these

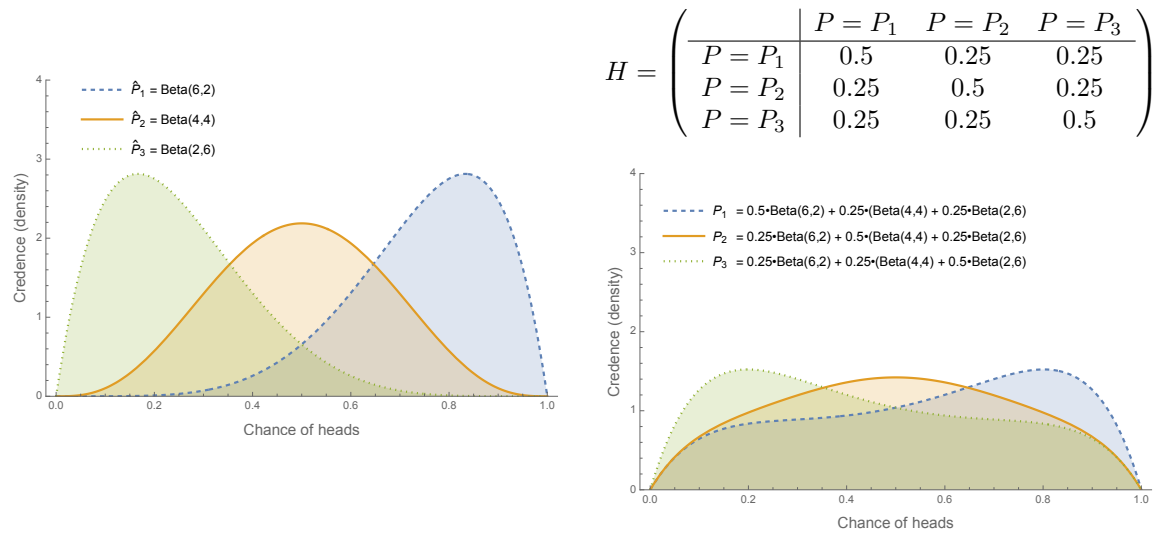


Figure 5.3: *Left:* Your three possible informed distributions over the bias of the spoon. *Right:* The top matrix shows your higher-order distributions, given each possible credence function; the bottom shows your induced distributions over the bias of the spoon, using Informed Reflection and the three possible credences.

Suppose in each case, your higher-order distribution assigns 50% to having your actual opinions, and 25% to the other two, so your higher-order distribution H is captured by the matrix in the top right of Figure 5.3. Thus your possible credence distributions over the biases are displayed in the bottom right of Figure 5.3. We get these curves by taking weighted averages of the informed distributions—for instance, the probability density that P_2 assigns to 0.5 equals $0.25 \cdot \text{Beta}(6, 2)(0.5) + 0.5 \cdot \text{Beta}(4, 4)(0.5) + 0.25 \cdot \text{Beta}(2, 6)(0.5) \approx 1.42$.

Notice a few things. First, you have both probabilistic uncertainty (about the bias of the spoon) and higher-order uncertainty (about what your distribution over the biases is). Second, this higher-order uncertainty leads your credences to be much less certain about the bias of the coin than your informed credences—the distributions are flatter. Third, your possible credences are not Beta distributions, since (generically) the average of two Beta distributions does not make another Beta distribution.

This last point is important, for it means that you update differently than you would if you were higher-order certain. Consider what happens when the spoon lands up. Focus on the possibility where your credence function is P_2 , so your informed distribution is $\text{Beta}(4, 4)$ over the bias. At this stage, both P_2 and \hat{P}_2 have the same mean, at 0.5—meaning they each start with an estimate of the bias of 50%. But your credences are more uncertain than your informed credences, so they

distributions to capture claims about what your distribution is, and whether the spoon will land up or down on the first toss. Thus our full algebra is $W = W_1 \cup W_2 \cup W_3$; W_i is the proposition that your informed distribution is \hat{P}_i (so for all $w \in W_i$, $\hat{P}_w = \hat{P}_i$), and each $W_i = \{i\} \times [0, 1] \times \{u, d\}$ are indexed copies of the possible biases between 0 and 1, crossed with whether the spoon will land up or down. So, for instance, the triple $(2, 0.47, u)$ is the world at which $\hat{P} = P_2$, the bias is 0.47, and the spoon lands up. Where B is a variable capturing the bias, we set $\hat{P}_i(u|B = a) = a$, so each informed distribution defers to the bias on how the spoon will land. Since the \hat{P}_i are informed, we set $\hat{P}_i(W_i) = 1$, i.e. $\hat{P}_i(\hat{P} = \hat{P}_i) = 1$.

