

Chapter 4

Standard Bayesianism and its Limits

Abstract

Almost all (precise) probabilistic models used in the behavioral sciences conform to a set of constraints that I call ‘Standard Bayesianism’. This is for good reason: under clarity, minimal rationality conditions require as much. This chapter introduces Standard Bayesianism, shows why it implies clarity, and sketches its limits.

4.1 Plan

This chapter introduces our foil. Standard Bayesianism is a constraint on probabilistic updates that is widely used throughout philosophy and the behavioral sciences—but its suite of assumptions are rarely stated fully. This chapter will state them in using concrete examples, without presupposing any mathematics. We will introduce *probability spaces* (§4.2) and explain why modeling updates requires enriching them to *probability frames* (§4.3).¹ Using probability frames, we’ll formalize Standard Bayesianism, explain why it is so standard, and show why it implies clarity. We’ll conclude by canvassing its limits (§4.5), paving the way for Ambiguous Bayesianism.

4.2 Probability spaces

Bayesianism is for modeling situations of uncertainty. Here’s one. I have two coins—one is fair, one is double-headed. I grabbed one randomly, flipped it, and wrote the outcome (heads or tails) on a slip of paper. In a moment, you’ll read what I wrote down, and update your beliefs.

There are a variety of things you’re unsure about. How did the coin land? There are two possibilities: heads or tails. What is the bias, X , of the coin I selected? There are again two possibilities: if it’s fair, $X = 0.5$; if it’s double-headed, $X = 1$.

¹A version of Harsanyi type spaces (Harsanyi 1967), a probabilistic generalization of Kripke frames (Kripke 1963).

Lots of this is now redundant with Ch. 3. How to trim?

Bayesians usually model an uncertain situation with a **probability space** (W, π) consisting of an outcome space W and a probability distribution π defined over it.²

The **outcome space** W is a set of ‘worlds’ (or ‘states’)—think of it as dividing logical possibilities into ones that are fine-grained enough for our modeling purposes.³ That means W must capture all the distinctions we’re modeling uncertainty about—learning what world $w \in W$ we’re in must answer all relevant questions. In our case, we need to include worlds where the coin lands heads vs. tails, and worlds with the various possible biases ($X = 0.5$ or $X = 1$). Since double-headed coins always land heads, that means we need at least 3 possibilities: fair-and-tails (f_t), fair-and-heads (f_h), and unfair-and-heads (\bar{f}_h). So let $W = \{f_t, f_h, \bar{f}_h\}$. Each world tells us how the coin landed—for example, the coin lands heads at f_h and \bar{f}_h , but lands tails at f_t ; and the bias is $X = 0.5$ at f_t and f_h , but $X = 1$ at \bar{f}_h .

Given an outcome space W , claims (or ‘propositions’, or ‘events’) are subsets of W that pick out a set of ways the world might be—for example, the claim that the coin will land heads is $\{f_h, \bar{f}_h\}$ (label that ‘ h ’), and the claim that it’s fair is $\{f_t, f_h\}$ (label that ‘ f ’). Propositions are assigned truth-values at worlds by set membership: q is true at a world w iff w is a member of q (written ‘ $w \in q$ ’). For instance, the proposition that the coin is fair is true at f_t and f_h , since $f_t \in \{f_t, f_h\}$ and $f_h \in \{f_t, f_h\}$. We can do Boolean logic with these claims using operations on sets—for example, the claim that the coin will *not* land heads, $\neg h = W \setminus \{f_h, \bar{f}_h\} = \{f_t\}$, and the claim that the coin is both fair *and* lands heads is $f \& h = \{f_t, f_h\} \cap \{f_h, \bar{f}_h\} = \{f_h\}$.

Most of the Bayesian action is in the **probability distribution**, π . Think of it as a way of spreading a unit mass of mud over W . The probability that π assigns to a claim $q \subseteq W$ is the total amount of mud that π puts on the worlds in w . We can think of π as an assignment of non-negative numbers to $w \in W$ that together sum up to 1. Formally, π is a function that takes subsets of W to real numbers that is *non-negative*, *normalized*, and *additive*.⁴ It’ll be convenient to choose a canonical ordering of W and then represent probability functions with *stochastic vectors*: a list of non-negative numbers (summing to 1) in which the i th entry gives the probability of the i th world. Let’s order the worlds (f_t, f_h, \bar{f}_h) .

In the case at hand, the reasonable credence function to start with is $\pi = \begin{pmatrix} f_t & f_h & \bar{f}_h \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$. In this vector, the top row is simply labeling to remind you of the ordering of the worlds—it says that π assigns 25%-credence to the coin being fair and landing tails (f_t), 25% to it being fair and landing heads (f_h), and 50% to it being unfair (double-headed) and landing heads (\bar{f}_h). Why is this the reasonable prior credence? Since you know I picked the coin randomly, you should be 50-50 on whether it’s fair or unfair—so the probabilities in each of $\{f_t, f_h\}$ and $\{\bar{f}_h\}$ must sum up to 0.5. Since you know that double-headed coins always land heads, the latter 50% is all on a heads possibility. But since you know fair coins land heads half the time, you should divide your credence, given f , evenly among the h - and t -possibilities. Metaphorically, your *conditional* credence in h

²Technically, a probability space needs an ‘algebra’ F that picks out the subsets of W that π is defined over. I’ll usually let F be all subsets of W , so we can omit it; we’ll need it occasionally when dealing with infinite models.

³Reminder: technical terms and symbols are bolded when defined; they are collected in the Technical Glossary [todo]. A bolded symbol (e.g. ‘ \mathbf{P} ’) is always synonymous with its unbolded counterpart (e.g. ‘ P ’).

⁴*Non-negativity* says for all $q \subseteq W$, $\pi(q) \geq 0$. *Normality* says that $\pi(W) = 1$. And *additivity* says that if q and r are mutually exclusive, then the probability of their union (disjunction) is the sum of their individual probabilities: if $q \cap r = \emptyset$, then $\pi(q \cup r) = \pi(q) + \pi(r)$.

given f , $\pi(h|f)$, is the proportion of mud in the f -region that is also in the h -region. Formally, it's defined by the **ratio formula**: $\pi(h|f) = \frac{\pi(h \cap f)}{\pi(f)}$. The sensible conditional credence to have is $\pi(h|f) = 0.5$ —so the 0.5 credence assigned to f is evenly split into 0.25 on f_h and 0.25 on f_t .⁵

Since W is fine-grained enough to capture the distinctions you're uncertain about, this assignment of probability to worlds implicitly determines the probabilities of everything else you're uncertain about—such as the bias of the coin. The bias X is a **random variable**—a function from worlds $w \in W$ to numbers X_w that capture the value of X at world w . In our case, X takes value 0.5 at f_t and f_h (where the coin is fair), and 1 at \bar{f}_h (where the coin is double-headed). Thus we can use X to make claims about what world we're in. For instance, $\langle X = 0.5 \rangle$ —the claim that the bias of the coin equals 0.5—is equivalent to the claim $\{f_t, f_h\}$ that the coin is fair; meanwhile $\langle X = 1 \rangle$ is equivalent to the claim $\{\bar{f}_h\}$ that it's double-headed. In other words, when we assert that a variable takes a given numeric value x , we are implicitly asserting that we're in one of worlds in which it gets value x .⁶ Thus such claims inherit probabilities from the probabilities assigned to worlds: $\pi(X = 1)$ is the probability that π assigns to the bias X taking the value 1. What is that? Since $\langle X = 1 \rangle$ is the event $\{\bar{f}_h\}$, that means that $\pi(X = 1) = \pi(\{\bar{f}_h\}) = 0.5$.

Two important lessons. First, *there are many different names for a given set of worlds*. For example, we earlier used ' f ' to pick out the claim $\{f_t, f_h\}$ that the coin is fair, and therefore ' $\neg f$ ' picks out the claim $\{\bar{f}_h\}$ that the coin is *unfair*. But there are many other ways to pick out the latter event. Since the unfair coin always lands heads, $\neg f \& h$ also picks out the same event $\{\bar{f}_h\}$. Plugging through the logic, so does ' $\neg\neg\neg f$ ' (since $\neg\neg q$ is always equivalent to q). And given how we defined X , $\langle X = 1 \rangle$ *also* picks out the claim $\{\bar{f}_h\}$ that the coin is unfair. In essence, random variables are just convenient ways to pick out sets of worlds—claims about their values can always be 'unpacked' simply into claims about what world we're in.

Second: *all uncertainty is modeled as uncertainty about what world you're in*. Uncertainty about whether the coin will land heads is uncertainty about whether you're at a world where h true (f_h or \bar{f}_h), or where it's false (f_t). Uncertainty about the bias of the coin is uncertainty about whether you're at a world where X takes value 0.5 (f_t or f_h), or where it takes value 1 (\bar{f}_h). Generally: if you're uncertain about the value of a quantity Y , then there must be worlds $w, v \in W$ that you assign positive probability to in which it takes different values, $Y_w \neq Y_v$.

For example, let the variable N be the number of times I'll toss the coin before showing it to you. As we described things, at all the worlds w in our model, $N_w = 1$; therefore $\langle N = 1 \rangle$ is the trivial claim W , i.e. is true at all worlds and so gets probability 1. (And $\langle N = 2 \rangle$ is true at no worlds and gets probability 0.) If we want to model uncertainty about N , we had better include worlds where it can take different values—for instance, we might split each of our possibilities w into w_1 (where I toss it once) and w_2 (where I toss it twice), yielding the bigger outcome space $W = \{f_{t_1}, f_{h_1}, \bar{f}_{h_1}, f_{t_2}, f_{h_2}, \bar{f}_{h_2}\}$. Much of the art of probability theory is figuring out how to elegantly represent and manipulate bigger state spaces that can capture more uncertainty about more variables.