update more when they see the spoon land up—see the right of Figure ?? . You ‘overreact’ to the evidence, compared to what you would do if you were higher-order certain: your estimate for the bias becomes 61%, rather than 55%

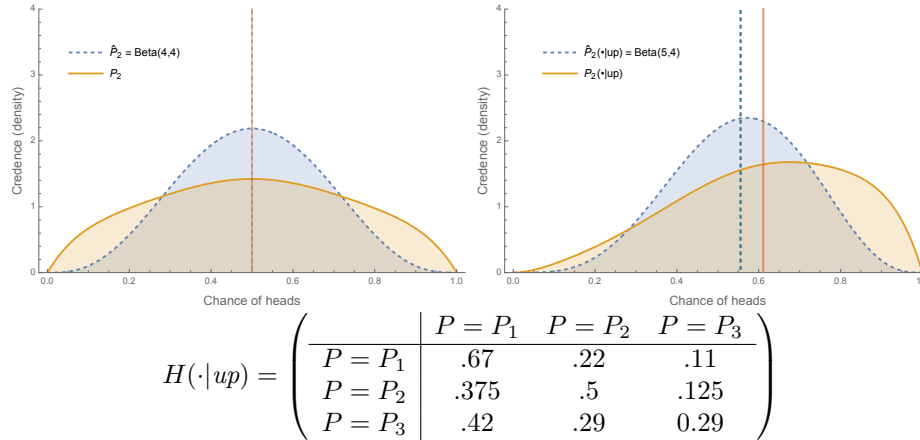


Figure 5.4: Higher-order uncertainty induces ‘overreaction’, compared with informed priors. Both P_2 and \hat{P}_2 begin with an estimate of 50% for the bias—but if it land sup, \hat{P}_2 shifts it’s estimate to 55%, while P_2 shifts to 61%.

Why? Hindsight bias! While \hat{P}_2 is only unsure about the bias, P_2 also unsure what its priors are: it leaves open that it has a higher ($P = P_1$) or lower ($P = P_3$) estimate of the bias. And since it thinks its credences are correlated with the truth, seeing the spoon land up provides evidence that it had the higher priors ($P = P_1$) to begin—shifting it from (0.25, 0.5, 0.25) to (0.375, 0.5, 0.125) over its possible priors. This hindsight effect for all three prior distributions is shown on the bottom of Figure 5.4.

This dynamic—of a hindsight effect leading to ‘over-reaction’ to evidence—is common under ambiguity, and will be a key piece part of the discussion in Chapter ?? on overconfidence. This is one example (of many) of how the connections between rational credence and truth start to look different under ambiguity.

5.8.3 Proofs†

[Under construction—ignore this]

Theorem 5.4.2 (Factorization Theorem). A self-aware synchronic frame (W, \mathfrak{a}, P) with finitely many candidates validates Informed Reflection iff it can be factored into two components:

1. A set of disjoint probability spaces $\{(W_1, \delta_1), \dots, (W_n, \delta_n)\}$ where $W = \bigcup_i W_i$, and for any $w_i \in W_i$, we define $\hat{P}_{w_i} := \delta_i$; and
2. For each space, a (unique⁴⁵) ‘higher-order’ distribution η_i defined over the partition $\{W_1, \dots, W_n\}$;

Such that for any world $w_j \in W_j$ and $q \subseteq W$: $P_{w_j}(q) = \mathbb{E}_{\eta_j}(\hat{P}(q))$, i.e. equals $\sum_{W_i} \eta_j(W_i) \cdot \delta_i(q)$.

⁴⁵It’s important (and easy to forget) that the higher-order distributions must be distinct. If $\eta_i = \eta_j$, then they’ll give rise to the same overall distribution P , meaning that learning what P is *won’t* determine what \hat{P} is, as it must.

Fact 5.4.3. If a synchronic frame (W, \mathfrak{a}, P) validates No Foregone Questions, then P is factorable.

Theorem 5.6.2 (Turtle Theorem). Suppose that P_w is certainty-certain. Then if P_w is uncertain what $P(\cdot)$ is (i.e. what $\mathbb{E}_P^1(\cdot)$ is), then for all n , P_w is uncertain what $\mathbb{E}_P^n(\cdot)$ is.

Proof. □

Theorem 5.5.1 (Greaves and Wallace 2006). If P_w is regular over \mathcal{Q} and A is strictly proper, then for all $u \neq u_w^c$, $\mathbb{E}_{P_w}(A(u_w^c)) > \mathbb{E}_{P_w}(A(u))$.

Proof. □

Corollary 5.5.2 (Conditioning Maximizes Expected Accuracy). If A is strictly proper, P is regular over \mathcal{Q} , and the frame $(W, \mathfrak{a}, P, P^+)$ maximizes expected accuracy with respect to \mathcal{Q} , then P^+ is the result of conditioning P on \mathcal{Q} : for all w , $P_w^+(\cdot) = P_w(\cdot | \mathcal{Q}_w)$.

Proof. □

5.8.4 Notational Niceties†

A friend once said that my biggest innovation was notation. He was (I hope) joking. But higher-order probability *is* ripe for linguistic confusions, usually due to the subtle line between which terms are *descriptions* (variables) and which are *rigid designators* (constants). So let's be crystal clear.

rewrite (delete?)
this section; add
 \hat{P}

$\mathbf{p}, \mathbf{q}, \mathbf{r}$, and so on will rigidly refer to claims (propositions, events,...), i.e. subsets of W . When I'm using other terms for claims (like u and b_u above), context will make it clear.

$\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}$, and so on will rigidly refer to worlds: $w \in W$.

$\mathbf{a}, \mathbf{b}, \mathbf{c}$, and so on will rigidly refer to real numbers: $a \in \mathbb{R}$.

π, δ, η , and so on will rigidly refer to probability distributions over W .

$\mathbf{P}, \mathbf{P}^+, \mathbf{P}^i$, and sometimes other capital roman letters (like T above) are *probability variables*: they describe a probability function, and therefore are modeled as functions from worlds w to distributions P_w over W .

$\mathbf{P}_w, \mathbf{P}_w^\ominus, \mathbf{P}_w^+$, and so on, which saturate a probability variable ($\mathbf{P}, \mathbf{P}^+, \dots$) with a world w are rigid designators for particular distributions, i.e. are vectors like π .

\mathbf{X}, \mathbf{Y} , and so on are *random variables*, i.e. descriptions of numbers, modeled as functions from worlds w to numbers X_w —so ' X_w ' (or ' $X(w)$ ') is a rigid designator for a number (like '17'). Note that since ' \mathbf{P} ' describes a probability function, ' $\mathbf{P}(\mathbf{q})$ ' describes a number—so is a function from worlds w to numbers $P_w(\mathbf{q})$ capturing your credence in \mathbf{q} at w .

$\langle \mathbf{P}(\mathbf{q}) = \mathbf{a} \rangle, \langle \mathbf{P} = \pi \rangle, \langle \mathbf{X} \leq \mathbf{b} \rangle$, and so on are propositions. They use variables to pick out sets of worlds: those where the variable ($\mathbf{P}(\mathbf{q})$, or \mathbf{P} , or \mathbf{X}) has the specified value ($P_w(\mathbf{q}) = \mathbf{a}$, or $P_w = \pi$, or $X_w \leq \mathbf{b}$).

Notice embedding a rigidly-designated probability function inside another is simply a notational shorthand for embedding a *number*: since $P_x(q)$ is a number (say, a), $P_{\textcircled{a}}(P_x(q) = b)$ is equivalent to $P_{\textcircled{a}}(a = b)$, which is either 0 or 1. Still, this can be useful—for example, if we want to refer to your actual credence that $P_{\textcircled{a}}(q)$ is your credence in q , we can write $P_{\textcircled{a}}(P(q) = P_{\textcircled{a}}(q))$, which equals $P_{\textcircled{a}}(\{w : P_w(q) = P_{\textcircled{a}}(q)\})$, and so can be nontrivial.