⁵Formally, the 'multiplication rule' says that $\pi(f \& h) = \pi(f) \cdot \pi(h|f) = 0.5 \cdot 0.5 = 0.25$. For those who want more background on mathematics of probability, I recommend Bertsekas and Tsitsiklis 2008—chapters 1 and 2 will teach you most of what need to know for this book. (No calculus needed.)

⁶Formally: $\langle X = x \rangle$ is the claim $\{w \in W : X_w = x\}$.

Crucial question: what if the quantity you're uncertain about, Y , is *itself* a probability—say, your future (or current!) credence that the coin is fair? Then it'd better be a variable that can take different values at different worlds—and we'd better find an elegant way to represent it.

4.3 Probability frames

4.3.1 Future Uncertainty

What will your future credence function be? That depends on what world you're in. As we said, you'll soon update on how the coin landed by reading what I wrote down—but which claim you'll condition on (h or t) obviously depends on how it landed. Let's introduce P^+ as a variable that picks out your posterior credence function (after seeing how it landed) in each world. Which distribution it picks out depends on what world you're in.

Suppose it lands heads—you're at f_h or \bar{f}_h . Then you'll update by **conditioning** your prior π on h , i.e. your posterior in any proposition q will be $\pi(q|h)$ —we can write this new probability function as $\pi(\cdot|h)$. By the ratio formula, this is equivalent to zero-ing out the $\neg h$ -possibilities, and 'renormalizing'—dividing by $\pi(h)$ —so that the remainder sums up to 1. In our case, it shifts you from the prior $\left(\begin{array}{c|ccc} f_t & f_h & \bar{f}_h \\ \hline 0.25 & 0.25 & 0.5 \end{array}\right)$ to $\left(\begin{array}{c|ccc} f_t & f_h & \bar{f}_h \\ \hline 0 & 1/3 & 2/3 \end{array}\right)$, leading you to be sure that the coin landed heads and $2/3$ -confident that the coin is double-headed.

Suppose the coin lands tails—you're at f_t . Then you'll update by conditioning on $\neg h$, shifting you from the prior $\left(\begin{array}{c|ccc} f_t & f_h & \bar{f}_h \\ \hline 0.25 & 0.25 & 0.5 \end{array}\right)$ to the posterior $\left(\begin{array}{c|ccc} f_t & f_h & \bar{f}_h \\ \hline 1 & 0 & 0 \end{array}\right)$. In other words, you become certain that it's the fair coin, since the double-headed one wouldn't have landed tails.

How can we efficiently encode all this information about your posterior? Since which world you're in determines what posterior you have, we can write it using a *stochastic matrix*—a list of stochastic vectors, where the row tells us what world you're in and the columns tell us what your posterior believes about what world you're in. In our case, your posterior is:

$$P^+ = \left(\begin{array}{c|ccc} & f_t & f_h & \bar{f}_h \\ \hline f_t & 1 & 0 & 0 \\ f_h & 0 & 1/3 & 2/3 \\ \bar{f}_h & 0 & 1/3 & 2/3 \end{array} \right)$$

This says: at world f_t , your posterior is certain that you're at f_t (first row); and at worlds f_h and \bar{f}_h , your posterior is certain the coin landed heads while being $1/3$ -confident that the coin is fair (you're at f_h) and $2/3$ -confident that it's unfair (you're at \bar{f}_h). We can thus think of your posterior P^+ is thus a vector-valued variable: given a world w , it yields a vector P_w^+ that encodes your posterior credence function at world w . In particular, $P_{f_t}^+ = (1, 0, 0)$, while $P_{f_h}^+ = P_{\bar{f}_h}^+ = (0, 1/3, 2/3)$.

Reflect on the differences between P^+ and π for a moment. π is a probability distribution that assigns numbers to worlds. P^+ is a *function* from worlds w to probability distributions P_w^+ that assign numbers to worlds. π is like a constant (like 7 or 2.4), and P^+ is like a variable (like X or Y). Call P^+ a **variable probability function**, on analogy with the distinction between a constant and a random variable. In philosopher's terminology: π is a rigid designator, while P^+ is

a description. The difference is crucial. Since P^+ can vary from world to world, it can be an object of uncertainty. Indeed, this is why P^+ *must* be a description: your prior is uncertain what your posterior will be, and so P^+ must vary amongst worlds that your prior leaves open—in our case, π is uncertain whether $P^+ = (1, 0, 0)$, or instead $P^+ = (0, 1/3, 2/3)$. In contrast: although π assigns different values to different worlds, it doesn't *vary* from world to world. $\pi(f_t) = 0.25$ is simply *true* in this model—it's not true in some worlds and false in others. Unlike P^+ , the values of π can't be objects of uncertainty.

This distinction is familiar in the case of numbers versus random variables. Suppose I in fact own 17 spoons, but you're uncertain whether I own 16 or 11 or 42. Then we (obviously) can't write down the claim that I own 17 spoons by writing ' $\langle 17 = 17 \rangle$ '—for that's a trivial claim that you know is true. Instead, we need to introduce a *variable* S which is a function from worlds w to numbers S_w capturing the number of spoons I have in world w . We can think of '17' as a rigid designator for a number, while S is a *description* of a number. Then we can write ' $\langle S = 17 \rangle$ ' to pick out the (nontrivial) set of worlds where I own 17 spoons: $\langle S = 17 \rangle := \{w \in W : S_w = 17\}$. This is the sort of thing you can be uncertain about, because you can be unsure whether you're in a world that's included in the set.

The same thing is happening with π and P^+ . π is a constant that picks out a given distribution; you can't be uncertain whether ' $\pi = (1, 0, 0)$ '. Meanwhile, P^+ is a variable that picks out different distributions in different worlds; therefore, you can be uncertain whether $P^+ = (1, 0, 0)$.

This highlights the importance of using notation that tracks which things you can be uncertain about, i.e. which things vary from world to world. There's no perfect system, but here's ours. When lowercase letters (x, y, z) are used for numbers, they are constants that are fixed and known. When uppercase letters (X, Y, S, \dots) are used for numbers ('quantities'), they are random variables that can be objects of uncertainty. ' X ' can mean something like 'the bias of the coin (whatever it is)', while ' x ' always means something like '0.5'. When we take a variable like X and saturate it with a world $w \in W$ (which are always picked out rigidly), we get a *number* X_w , which we could equally write $X(w)$ in standard function notation. Notice that while X is a variable, X_w is a rigid designator for the value that X takes at w . In our above model, $X_{f_t} = 0.5 = X_{f_h}$, while $X_{\bar{f}_h} = 1$. Finally, we can use variables like X to specify sets of worlds that get assigned probabilities—for example $\langle X = 0.5 \rangle = \{f_t, f_h\}$, and hence $\pi(X = 0.5) = \pi(\{f_t, f_h\}) = 0.25 + 0.25 = 0.5$.⁷

Similarly, when lowercase Greek letters (π, δ, η, \dots) are used for probability distributions, they are constants that are fixed and known. When *uppercase* Roman letters (P, P^+, C, \dots) are used for probability functions, they are variables (descriptions of probability functions) that can be objects of uncertainty. ' P^+ ' can mean something like 'your future credence function (whatever it is)', while ' π ' always means something like '(0.25, 0.25, 0.5)'. When we take a variable probability function P^+ and saturate it with a world $w \in W$, we get a particular probability distribution P_w^+ . Notice that while P^+ is a variable, P_w^+ is a rigid designator for the probability distribution that P^+ picks out at w . In our above model, $P_{f_t}^+$ is $(1, 0, 0)$, while $P_{f_h}^+$ and $P_{\bar{f}_h}^+$ both equal $(0, 1/3, 2/3)$. Finally, we can use variables like P^+ to specify sets of worlds that get assigned probabilities—for example, $\langle P^+ = (1, 0, 0) \rangle$ is the set of worlds where your posterior is certain it's at f_t , i.e. $\{f_t\}$, and hence

⁷When angle-bracket events like $\langle X = 0.5 \rangle$ are embedded inside functions, we omit the brackets for readability: $\pi(X = 0.5)$, rather than $\pi(\langle X = 0.5 \rangle)$.

$\pi(P^+ = (1, 0, 0)) = \pi(f_t) = 0.25$.

Using P^+ , we can specify all the claims about your posterior that we care about. For example, we can saturate it with a proposition f to get a random variable $P^+(f)$ which picks out your posterior that the coin is fair: at each world w , it takes the value that P_w^+ assign to f , i.e. $P_w^+(f)$. Thus $P^+(f)$ equals 1 when you see the coin land tails (i.e. $P_{f_t}^+(f) = 1$), and equals $1/3$ when you see the coin land heads (i.e. $P_{f_h}^+(f) = 1/3$ and $P_{\bar{f}_h}^+(f) = 1/3$). And thus we can use $P^+(f)$ to pick out events that you can be uncertain about—for instance, $\langle P^+(f) = 1 \rangle$ is equivalent to the claim that the coin will land tails, $\{f_t\}$, and hence $\pi(P^+(f) = 1) = \pi(f_t) = 0.25$.