Finally: one of the most important notions in Bayesian models is the *expected value* of a variable: the probability-weighted average value that a bunch of independent copies of that variable would (likely) yield, according to a given probability function. Since expected values are always relative to probability functions, and probability functions can be picked out either *rigidly* (π) or *descriptively* (P), that means we likewise need to track the distinction between a *rigidly* and *descriptively* designated expectation:

\mathbb{E}_{π} vs. \mathbb{E}_P vs. \mathbb{E}_{P_w} : \mathbb{E}_{π} is a rigid designator for an expectation function: give it a variable X , and it'll output a number $\mathbb{E}_{\pi}(X)$ that is π 's expectation of X . $\mathbb{E}_P(X)$ is a *description* of an expectation function: give it a variables X , and it'll output a function $\mathbb{E}_P(X)$ from worlds w to P_w 's expectation of X : $\mathbb{E}_{P_w}(X)$. Finally, \mathbb{E}_{P_w} is a rigid designator for P_w 's expectation function: give it a variable X , and it'll output a number $\mathbb{E}_{P_w}(X)$.

How are expectations calculated? Several ways. The simplest is to use the probability function (π , or P_w) to weight each world w , multiply that weight by the value of X at w , and sum them up:

$$\mathbb{E}_{\pi}(X) := \sum_{x \in W} \pi(x) \cdot X_x \qquad \mathbb{E}_{P_w}(X) := \sum_{x \in W} P_w(x) \cdot X_x$$

Thought of this way, it becomes useful to *also* think of a variable X defined over W as a vector—having ordered the worlds (w_1, \dots, w_n) , the i th entry of X is the value X assigns to the i th world, X_{w_i} . For example, in the above spoon-flipping model, where $W = (u_u, u_d, d_d, d_u)$ (i.e. *up and biased-up, up and biased-down...*), then the variable X for *the chance of landing up* is the vector $X = \begin{pmatrix} u_u & u_d & d_d & d_u \\ 2/3 & 1/3 & 1/3 & 2/3 \end{pmatrix}$. (As with probability vectors, the top row is labeling, and is not officially part of the vector.) When we treat both probability distributions and variables as vectors, expectations become *dot products*: π 's expectation of X , $\mathbb{E}_{\pi}(X)$, is (thinking of them as vectors) the dot product of $\pi \cdot X := \pi(w_1)X_{w_1} + \dots + \pi(w_n)X_{w_n}$. For instance, according to your prior $\pi = (2/6, 1/6, 2/6, 1/6)$ before learning the bias of the spoon, your expectation for it's bias was the weighted average $\pi \cdot X = (\frac{2}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}) \cdot (\frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}) = \frac{2}{6}(\frac{2}{3}) + \frac{1}{6}(\frac{1}{3}) + \frac{2}{6}(\frac{1}{3}) + \frac{1}{6}(\frac{2}{3}) = \frac{1}{2}$. As we'll see, thinking of expectations this way helps with both calculations and conceptual reasoning.

Last point: while $\mathbb{E}_{\pi}(X)$ is a rigid designator, $\mathbb{E}_P(X)$ is a description. Thus, like P , we can use \mathbb{E}_P to form claims about your expectations that may be true or false—they pick out nontrivial sets of worlds, and so can be objects of uncertainty. For example, $\langle \mathbb{E}_P(X) = 2/3 \rangle$ is the set of worlds where your expectation of X is $2/3$, i.e. $\{w \in W : \mathbb{E}_{P_w}(X) = 2/3\}$. In the biased-spoon model, this was true at u_u and d_u (where you know the spoon is biased-up), and false at u_d and d_d (where you know the spoon is biased-down). Thus, for instance, Teddy assigns intermediate credence to $\langle \mathbb{E}_P(X) = 2/3 \rangle$: at the actual world u_d , he's $2/3$ -confident of it: $T_{u_d}(\mathbb{E}_P(X) = 2/3) = T_{u_d}(b_u) = 2/3$. We'll make much of these nontrivial beliefs about your expectations in what follows.

5.8.5 Probabilistic Reasoning Under Ambiguity[†]

This is a reference section, briefly which forms of probabilistic reasoning are safe to use in probability frames. The short story is: all of them. (The point is to show just how little we've changed.) The longer story involves some care when we embed probability functions within others, to track whether they are being read rigidly (in which case: everything's as normal) or descriptively (in which case, be careful!). Let w be an arbitrary world in an arbitrary probability frame, $q, r \subseteq W$ be arbitrary propositions, and $@$ be the actual world:

rewrite.
Trim? Focus on
lower rule?

Additivity: Probabilities are (certain to be) additive:

$$\begin{aligned} \text{If } q \cap r = \emptyset, \text{ then } P_w(q \cup r) &= P_w(q) + P_w(r) \\ P_{@}(P(q \vee r) = P(q) + P(r)) &= 1 \end{aligned}$$

Ratio Formula: Conditional probabilities (are certain to) satisfy the ratio formula:

$$\begin{aligned} \text{If } P_w(r) > 0, \text{ then } P_w(q|r) &= \frac{P_w(q \cap r)}{P_w(r)} \\ P_{@}(P(q|r) = \frac{P(q \cap r)}{P(r)}) &= 1 \end{aligned}$$

Total Probability: For any partition $\mathcal{Q} = \{q_1, \dots, q_n\}$ of W , your credence in r is (certain to be) an average of its conditional credence given the true cell of \mathcal{Q} :

$$\begin{aligned} P_w(r) &= P_w(q_1) \cdot P_w(r|q_1) + \dots + P_w(q_n) \cdot P_w(r|q_n) = \sum_{q_i \in \mathcal{Q}} P_w(q_i) \cdot P_w(r|q_i) \\ P_{@}(P(r) = P(q_1) \cdot P(r|q_1) + \dots + P(q_n) \cdot P(r|q_n)) &= 1 \end{aligned}$$

Bayes Theorem: The conditional probability of q given e is (certain to be) proportional to how likely q makes e :

$$\begin{aligned} P_w(q|e) &= \frac{P_w(q) \cdot P_w(e|q)}{P_w(e)} \\ P_{@}(P(q|e) = \frac{P(q) \cdot P(e|q)}{P(e)}) &= 1 \end{aligned}$$

Expectations: For any random variable (functions from worlds $w \in W$ to numbers X_w), the expectation of X according to P is (certain to be) a probability-weighted (according to P) average of X 's possible values:

$$\begin{aligned} \mathbb{E}_{P_w}(X) &= \sum_{z \in W} P_w(z) \cdot X_z = \sum_{a \in \mathbb{R}} P_w(X = a) \cdot a \\ P_{@}(\mathbb{E}_P(X) = \sum_{z \in W} P(z) \cdot X_z) &= 1, \text{ and } P_{@}(\mathbb{E}_P(X) = \sum_{a \in \mathbb{R}} P_w(X = a) \cdot a) = 1 \end{aligned}$$

Total Expectation: For any partition $\mathcal{Q} = \{q_1, \dots, q_n\}$.⁴⁶ your expectation for X is (certain to be) a probability-weighted average of your conditional expectation, given the true cell of \mathcal{Q} :

Let $\mathbb{E}_{P_w}(X|q)$ be P_w 's conditional expectation of X given q : $\sum_{z \in W} P_w(z|q) \cdot X_z$. Then:

$$\begin{aligned} \mathbb{E}_{P_w}(X) &= P_w(q_1) \mathbb{E}_{P_w}(X|q_1) + \dots + P_w(q_n) \mathbb{E}_{P_w}(X|q_n); \text{ and} \\ P_{@}(\mathbb{E}_P(X) = P(q_1) \mathbb{E}_P(X|q_1) + \dots + P(q_n) \mathbb{E}_P(X|q_n)) &= 1 \end{aligned}$$

This is a tricky point. Suppose we follow standard notation and define the *variable* for P_w 's conditional expectation of X given the true answer to \mathcal{Q} : $\mathbb{E}_{P_w}(X|\mathcal{Q}) := \mathbb{E}_{P_w}(X|q_z)$ at world z , where q_z is the true cell of \mathcal{Q} at z . Then total expectation can be equivalently stated as:

$$\mathbb{E}_{P_w}(\mathbb{E}_{P_w}(X|\mathcal{Q})) = \mathbb{E}_{P_w}(X) \quad (\text{tower rule})$$

⁴⁶Or variable Y , since $\{\langle Y = a_1 \rangle, \dots, \langle Y = a_n \rangle\}$; remember that we're usually focusing on finite W .