Given a random variable like X (the bias of the coin), an important quantity is π 's **expectation** of that variable, $\mathbb{E}_\pi(X)$. This is a weighted average of the possible values of X , with weights determined by how likely π thinks they are. Formally, $\{x_1, \dots, x_n\}$ are the possible values of X that π leaves open, then $\mathbb{E}_\pi(X) = \sum_{x_i} \pi(X = x_i) \cdot x_i$. In our case, π is 50-50 between the bias of the coin being 0.5 ($X = 0.5$) and being 1 ($X = 1$), so it's expectation is the average of these $\mathbb{E}_\pi(X) = 0.75$. An equivalent way to calculate expectations—which will be more familiar to some—is to sum across *worlds*: $\mathbb{E}_\pi(X) = \sum_{w \in W} \pi(w) \cdot X_w$. In our case, that's $\pi(f_t) \cdot 0.5 + \pi(f_h) \cdot 0.5 + \pi(\bar{f}_h) \cdot 1 = 0.25(0.5) + 0.25(0.5) + 0.5(1) = 0.75$.

Notice that you needn't be confident that the value of the variable is close to the expectation. Indeed, your expectation of X can be a value that you know it won't take—the coin's bias is either 0.5 or 1, yet π 's expectation of this is 0.75. Instead, your expectation of X is what you're confident the *average* of a bunch of independent copies of X would be. If I repeatedly grabbed coins at random that were each 50-50 likely to be 0.5- or 1-biased toward heads, then you're confident that the *average* bias of those coins would be around 0.75.

Since your prior π has opinions about the possible values of your posterior P^+ , it thereby has an expectation of them. For instance, π 's expectation of your posterior credence that the coin is fair is $\mathbb{E}_\pi(P^+(f)) = \pi(P^+(f) = 1)(1) + \pi(P^+(f) = \frac{1}{3})(\frac{1}{3}) = 0.25(1) + \frac{3}{4}(\frac{1}{3}) = 0.5$. Huh. That equals your prior that the coin is fair, $\pi(f) = 0.5$. Let's try another one: your prior that the coin *either lands tails, or is unfair* is $\pi(\{f_t, \bar{f}_h\}) = 0.75$. And your expectation of your posterior in that is $\mathbb{E}_\pi(P^+(\{f_t, \bar{f}_h\})) = \pi(f_t)(1) + \pi(f_h)(\frac{2}{3}) + \pi(\bar{f}_h)(\frac{2}{3}) = 0.25(1) + 0.25(\frac{2}{3}) + 0.5(\frac{2}{3}) = 0.75$.

This is no accident. A core feature of Standard Bayesian models—sometimes called the 'Martingale' principle—is that your prior in a claim q always equals your expectation of your posterior in q : $\pi(q) = \mathbb{E}_\pi(P^+(q))$. Martingale turns out to be the crux of why Standard-Bayesian updates are unbiased and tend to converge to the truth. And in many ways, the crux of this book is that under ambiguity, Martingale will inevitably fail. But hold that thought.

4.3.2 Current Uncertainty

Consider what we've done. Your future credences are objects of uncertainty. To represent this properly, we must model them with a **variable probability function** P^+ —a function from worlds w to probability functions P_w^+ . In doing so, we implicitly define claims about your future probability function as subsets of W . For instance, $\langle P^+(f) = 1/3 \rangle$ is the set of worlds where your future credence is $1/3$, i.e. the set of worlds where the coin lands heads, i.e. $\{f_h, \bar{f}_h\}$.

Notice something. We did this so that your prior, π could have well-defined opinions about your

posterior. But in doing so, we've implicitly guaranteed that *your posterior* P^+ also has opinions about your posterior. After all, your posterior has opinions about whether the coin landed heads, i.e. about whether $\{f_h, \bar{f}_h\}$ is true, i.e. about whether $\langle P^+(f) = 1/3 \rangle$ is true, i.e. about whether your posterior in heads is $1/3$. If you're in the world where the coin landed tails (f_t), then your posterior is certain that the coin landed tails: $P^+ = \begin{pmatrix} f_t & f_h & \bar{f}_h \\ 1 & 0 & 0 \end{pmatrix}$, hence you assign zero credence to the claim that your posterior is $1/3$ in the coin being fair: $P_{f_t}^+(P^+(f) = 1/3) = 0$. Meanwhile, if you're in a world where the coin landed heads (f_h or \bar{f}_h), then your posterior is certain that the coin landed heads: $P^+ = \begin{pmatrix} f_t & f_h & \bar{f}_h \\ 0 & 1/3 & 2/3 \end{pmatrix}$, hence you assign credence 1 to the claim that your posterior is $1/3$ in the coin being fair: $P_{f_h}^+(P^+(f) = 1/3) = 1$, and likewise for $P_{\bar{f}_h}^+$. Thus your posterior is always certain about whether your credence in the coin being fair is $1/3$.

The same is true for any other claim about your future credence function, specified by the form $\langle P^+ \text{ has property } \phi \rangle$. Whether P^+ has such a property is determined by which probability function (stochastic vector) P^+ is; and in the worlds where $P^+ = (1, 0, 0)$ (namely, f_t), P^+ is certain of this; while in the worlds where $P^+ = (0, 1/3, 2/3)$, P^+ is certain of *this*. An easy way to see this is that the above stochastic matrix is 'block diagonal': the different blocks where P^+ takes different values ($\{f_t\}$ versus $\{f_h, \bar{f}_h\}$) each assign credence 0 to each other.

Upshot: the way we've defined your posterior has implicitly made its opinions *clear*. Formally:

Clarity. P^+ 's opinion about q are *clear* at w iff it is certain of what it's opinion in q is:

There is a number x such that $P_w^+(P^+(q) = x) = 1$.

P^+ is *clear* (period) iff it's opinion in every claim q is clear at every world w .

So defined, your posterior P^+ is clear in the above model.

What about your *prior*? As we defined it, your prior is a constant distribution π , meaning that it can't be an object of uncertainty. But that's an omission. In this case you *do* have opinions about your prior—at the prior time, you are certain that you are 50-50 between the coin being fair or double-headed, and at the posterior time you *remember* that at the prior time you were 50-50 between the coin being fair or double-headed.

How can we represent facts about your prior as potential objects of uncertainty? The same way: we introduce a variable probability function P for your prior: a function from worlds w to probability functions P_w that captures your *prior* opinions at w . In this case you are certain of what your prior is, so it doesn't vary across worlds: at all worlds w , $P_w = \pi$, i.e. $P_w = (0.25, 0.25, 0.5)$. We can represent this fact explicitly with another stochastic matrix. Here they are together:

$$P = \left(\begin{array}{c|ccc} & f_t & f_h & \bar{f}_h \\ \hline f_t & 0.25 & 0.25 & 0.5 \\ f_h & 0.25 & 0.25 & 0.5 \\ \bar{f}_h & 0.25 & 0.25 & 0.5 \end{array} \right) \qquad P^+ = \left(\begin{array}{c|ccc} & f_t & f_h & \bar{f}_h \\ \hline f_t & 1 & 0 & 0 \\ f_h & 0 & 1/3 & 2/3 \\ \bar{f}_h & 0 & 1/3 & 2/3 \end{array} \right)$$

In this case, your prior is clear, and moreover doesn't vary across worlds, so the extra information is redundant. But in many cases—even under clarity—it's not redundant. For instance, suppose that although I *in fact* told you I was going to grab either a fair- or double-headed coin, I *might've instead* told you that I was definitely going to grab the double-headed coin—call that

possibility d_h . Then the two possible priors you might've had are $P = \left(\begin{array}{c|ccc} f_t & f_h & \bar{f}_h & d_h \\ \hline 0.25 & 0.25 & 0.5 & 0 \end{array} \right)$, and $P = \left(\begin{array}{c|ccc} f_t & f_h & \bar{f}_h & d_h \\ \hline 0 & 0 & 0 & 1 \end{array} \right)$. To represent someone (say, Teddy) who was unsure which prior you had, we'd need to allow your prior to vary across worlds.⁸ Likewise if we wanted to allow your future self to *forget* what your prior was, even if it was clear to you at the time.

Once we treat your prior P as a variable, it—just like your posterior P^+ —automatically has opinions about itself. For instance, you are certain that your prior in the coin landing tails is 0.25: at each world w , $P_w(P(t) = 0.25) = 1$. In this model, P is clear just like P^+ is.

Once your prior is a variable, we need to specify what the actual world is so that we know what *actual* distribution you have. We can do that by selecting a world $\mathbf{a} \in W$ —I'll reserve the fancy a , \mathbf{a} , for the actual world. Your *actual* prior credence function (rigidly specified) is $\mathbf{P}_{\mathbf{a}}$, your actual posterior (rigidly specified) is $\mathbf{P}_{\mathbf{a}}^+$.

In sum: suppose we want to represent a Bayesian agent who begins with some uncertainty, goes through an update, and has opinions about what opinions they started with and ended up with. Then what we need is a **probability frame** (W, \mathbf{a}, P, P^+) consisting of an outcome space W , an actual world $\mathbf{a} \in W$, and two variable probability functions: P for the prior, and P^+ for the posterior.⁹

4.4 Standard Bayesianism

Probability frames are an extremely broad class of Bayesian models. They (or close variants of them) have been used by many authors in many fields to tractably model the potentially-infinite hierarchy that emerges when we model beliefs about beliefs (about beliefs about...)¹⁰ They build in only the assumption that you are certain to have a (precise) probability function, defined over all relevant propositions—where 'relevant propositions' include claims about your first- and higher-order opinions about the propositions in question (see §6.5.1). Since P and P^+ can be *any* function from worlds to distributions, they don't build in any assumptions about how accurate your priors are, or how you update them, or how confident you are about what priors you have. This flexibility is nice mathematically; but to make them useful in practice, we'll need to impose further constraints.

A basic constraint is that your credences are *self-aware*: they always assign positive probability to having the values they actually have. Formally:

Self-Awareness. P is *self-aware* iff never rules out having its actual opinions:

For all w , $P_w(P = P_w) > 0$.

⁸In the case, we'd represent your prior with the matrix $P = \left(\begin{array}{c|ccc} f_t & f_h & \bar{f}_h & d_h \\ \hline f_t & 0.25 & 0.25 & 0.5 & 0 \\ f_h & 0.25 & 0.25 & 0.5 & 0 \\ \bar{f}_h & 0.25 & 0.25 & 0.5 & 0 \\ d_h & 0 & 0 & 0 & 1 \end{array} \right)$. If Teddy were 50-50 on whether I told you it was double-headed, or was selected randomly, then his prior would be $\tau = (0.125, 0.125, 0.25, 0.5)$.

⁹In uncountably-infinite cases, we'll also need an algebra F . When omitted, F is always the power set of W .

¹⁰See e.g. Kripke 1963; Hintikka 1962; Harsanyi 1967; Gärdenfors 1975; Aumann 1976, 1999; Gaifman 1988; Geanakoplos 1989; Samet 1999, 2000; Williamson 2000, 2008, 2014; Schervish et al. 2004; Lasonen-Aarnio 2015; Salow 2018, 2019; Das 2022; Dorst 2020a; Dorst et al. 2021; Levinstein 2023; Levinstein and Spencer 2024.

This definition illustrates the importance of distinguishing constants from variables. P is a variable, that can take different values; P_w is a constant which always specifies the probability function that P picks out at w . So what this definition is saying is that at each world w , your probability function (P_w) leaves open that your probability function (P) has the values it actually has. The example from §4.3.2 was self-aware; a footnote gives an example that is not.¹¹

In this book, I'll restrict attention to frames where both P and P^+ are self-aware. Equivalently, if you are certain about some claim about your own probabilities, then that claim is true: if $P_w(P(q) = x) = 1$, then $P_w(q) = x$. A sufficient condition for this is that P and P^+ always assign positive credence to the world they're in.¹²

Standard Bayesianism is an additional constraint on frames. It can be formalized in several equivalent ways, but here's a natural one. First, we assume that your prior P is self-aware and clear. Second, we assume that there is some **partition** \mathcal{Q} that captures your new evidence—think of it like a question that you're learning the answer to. In our above example, \mathcal{Q} was the question, 'How will the coin land?', with possible answers $\{tails, heads\}$, i.e. $\{\{f_t\}, \{f_h, \bar{f}_h\}\}$. More generally, a partition is a way of dividing W into mutually exclusive and collectively exhaustive sets, so each world is in exactly one 'cell' of the partition. \mathcal{Q}_w is the cell that w is in. Formalized:

Standard Bayesianism: A frame $(W, \mathfrak{a}, P, P^+)$ is *Standard-Bayesian* iff:

- i) P is self-aware and clear;¹³ 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 , $P_w^+(q) = P_w(q|\mathcal{Q}_w)$.

Standard Bayesianism is a constraint on probabilistic updating that a given model might satisfy or violate. It is so-called because, in practice, 99% of Bayesian models satisfy its assumptions. This is for two reasons. First, almost all Bayesian models presuppose that the prior is clear and assigns positive credence to having its actual values. This is often done implicitly, by referring to the prior with a constant π rather than a variable P —since π 's values don't vary across worlds, every probability function in the model is certain of what π (i.e. your prior) is. Second, for some good reasons (and some bad), it is standardly assumed that when you update your credences, there must be some *signal* that drove the update. The question, 'Which signal (if any) did you receive?' always forms a partition, so it's natural to identify the signal simply with the partition-cell in which you received that signal. Finally, there are compelling arguments that under such conditions, the optimal way to update your credences is via conditioning (e.g. Greaves and Wallace 2006; Huttegger 2013). Hence Standard Bayesianism.

In any Standard-Bayesian model, your prior and posterior are clear—just like in the coin-flip example:

Fact 4.4.1. If a finite probability frame is Standard Bayesian, then P and P^+ are both clear.

¹¹If $P = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0.4 & 0.6 \\ 0 & 0.4 & 0.6 \end{pmatrix}$, then P is not self-aware. Letting w be the first world, $P_w = (0, 0.5, 0.5)$. But P_w only leaves open the second two worlds, where $P = (0, 0.4, 0.6)$. So $P_w(P = (0, 0.5, 0.5)) = P_w(P = P_w) = 0$.

¹²I.e. for all w , $P_w(w) > 0$ and $P_w^+(w) > 0$.

¹³Equivalently: P is always correctly certain about itself—self-awareness rules out P being certain but mistaken.

Is it worth stating with a rider that \mathcal{Q} just partitions the communicating classes? That might make various equivalences later easier, since won't need to assert reflexivity

(I restrict attention to finite frames for simplicity. All proofs are in the Appendix, §4.6.)

Why? By definition, your prior is clear. Thus it is certain of what its conditional probabilities are—for all possible bits of evidence \mathcal{Q}_v you might get, if $P_w(q|\mathcal{Q}_v) = x$, then $P_w(P(q|\mathcal{Q}_v) = x) = 1$. And since \mathcal{Q} is a partition, you are always certain of what evidence you received: if at world w you get evidence \mathcal{Q}_w and you leave open world v , then v and w are in the same partition-cell, so $\mathcal{Q}_w = \mathcal{Q}_v$. Thus in a Standard Bayesian model, you're always sure what your prior conditional credences were, and you're always sure *which* conditional credence function your posterior should match—so you're always sure of what your posterior is.

In a way, Standard Bayesianism is the theory of rational updating we get if we assume that (1) your priors and posteriors are always clear, and (2) you update in a minimally-rational way, given that fact. Let's see why.

4.4.1 Why Partitions?

Old Bayesians will be familiar with partitional updating, but it can look strange the first time you see it. And for those coming from certain philosophical backgrounds (Williamson 2000), it's natural to worry about what partitionality presupposes. Indeed, there's a lively debate in the philosophical literature about how 'externalist' Bayesians—those who think we can rationally be mistaken about what our evidence is—should respond to apparent cases of non-partitional evidence.¹⁴ I'm largely going to avoid that debate, other than to make a few points that we need for our discussion.

The most important one is this: *under clarity, all rational updating is partitional updating*. That is: if you're certain that your rational current- and future-self know what their credences are, then so long as you're (sure to be) minimally rational, the update must be Standard-Bayesian and proceed via conditioning on a partition. This point has been made in a variety of different guises by a variety of different authors over the years;¹⁵ but let's state it clearly for our setting.

Start simple. Suppose I announce that if the coin lands tails, I'll tell you how it landed; but if it lands heads, I won't tell you anything. Why doesn't that make your evidence non-partitional? In particular: when the coin lands tails, you learn $\{f_t\}$, but when it lands heads you learn nothing—your evidence is the trivial proposition $\{f_t, f_h, \bar{f}_h\}$. Since these two propositions overlap (at f_t), and they are the possible evidence you might receive, your possible evidence *doesn't* form a partition. Right?

Obviously not. Those two propositions *aren't* the possible pieces of evidence you might receive. When a Bayesian talks about what evidence they might receive, they're usually talking about the *total* (new) evidence—the logically strongest proposition that you learn. And in this setting: when I tell you nothing, you learn *more* than nothing. After all, you realize that I told you nothing, and you know that I would tell you nothing only if the coin landed heads. Thus you can infer that it *did* land heads, and so should condition on $\{f_h, \bar{f}_h\}$. In other words, the possible total bodies of evidence you receive in this context are the simply $\{tails, heads\}$, i.e. $\{\{f_t\}, \{f_h, \bar{f}_h\}\}$ —

¹⁴ See e.g. Geanakoplos 1989; Hild 1998; Samet 1999; Williamson 2000, 2014, 2019; Horowitz 2014; Schoenfield 2017a, 2018; Salow 2018, 2019; Ahmed and Salow 2019; Gallow 2019, 2021; Isaacs and Russell 2022; Zendejas Medina 2022; Isaacs and Levinstein 2024; Zhang and Rand 2025.

¹⁵E.g. Skyrms 1980, 1997, 2006; Hild 1998; Samet 1999; Huttegger 2014; Salow 2018

I've simply concocted an annoyingly-indirect way to provide you with that information.¹⁶ Call this the *partition-recovery* strategy.

Partition-recovery doesn't require you to know the setup. Suppose *unknownst to you*, I decide to flip the coin and say 'heads' if the coin landed tails, but sit there like a doofus, saying nothing, if the coin landed heads. Then do your possible bits of evidence fail to be partitional? No! For if the coin lands heads, you'll learn *that I said, 'heads'*; and if the coin lands tails, you'll learn *that I said nothing* (and, indeed, *that I sat there like a doofus*). Those two propositions are exclusive; and if we list out *all* the possible things I might've done that you would've learned (*Kevin said, 'heads', Kevin said nothing, Kevin ate the coin,...*), we'll form a partition.

What partition-recovery *does* require is that you are certain of what you learned—at least in the sense that you are certain what effect updating has had on your credence function. You need to be able to be (rationally) certain that what you learned was that I said nothing, or said 'bloop', or ate the coin, or what not. 'What you learned' is a question that always forms a partition—so if you always know the answer to it, you have a partition to update on.

?? should I mention 'learn' and factivity?

Here's the kicker: under clarity, you *always* know the answer to it. For 'What posterior did you end up with?' always forms a partition—indeed, it forms a partition that captures all the possible effects that the learning experience might have on your opinions. If you have clarity, your posterior will always know the answer to that question. So we can always do partition-recovery by treating your posterior P^+ as the result of conditioning on the true cell of the partition associated with the question, *What is P^+ ?*¹⁷ Thus under clarity you can always do partition-recovery: by observing what happened to your credences, you can back-infer what evidence you must've received.