This is the ‘tower rule’ of expectations. It is true, but we have to proceed with caution (recall footnote ??). This statement keeps the probability function P_w rigid in the embedded expectation. That’s crucial—if we change the embedded expectation to a *description*, using \mathbb{E}_P instead of \mathbb{E}_{P_w} , the result is often false. Often

$$\mathbb{E}_{P_w}(\mathbb{E}_P(X|\mathcal{Q})) \neq \mathbb{E}_{P_w}(X)$$

For example, this equality fails in our spoon frame (§??): let $w = s_l$, \mathcal{Q} be the trivial partition $\{W\}$, and $X = \mathbb{1}_s$, then $\mathbb{E}_{P_w}(\mathbb{E}_P(X|\mathcal{Q})) = \mathbb{E}_{P_{s_l}}(P(s)) = 0.55 > 0.5 = P_{s_l}(s) = \mathbb{E}_{P_{s_l}}(\mathbb{1}_s)$.

With this in mind, you might be tempted to look at the tower rule, note that it’s true at all worlds, and infer that the proposition $\langle \mathbb{E}_P(\mathbb{E}_P(X|\mathcal{Q})) = \mathbb{E}_P(X) \rangle$ is true at all worlds. But it’s not: we’ve erased the embedded w within the expectation operator, changing it to P_w ’s expectation conditional on the true cell of \mathcal{Q} to P ’s expectation conditional on the true cell of \mathcal{Q} —since the latter can vary from world to world, it needn’t equal P_w .

This is a general phenomenon: when we form propositions with embedded probability functions, we standardly only saturate the outer one, on pain of changing the meaning. Teddy is unsure what your credence is that the spoon will land up: $\langle T(P(u) = 2/3) = 2/3 \rangle$. Where is this proposition true? At the set of worlds w at which $T_w(P(u) = 2/3) = 2/3$, *not* the set of worlds w where $T_w(P_w(u) = 2/3) = 2/3$, for that makes the embedded claim no longer the sort of thing Teddy is unsure about: he assigns it 0 or 1.

Upshot: the tower rule holds in probability frames, and indeed is certain to hold. But the best way to assert this is to explicitly write out the sum, as I did above—the notation $\mathbb{E}_{P_w}(X|\mathcal{Q})$ is ripe for confusion.

Linearity and additivity of expectations: Your expectations are certain to be linear and additive. For any variables X, Y , and real numbers a, b :

$$\mathbb{E}_{P_w}(aX + b) = a \cdot \mathbb{E}_{P_w}(X) + b, \text{ so}$$

$$P_{\textcircled{Q}}\left(\mathbb{E}_P(aX + b) = a \cdot \mathbb{E}_P(X) + b\right) = 1.$$

And $\mathbb{E}_{P_w}(X + Y) = \mathbb{E}_{P_w}(X) + \mathbb{E}_{P_w}(Y)$, so

$$P_{\textcircled{Q}}\left(\mathbb{E}_P(X + Y) = \mathbb{E}_P(X) + \mathbb{E}_P(Y)\right) = 1$$

The law of large numbers: If $\{X_1, \dots, X_n\}$ is a large set of variables that you treat as independent and identically distributed, then you’re confident that their average is roughly equal to their expectation:

If the X_i are iid according to P_w , with $\mathbb{E}_{P_w}(X_i) = a$ and finite variance, then as $n \rightarrow \infty$, $P_w\left(\frac{\sum_i X_i}{n} \approx a\right) \rightarrow 1$.

So if $P_{\textcircled{Q}}$ is certain that the X_i obey these constraints for P , then as $n \rightarrow \infty$, then

$$P_{\textcircled{Q}}\left(P\left(\frac{\sum_i X_i}{n} \approx a\right) \approx 1\right) \rightarrow 1$$

Central Limit Theorem: If $\{X_1, \dots, X_n\}$ is a large set of variables that you treat as iid, then as

$n \rightarrow \infty$, your distribution over their normalized sum is approximately Gaussian:

If the X_i are iid according to P_w , with $\mathbb{E}_{P_w}(X_i) = a$, $\text{Var}_{P_w}(X_i) = b$, and $\text{Var}_{P_w}(X_i^2) < \infty$, then as $n \rightarrow \infty$, P_w 's distribution for the variable $\frac{\sum X_i - a}{b\sqrt{n}}$ becomes approximately Gaussian.

Again, if $P_{\textcircled{a}}$ is certain that these constraints hold, then $P_{\textcircled{a}}$ is certain that P 's distribution for the normalized sum is approximately Gaussian.

In sum: probability theory works just like it always has.