4.4.1.1 The Monty Hall Puzzle

Let's make this concrete. Many probability puzzles work by implicitly assuming clarity, but failing to make proper use of that fact in specifying what evidence you received. Here's one (cf. Gong et al. 2022; Seidenfeld et al. 2024).

Anna, Bob, and Chiara are the last contestants left on the Island of Trials. Only one of them will win crown—the possible outcomes are a (Anna wins), b (Bob wins), and c (Chiara wins). The judges have already decided, but no information has been given yet. Bob's clear prior is that each contestant is $1/3$ -likely to win. But he wants to be optimistic, so he devises a scheme to increase his credence. He privately asks the host which of the *other two* contestants will lose the crown. Since he always knows that (at least) one of them will lose, he reasons, this won't reveal any illicit information.

¹⁶There are different ways to use the term 'evidence', and some of the disagreement about non-partitional updating may be terminological. Schoenfeld (2017b) understands 'exogenous' evidence to be (metaphorically) 'what the world flings at you'—in economists' terminology, the signal you receive. And indeed the two signals you might receive might well be either 'heads' or nothing—or indeed, they might be nonsense; perhaps I'll just say either 'bleep' or 'bloop'. Nonsense never forms a partition. But if you know this setup—namely, that you will be told nothing (or hear 'bloop') iff the coin lands heads—then you can infer that nothing (or 'bloop') implies that the coin landed heads.

¹⁷Is there something circular about getting to P^+ by conditioning on what P^+ will be? Recall that $\langle P^+ = \delta \rangle$ is one among many names for the same proposition. If your posterior is δ iff you get signal s_i , then $\langle P^+ = \delta \rangle$ is the same proposition as *you received s_i* —so when we say that you condition on $\langle P^+ = \delta \rangle$ that's just a shorthand for saying that you condition on *receiving s_i* . Likewise in more complicated setups.

The host shrugs, and tells him ‘Chiara didn’t win’. So Bob conditions on $\neg c$, and therefore becomes 50-50 that either he or Anna won. But wait. If the host had instead told him ‘Anna didn’t win’, Bob would’ve conditioned on $\neg a$, and so become 50-50 that either he or Chiara would win. Either way, he would’ve increased his credence from $1/3$ to $1/2$ that *he’d* win! What went wrong?

Bob violated partitionality, despite having clarity—so he didn’t condition on his total evidence. When told ‘Chiara didn’t win’, he conditions on $\neg c$; when told ‘Anna didn’t win’, he conditions on $\neg a$. But these two propositions are compatible—they are both true if *Bob* wins, b , which is why the strategy raises his credence in b .

This is a variant of the Monty Hall problem. The standard diagnosis is that Bob failed to condition on the fact that the host *told* him Chiara (or Anna) didn’t win. But a more-general way to put the point is that Bob failed to make use of information he had about which posterior he would end up with in different possibilities.

We can see this clearly if we write down a probability frame for Bob’s update. At first, we might say there are just three possibilities— a , b , and c . But remember: worlds need to be fine-grained enough to answer all relevant questions, *including* what Bob’s posteriors are. Yet b doesn’t determine how he updates—it’s consistent with him being told either ‘Chiara didn’t win’ or ‘Anna didn’t win’. Thus we need to split b into two possibilities: one in which he’s told it wasn’t Chiara ($b_{\bar{c}}$), and the other in which he’s told it wasn’t Anna ($b_{\bar{a}}$). Suppose that Bob thinks, conditional on *him* winning, that it’s 50-50 whether the host will say ‘Anna didn’t win’ or ‘Chiara didn’t win’, so he’s $\frac{1}{3} \cdot 0.5 = \frac{1}{6}$ in each of $b_{\bar{a}}$ and $b_{\bar{c}}$. Then we can write doesn’t his update in a probability frame. Subscripting each world by what he’s told, $W = \{a_{\bar{c}}, b_{\bar{c}}, b_{\bar{a}}, c_{\bar{a}}\}$. Let’s suppose the actual world is $\mathbf{a} = a_{\bar{c}}$; we’ll bold this in the diagrams. Then Bob’s priors and posteriors are as follows:

$$P = \left(\begin{array}{c|cccc} & a_{\bar{c}} & b_{\bar{c}} & b_{\bar{a}} & c_{\bar{a}} \\ \hline \mathbf{a}_{\bar{c}} & 1/3 & 1/6 & 1/6 & 1/3 \\ b_{\bar{c}} & 1/3 & 1/6 & 1/6 & 1/3 \\ b_{\bar{a}} & 1/3 & 1/6 & 1/6 & 1/3 \\ c_{\bar{a}} & 1/3 & 1/6 & 1/6 & 1/3 \end{array} \right) \quad P^+ = \left(\begin{array}{c|cccc} & a_{\bar{c}} & b_{\bar{c}} & b_{\bar{a}} & c_{\bar{a}} \\ \hline \mathbf{a}_{\bar{c}} & 1/2 & 1/4 & 1/4 & 0 \\ b_{\bar{c}} & 1/2 & 1/4 & 1/4 & 0 \\ b_{\bar{a}} & 0 & 1/4 & 1/4 & 1/2 \\ c_{\bar{a}} & 0 & 1/4 & 1/4 & 1/2 \end{array} \right)$$

He starts out $(1/3, 1/6, 1/6, 1/3)$ over the four possibilities, and so is $1/6 + 1/6 = 1/3$ -confident that he’ll win. If he’s told ‘Chiara didn’t win’ (top two rows), he conditions on $\neg c$, zeroing out the last possibility and shifting to $P^+ = (1/2, 1/4, 1/4, 0)$. If he’s told, ‘Anna didn’t win’ (bottom two rows), he conditions on $\neg a$, shifting to $P^+ = (0, 1/4, 1/4, 1/2)$. Either way, he ends up $1/4 + 1/4 = 1/2$ -confident that he’ll win.

Now, this is a *mathematically* coherent update—Bob’s prior and posterior are always probability functions. It’s non-partitional since the things he might become certain of— \bar{c} and \bar{a} —don’t form a partition; they overlap in the b -worlds. (Compare the ‘clock-like’ updates in e.g. Salow 2018.)

The problem with this update is not mathematical. The problem is that it doesn’t accurately represent Bob’s situation. It implicitly implies that his posterior opinions are *ambiguous*—in particular, that he doesn’t know what he was told, and so doesn’t know what his posterior is. Why? He’s told different things in $a_{\bar{c}}$ and $b_{\bar{a}}$ (‘Chiara didn’t win’ and ‘Alice didn’t win’, respectively), and so has different posteriors in the two worlds $((1/2, 1/4, 1/4, 0)$ and $(0, 1/4, 1/4, 1/2)$, respectively). The

model therefore implies that if $a_{\bar{c}}$ is the actual world—so Bob was told, ‘Chiara didn’t win’—Bob leaves open that he’s at $b_{\bar{a}}$, i.e. he leaves open that he was told ‘Alice didn’t win’. But that’s wrong. In our description of the case, Bob is certain of what the host tells him, and of the credence function it induces. In conditioning on either $\neg a$ or $\neg c$, he’s failed to use his total evidence.

How to fix the model? Notice that Bob’s possible posteriors partition the space: at $a_{\bar{c}}$ and $b_{\bar{c}}$ he has $(1/2, 1/4, 1/4, 0)$, while at $b_{\bar{a}}$ and $c_{\bar{a}}$ he has $(0, 1/4, 1/4, 1/2)$. Thus if Bob can *introspect* what his credences are, he can learn whether he’s in $\{a_{\bar{c}}, b_{\bar{c}}\}$ or instead in $\{b_{\bar{a}}, c_{\bar{a}}\}$. If he were to condition on this further information in each world, his posteriors would then be:

$$P^{++} = \left(\begin{array}{c|cccc} & a_{\bar{c}} & b_{\bar{c}} & b_{\bar{a}} & c_{\bar{a}} \\ \hline a_{\bar{c}} & 2/3 & 1/3 & 0 & 0 \\ b_{\bar{c}} & 2/3 & 1/3 & 0 & 0 \\ b_{\bar{a}} & 0 & 0 & 1/3 & 2/3 \\ c_{\bar{a}} & 0 & 0 & 1/3 & 2/3 \end{array} \right)$$

Thus, since Bob has clarity about his posteriors, he can recover partitionality by introspecting on what happened to them. In doing so, we remove the paradox: whether the host tells him ‘Chiara didn’t win’ or ‘Alice didn’t win’, Bob remains $1/3$ that he will win. The local reason is that being told ‘Chiara didn’t win’ rules out won of the ways Bob could win (namely, if Bob wins and is told ‘Alice didn’t win’); but it doesn’t rule out any of the ways Alice could win. Vice versa with being told ‘Alice didn’t win’. The global reason is that conditioning a clear prior on a partition always generates updates that satisfy strong ‘Reflection’ principles, as we’ll see in a moment.

4.4.1.2 Clarity Guarantees Partitionality

The trick we just pulled with Bob can be pulled *whenever* your prior and posterior are clear. Under minimal rationality conditions, clarity guarantees that updates are partitional.

We’ll see one version of this in Chapter 6, which will give an extended discussion of the rationality assumption at the heart of Bayesianism. This assumption—which I’ll call the *value of rationality*—says that an update is rational only if it is *valuable* in the sense that it can be expected to lead to better beliefs and decisions, no matter your priorities. If P and P^+ are clear and self-aware, then P values P^+ if and only if P^+ is the result of conditioning P on the true cell of a partition—if and only if the update is a Standard-Bayesian one. This result goes some way to explaining why Standard Bayesianism is so standard—under the implicit assumption of clarity, it follows from standard rationality assumptions.

But we don’t need assumptions nearly as strong as the value of rationality in order to derive partitionality from clarity. One way to do so is to assume that there’s always *some* true proposition you condition on, though the set of such propositions might not form a partition. Under clarity, they must do so.¹⁸

A more general approach is to assume merely that rational updating is minimally reliable, in the sense that if you’re *certain* that a rational update will lead you to have a high estimate for a

¹⁸Precisely: if (W, α, P, P^+) is clear and self-aware, and for all w there is a proposition e_w such that $w \in e_w$ and $P_w^+(\cdot) = P_w(\cdot|e_w)$, then the update is Standard Bayesianism. See Salow 2018.

quantity, then you should already have a high estimate for it. Precisely:

No Foregone Conclusions: If you're certain that your posterior estimate for X will be high, you should already have a high estimate for X .

Formally: if $P_w(\mathbb{E}_{P^+}(X) \geq x) = 1$, then $\mathbb{E}_{P_w}(X) \geq x$.¹⁹

Violating No Foregone Conclusions would exhibit a clear lack of trust for your future self. For instance, suppose $\mathbb{E}_{P_a}(X) = 0.5$, but $P_a(\mathbb{E}_{P^+}(X) \geq 0.6) = 1$: you currently estimate the value of X to be 0.5, but you're certain that after updating you'll estimate it to be at least 0.6. Then despite being certain that your future-self will have an estimate of at least 0.6 for X , you still have a lower estimate for it. You treat your future-self much the same way you treat some benighted fool who you know you disagree with.

No Foregone Conclusions is a much weaker principle, than the value of rationality. But *under clarity*—and the background assumption that you are not certain of relevant falsehoods—it suffices to prove Standard Bayesianism (and hence suffices to prove the value of rationality). Say that P is **reflexive** iff it is not certain of any falsehoods.²⁰ Then:

Fact 4.4.2. Suppose a finite probability frame is clear and P is reflexive. Then if it validates²¹ No Foregone Conclusions, it is Standard Bayesian.

Upshot: under clarity, all reasonable roads lead to Standard Bayesianism. I think this helps to explain its hegemonic status within the behavioral and social sciences: if you assume that opinions can be represented by probabilities and that those probabilities are clear, Standard Bayesianism is the only sensible game in town.

4.4.1.3 Non-partitional updates?

Under clarity, all sensible updating is partitional updating. This book is about how we can and should deny clarity. So you might reasonably assume that we'll be doing all sorts of crazy, non-partitional Bayesian updates.

We won't. That's not because I think non-partitional updates are irrational. In the ongoing debate about what 'externalism' (or, as I prefer, 'rational ambiguity') implies for updating (footnote 14), I side with those who think it makes non-partitional updating often rational. Indeed, earlier versions of the arguments in this book presupposed as much (Dorst 2023b).

Nevertheless, in this book all updating will be partitional. Ambiguity will slip in via the prior: if you are unsure what your priors are, and you do a partitional update, you'll usually be uncertain what your posteriors are. The reasons are both strategic and explanatory.

Strategically: I want to show that debates about partitional updating are largely orthogonal to the most interesting questions raised by ambiguity. Everyone agrees that when what you learn is the clear answer to a fixed question (when your evidence is partitional), you should update by

¹⁹Recall that P_w is a constant distribution which P^+ is a variable one. Therefore $\mathbb{E}_{P_w}(X)$ is a constant number (P_w 's expectation for X), while \mathbb{E}_{P^+} is a variable (P^+ 's expectation for X , whatever it might be).

²⁰Formally, for all q , $\langle P(q) = 1 \rangle \rightarrow q$ is true at all worlds. Equivalently, for all w , $P_w(w) > 0$.

²¹A frame **validates** a principle iff the principle is true at all worlds, for all instantiations of its variables. In this case, that means that for all worlds w , variables X , and numbers x , if $P_w(\mathbb{E}_{P^+}(X) \geq x) = 1$, then $\mathbb{E}_{P_w}(X) \geq x$.

This is violated in the 'agreeing to disagree' AB models. Weaken it somehow?

Mention alternatives:
- 'ambiguity'
- from econ
- imprecise probability
- ??

conditioning your prior on the answer. Yet it turns out that when your priors are ambiguous, such updating will lead rational people’s beliefs to evolve in very different ways than we’re used to—they will exhibit biases, violate deference principles, and systematically fail to converge to the truth. These are the phenomena of most interest.

Explanatorily: updating by conditioning on the true cell of a partition is a well-understood and well-behaved process. We *know* when and why you should do these updates, and we can induce them in the laboratory: if you know that all you’ll learn is whether the coin landed heads or tails, then you should update by conditioning on the true cell of $\{heads, tails\}$. Meanwhile, non-partitional updating is (currently) the wild west: we have no agreed-upon answers to when (or if!) it should happen, and exactly how. It is not currently a tractable way to test for (ir)rationality.

That’s not to say it can’t be done. I consider it a live and interesting research question how properly constrained non-partitional updating might affect (and, I bet, exacerbate) the results of this book. But our plate is full enough as it is.

4.4.2 Reflection Implies Clarity

There is another natural route to Standard Bayesianism. It is to directly impose strong constraints on how your prior P defers to your posterior P^+ . The mathematical fact is that if we impose a version of the famous ‘Reflection’ principle (van Fraassen 1984), Standard Bayesianism follows. It’s often concluded that, since rationality *does* demand Reflection, rationality demands Standard Bayesianism.²²

But that is the wrong conclusion. It equally follows that you can satisfy the Reflection principle (for all propositions) only if *you are certain that your posterior will be clear*. But reasonable people often have ambiguous posteriors. So for any notion of ‘rationality’ applicable to agents like us—who should *not* be certain that their posterior will be clear—the correct conclusion is that Reflection must and should fail. We will spend much of the next two chapters discussing why this is so, and what principle should replace Reflection. But for now, let’s just say what the result is.

Reflection is a constraint on what your prior thinks of your posterior. In particular, it says that your prior *strongly defers* to your posterior in the following sense: conditional on your posterior having a given set of credences, your prior should (conditionally) adopt those credences.

This can get confusing. Let’s see how it works in the coin case—your prior and possible posteriors are repeated below. Your prior is 50-50 on the claim that the coin is fair: $P_a(f) = 0.5$. Your prior is unsure whether your posterior will be 100%-confident of this (if the coin lands tails), or 1/3-confident of this (if the coin lands heads). Obviously, your prior can’t just match your posterior—it doesn’t know what your posterior is!

$$P = \left(\begin{array}{c|ccc} & f_t & f_h & \bar{f}_h \\ \hline f_t & 0.25 & 0.25 & 0.5 \\ f_h & 0.25 & 0.25 & 0.5 \\ \bar{f}_h & 0.25 & 0.25 & 0.5 \end{array} \right) \qquad P^+ = \left(\begin{array}{c|ccc} & f_t & f_h & \bar{f}_h \\ \hline f_t & 1 & 0 & 0 \\ f_h & 0 & 1/3 & 2/3 \\ \bar{f}_h & 0 & 1/3 & 2/3 \end{array} \right)$$

²²E.g. Gaifman 1988; Samet 1999; Van Fraassen 1999; Skyrms 2006; Weisberg 2007; Briggs 2009; Huttegger 2015.

But your prior can *conditionally* match your posterior. That is: your prior's *conditional* credences, given a condition that says what your posterior is, can match your posterior. Since $\langle P^+(f) = 1/3 \rangle$ is a proposition your prior has opinions about, then your prior thereby has conditional opinions, given it. In particular, your credence that the coin is fair, given that your posterior is $1/3$ -confident of it, is well defined: $P_a(f|P^+(f) = 1/3)$. What is it? By the ratio formula, it is the ratio of the probability that the coin is fair *and* your posterior is $1/3$, divided by the probability that your posterior is $1/3$.²³ The former probability is that of the coin being fair and landing heads, f_h , which is 0.25. And the latter probability is that of the coin landing heads, $\{f_h, \bar{f}_h\}$, i.e. 0.75. So the *conditional* probability is $P_a(f|P^+(f) = 1/3) = \frac{0.25}{0.75} = 1/3$. In other words: you don't know what your posterior will be; but *conditional* on your posterior having credence $1/3$ in f , you infer that the coin must land heads—and therefore drop your (conditional) credence to $1/3$ to match your posterior. Meanwhile, conditional on your posterior having credence 1 in f , you infer that the coin landed tails, and so raise your (conditional) credence to 1 to match your posterior: $P_a(f|P^+(f) = 1) = 1$.

Reflection is the constraint that this *always* holds. For any claim q , conditional on your future-self being x -confident of q , you should be (conditionally) x -confident of it: $P_a(q|P(q) = x) = x$. The most natural formalization of Reflection generalizes this constraint to learning about your posterior's opinions in multiple propositions at once:

Reflection: Conditional on your posterior having credence x in q and y in p , have (conditional) credence x in q .

Formally: $P_w(q|\langle P^+(q) = x \rangle \& \langle P^+(p) = y \rangle) = x$.

Say that a prior P_w *reflects* a posterior P^+ iff it obeys this principle for all propositions and numbers such that the conditional probability is well-defined. Say that P is **reflexive** iff it is not certain of any falsehoods.²⁴ The mathematical result is that if P_w reflects both the prior P and the posterior P^+ , then the update is a Standard-Bayesian one²⁵:

Fact 4.4.3. In a finite probability frame in which P is reflexive, if for all w , P_w reflects both P and P^+ , then the frame is Standard-Bayesian.

This might seem like a result that arrives at Standard Bayesianism without presupposing clarity. But it's not: the Reflection principle implicitly implies that your posterior must be clear—you can't reflect a posterior if you leave open that it might be ambiguous:

Fact 4.4.4. If P_w reflects P^+ in a finite probability frame, then P_w is certain that P^+ is clear.²⁶

Indeed, the proof of Fact 4.4.3 goes *via* clarity. It is the fact that Reflection implies clarity that allows us to establish that P^+ can always be thought of as conditioning P on the answer to 'What posterior did I end up with?', as discussed above.

²³I.e. $P_a(f|P^+(f) = 1/3) = \frac{P_a(f \& \langle P^+(f) = 1/3 \rangle)}{P_a(\langle P^+(f) = 1/3 \rangle)}$.

²⁴Formally, for all q , $\langle P(q) = 1 \rangle \rightarrow q$ is true at all worlds. Equivalently, for all w , $P_w(w) > 0$.

²⁵Versions of this have been proven by, for example, Hild 1998; Samet 1999; Van Fraassen 1999 and Salow 2018.

²⁶Versions of this have been proven for different formulations of Reflection, e.g. Gaifman 1988; Hall 1994; Lewis 1994; Elga 2013; Dorst 2020a; Dorst et al. 2021.

Explain that the converse is true too.

Why does Reflection imply clarity? The intuitive reason is this (see Elga 2013). If P^+ is ambiguous, then it doesn't know its own values. Thus it might assign credence x to q *only because* it isn't sure what its credences are—were it informed that $\langle P^+(q) = x \rangle$, that would provide new information which might *change* P^+ 's credence in q . But your conditional credence $P_w(q|P^+(q) = x)$ is (roughly) the credence you'd adopt in q if you were *told* that $\langle P^+(q) = x \rangle$. If $P^+(q)$ is ambiguous, then when you learn that $\langle P^+(q) = x \rangle$, you know more than P^+ does. Thus you shouldn't necessarily adopt the P^+ 's credence.

In particular, suppose you were to learn that your future-credences were higher-order uncertain in a particular way: say, your future credence assigns 0.6 to p , but only is 40%-confident of that fact: $\langle P^+(p) = 0.6 \rangle$ and $\langle P^+(P^+(p) = 0.6) = 0.4 \rangle$. What credences should you have, upon learning this conjunction of facts? Substituting $\langle P^+(p) = 0.6 \rangle$ in for q , Reflection implies that you should be only 40%-confident that $\langle P^+(p) = 0.6 \rangle$, for an instance of Reflection is

$$P_w\left(P^+(p) = 0.6 \mid \langle P^+(P^+(p) = 0.6) = 0.4 \rangle \ \& \ \langle P^+(p) = 0.6 \rangle\right) = 0.4$$

But that's wrong. The other conjunct *tells* you that $\langle P^+(q) = 0.6 \rangle$ —thus you should be *certain* of this, not 40%-confident of it. Indeed, that follows from the ratio formula. Put the other way: if you are to obey Reflection, you must never assign positive credence to a conjunction like this. That is, you must be certain that your posteriors P^+ will be clear.

We will spend much of the following two chapters getting clearer on why this is happening, and what we should replace Reflection with under ambiguity. But the point, for now, is simply that Reflection does *not* give us a route to Standard Bayesianism without assuming clarity. Instead, Reflection implicitly *requires* clarity.

This is yet another way in which clarity can sneak in the back door. Anytime someone writes down a Bayesian update, if they (implicitly or explicitly) presuppose that the prior reflects the posterior (for all propositions), then they have implicitly imposed clarity.

4.5 Its Limits

Standard Bayesianism is used throughout the behavioral sciences—both to predict people's behavior, and as a yardstick to judge their rationality. The gold-standard way to show that a given 'bias' is rational is to show that Standard Bayesians would exhibit it, and show that they'd do so in a way similar to how real people do. Economics and cognitive science are full of papers doing exactly that.²⁷

But there are sharp limits on what Standard Bayesianism can rationalize. For it implies clarity—and rational thinking under clarity looks very different from the way you and I think.

²⁷ *Examples:* Biased generation and assimilation of evidence (Feeney et al. 2000; Hahn and Oaksford 2007; Jern et al. 2014; Benoit and Dubra 2019; Gershman 2019; Henderson and Gebharter 2021) HartmannXX, seeking confirmation (Oaksford and Chater 1994, 2003; Navarro and Perfors 2011; Hahn and Harris 2014), asking non-diagnostic questions (Feeney et al. 2008; Crupi et al. 2009), discriminating against those who are harder to interpret (Phelps 1972; Aigner and Cain 1977; Cornell and Welch 1996; Hedden 2021), being miscalibrated on tricky questions (Juslin 1994; Juslin et al. 2000; Moore and Healy 2008; Moore 2020; Dorst 2023a), ranking conjunctions as more 'likely' than their conjuncts (Levi 1985; Dorst and Mandelkern 2022), committing the gambler's fallacy (Rabin and Vayanos 2010; Dorst 2025), having selective memories (Wilson 2014), and so on.

Since Standard Bayesians have clarity, they won't exhibit noisy judgments (Ch. 2). After all, they are *certain* of what their own opinions are. So if you ask them how likely they think it is that *I own a dozen spoons* (d), i.e. what $P(d)$ is, then they are certain of the answer: there is an x such that $P(P(d) = x) = 1$. For them, this question is analogous to the question, 'How likely do you think it is that the 27th digit of pi is between 1–6?'. They know, of a particular number (in that case, 0.6) that it accurately captures their credence. Rational people who are certain of the answer and want to be helpful will reliably give the answer. (For example, perhaps they can draw a single sample from their distribution and report its value.) But *real* people who want to be helpful are unreliable and noisy when answering questions like, 'How likely do you think it is that I own a dozen spoons?' Standard Bayesianism can't rationalize that.²⁸

Relatedly, Standard Bayesians will not exhibit self-doubt. Of course, they will often be uncertain whether the action they are taking will work out for the best. But they will never be uncertain which action is best *by their own lights*. To state that more carefully, let's assume that our agent is certain that the rational thing for them to do, given their (rational) opinions, is to take an option that maximizes expected value.²⁹ That is, if their options are $\{o_1, \dots, o_n\}$ and their utility function is U —so that $U(o_i)$ is a random variable saying how much value option o_i would realize at each world—suppose that they are certain that they should take an option o_i that maximizes $\mathbb{E}_P(U(o_i))$. Since they have clarity, they know what P is, and therefore know what $\mathbb{E}_P(X)$ is for any variable X —including $U(o_1)$, $U(o_2)$, etc. Thus they are certain what expected value each option has, and hence are certain which options maximize expected value. So they are certain of which options are best by their own lights.

But real people have self-doubt. Anyone who's made a difficult career or personal decision (i.e. anyone) knows that often we are genuinely unsure what is (expectedly) best by our own lights. When you're deciding which graduate school to go to, you are genuinely unsure what you expect is likely to happen if you choose MIT versus Princeton. So you are unsure whether the decision you're making is the one you expect to be best.³⁰ Standard Bayesianism can't rationalize this.

So Standard Bayesians don't think and act like we do: they don't exhibit noise or self-doubt. But our focus is on the ways we think and act—the biases—that lead to polarization and radicalization. Part III will explain at length why Standard Bayesian models can't explain these biases, at least not in the way that real people exhibit them. For now, I'll just give an informal sketch of why.

First, Standard-Bayesian models can't rationalize *hindsight bias*. This is the tendency for learning things to increase your estimate for your prior probability in them. ('I knew it all along!') For a

²⁸Of course, you can write down Standard-Bayesian models and hard-code in procedures that generate noise. Theorists do that all the time. For example, perhaps you have a Standard Bayesian that does outcome-sampling to report their credences. My claim is not that such procedures are impossible, but that they *make no sense*—since our Bayesian is certain of their own credences, they have better available reporting options. For example: instead of outcome-sampling they can estimate-sample, always accurately reporting their credence, as discussed in Ch. 2. Under clarity, estimate-sampling always elicits your true credence, while outcome-sampling elicits a noisy-indicator of your true credence. Thus, under clarity, estimate-sampling always has higher expected accuracy (and greater expected utility) than outcome-sampling.

²⁹Or maximizes risk-weighted expected value (Quiggin 1982; Tversky and Kahneman 1992; Buchak 2013). Or other decision rules. What matters is just that they are certain, of some function from opinions and options to actions, that it is the rational procedure to follow in their situation.

³⁰Psst—it's MIT.

Standard Bayesian, at the initial time they have clarity—they are certain of what their prior probability is. And they always update by conditioning, which preserves certainties—if initially they are certain that $P(q) = 0.6$, then at all later times they will also be certain of that. So Standard Bayesians won't exhibit hindsight bias the way real people do.

Second, Standard-Bayesian models can't rationalize *confirmation bias*. This is the tendency to seek and interpret evidence in a way that predictably strengthens your beliefs. Chapter 8 will argue that the best way to formalize this is as a violation of the 'Martingale principle': you exhibit confirmation bias on q when you expect your search for evidence to, on average, increase your credence in q .³¹ Standard Bayesians will won't exhibit confirmation bias in this way, for their updates *always* satisfy the Martingale principle—it follows from the fact that they obey Reflection toward their future self.

Third, Standard-Bayesian models can't rationalize *predictable polarization*. This is the tendency to predictably fail to converge to the truth, instead falling into persistently and sharply disagreeing camps. Chapter 9 will show that various well-known 'convergence to the truth' results prevent Standard Bayesians from predictably polarizing like this.

Fourth, Standard-Bayesian models can't rationalize *predictable overconfidence*. This is the tendency to predictably become miscalibrated, wherein judgments that you are (say) 90%-confident in are only true 60% of the time. Chapter 10 will show that, since Standard Bayesians obey the Reflection principle toward their future-selves, they can't predictably become miscalibrated in the ways that real people do.

Zooming out, the point is that the 'optimistic' results of the standard rational models are deeply misleading—both about rationality, and about psychology, politics, and society.

Let's see what we need to do to fix those models.

4.6 Appendix[†]

4.6.1 Practice Problems

[Answer key under construction]

1. I grabbed a coin at random from a bucket containing 30 $\frac{1}{3}$ - and 60 $\frac{2}{3}$ -biased (towards heads) coins. I'm going to flip it once and update on the flip. Write down a probability frame that captures my prior and posteriors about (i) the first flip of the coin, and (ii) the bias of the coin.
2. There are 3 doors, A, B and C. There is a car behind one of them and a goat behind the other two. Suppose I'm $\frac{1}{3}$ -confident the car is behind each door.
I pick door B. I knew Monty was going to pick a door that had a goat behind it to show to me. He picks door A. What are my updated probabilities for the car being behind doors B and C? Draw a probability frame explaining why.

³¹Formally, your expectation for your posterior credence in q is higher than your prior in q : $\mathbb{E}_P(P^+(q)) > P(q)$.

3. Suppose P^0 is (clear and) certain that a fair coin will be tossed twice. P^1 is the result of conditioning P^0 on how the first toss landed, and P^2 is the result of conditioning P^0 on how the second toss landed. Draw a probability frame (W, P^0, P^1, P^2) representing this scenario. Show that P^0 obeys Reflection toward both P^1 and P^2 .

Generalize this: show that if P^0 is clear and P^1 and P^2 are each the result of updating P^0 on two (potentially different) different partitions \mathcal{Q}_1 and \mathcal{Q}_2 , then P^0 obeys Reflection toward both P^1 and P^2 .

Gallow 2018 shows that no one can “serve two epistemic masters” in the sense of (1) obeying Reflection toward each of their (uncertain) probabilities, (2) leaving open that they’ll disagree, and (3) always being a weighted average between their two opinions when you learn that they disagree. Which of these premises fails in the two-coin-toss case? What lesson should we draw from this?

4. Show that the Reflection principle implies Martingale: if P_a reflects P^+ —so for all x , and q : $P_a(q | \langle P^+(q) = x \rangle) = x$ —then for any q , $\mathbb{E}_{P_a}(P^+(q)) = P_a(q)$.
5. Assume higher-order-certainty certainty, i.e. that P_w is certain of what P is and is not certain of. (So if $P_w(v) > 0$, then P_w and P_v assign positive probability to the same worlds.) Show that if P_w obeys *Self-Martingale*—i.e. for all q , $\mathbb{E}_{P_w}(P(q)) = P_w(q)$ —then P_w is clear. (This is a simplified version of the result in Samet 2000.)
6. Consider a synchronic version of Reflection: $P_w(q | P(q) = x) = x$. Show that if this holds for all x and q at a given world w , then P_w must be clear. (Hint: focus on the probabilities P_w assigns to various worlds $v \in W$. the worlds.)
7. Give an example of a frame that is self-aware (i.e. for all w , $P_w(P = P_w) > 0$), but is not reflexive (so there is a world w such that $P_w(w) = 0$).

4.6.2 Proofs

This section contains proofs of all results stated in the chapter. I restrict attention to finite probability frames for simplicity—many of the results generalize to infinite ones as well, with proper care taken for handling probability-0-but-possible events.

Fact 4.4.1. If a finite probability frame is Standard Bayesian, then P and P^+ are both clear.

Proof. By definition, the prior P is clear. Taking a world w in which P_w^+ is well-defined, we’ll show that P_w^+ is clear too. By definition, there is a unique partition-cell \mathcal{Q}_w that w is in such that for all v in \mathcal{Q}_w , $P_v^+(\cdot) = P_v^+(\cdot | \mathcal{Q}_w)$. Since P_w^+ is well-defined, $P_w(\mathcal{Q}_w) > 0$. Now consider any v such that $P_w^+(v) > 0$. Since $P_w^+(\mathcal{Q}_w) = 1$, it follows that $P_v^+(\cdot) = P_v(\cdot | \mathcal{Q}_w)$. Meanwhile, since P_w is clear, there is a unique π such that $P_w(P = \pi) = 1$; and since conditioning preserves certainties, likewise $P_w^+(P = \pi) = 1$. Thus for any v that P_w^+ leaves open, $P_v = \pi$ and $P_v^+(\cdot) = P_v(\cdot | \mathcal{Q}_w) = \pi(\cdot | \mathcal{Q}_w)$. Thus for any event $q \subseteq W$, we have that $P_w^+(P^+(q) = \pi(q | \mathcal{Q}_w)) = 1$. So P^+ is clear. \square

Fact 4.4.2. Suppose a finite probability frame is clear and self-aware, and P is reflexive. Then if it validates No Foregone Conclusions, it is Standard Bayesian.

Proof. We know that P is clear and self-aware, so what we need to show is that there is a partition \mathcal{Q} such that for all v , $P_v^+(q) = P_v(q|\mathcal{Q}_w)$. Consider the partition of the possible posteriors: $\mathcal{Q} = \{\langle P^+ = \pi_1 \rangle, \dots, \langle P^+ = \pi_n \rangle\}$. Since P is reflexive, every world v is seen by some world w (there is a w such that $P_w(v) > 0$). So take an arbitrary w and v such that $P_w(v) > 0$. It'll suffice to show that $P_v^+(\cdot) = P_v(\cdot|P^+ = P_v^+)$, for—recalling that $\langle P^+ = P_v^+ \rangle$ is another name for the proposition $\langle P^+ = \pi_i \rangle$ for the π_i equal to P_v^+ —this will show that P^+ is obtained by conditioning P on the true cell of \mathcal{Q} . By clarity and self-awareness, $P_w(P^+ = P_w) = 1$, so since $P_w(v) > 0$, $P_v = P_w$. Thus it suffices to show that $P_w(\cdot|P^+ = \pi_v) = P_v^+(\cdot)$.

By No Foregone Conclusions, if $P_w(\mathbb{E}_{P^+}(X) \geq x) = 1$, then $\mathbb{E}_{P_w}(X) \geq x$ for all random variables X and numbers x . It follows by the hyperplane separation theorem³² that P_w is in the convex hull of the set of possible posteriors, π_1, \dots, π_n . That is to say: there is a set of non-negative numbers $\lambda_i \geq 0$ summing to 1 such that for all q , $P_w(q) = \sum_i \lambda_i \pi_i(q)$. We use this, combined with the clarity and self-awareness of P^+ , to show that $P_w(\cdot|P^+ = \pi_v) = P_v^+(\cdot)$. Take an arbitrary q . By the ratio formula, $P_w(q|P^+ = \pi_v) = \frac{P_w(q \cap \langle P^+ = \pi_v \rangle)}{P_w(P^+ = \pi_v)}$. Since $P_w(v) > 0$, the denominator is nonzero. Applying the above convex decomposition, we have:

$$P_w(q|P^+ = \pi_v) = \frac{\sum_i \lambda_i \pi_i(q \cap \langle P^+ = \pi_v \rangle)}{\sum_i \lambda_i \pi_i(P^+ = \pi_v)}$$

Since all the candidate π_i s are clear and self-aware, they all assign probability 1 to $\langle P^+ = \pi_i \rangle$ and 0 to $\langle P^+ = \pi_j \rangle$ for $j \neq i$. Thus all the terms are zero except the one for π_v with weight λ_v :

$$= \frac{\lambda_v \pi_v(q \cap \langle P^+ = \pi_v \rangle)}{\lambda_v \pi_v(P^+ = \pi_v)}$$

Canceling λ_v , the denominator is 1 and the numerator is $\pi_v(q)$, i.e. $P_v^+(q)$, as desired. \square

Fact 4.4.3. In a finite probability frame in which P is reflexive, if for all w , P_w reflects both P and P^+ , then the frame is Standard-Bayesian.

Proof. By Fact 4.4.4 below, since P reflects both P and P^+ , for all w P_w is certain that P and P^+ are clear. Since P is reflexive, for all w , $P_w(w) > 0$; hence anything that P_w is certain of is true at w , meaning P and P^+ are both clear at all worlds. Moreover, since P reflects P^+ at all worlds, anything P^+ is certain of is true. (If not, there's a world v where $P^+(v) = 0$. Since $P_v(v) > 0$, it follows that $P_v(v|P^+(v) = 0)$ is well-defined and not equal to 0, violating Reflection.)

What we must show is that there is a partition \mathcal{Q} such that for all w , $P_w^+(\cdot) = P_w(\cdot|\mathcal{Q}_w)$. Consider the partition induced by the possible posteriors: $\mathcal{Q} = \{\langle P^+ = \pi_1 \rangle, \dots, \langle P^+ = \pi_n \rangle\}$. Suppose, for reductio, there is a w such that $P_w^+(\cdot) \neq P_w(\cdot|P^+ = \pi_w)$, where $\pi_w = P_w^+$. By definition $w \in \langle P^+ = \pi_w \rangle$, so by reflexivity $P_w(P^+ = \pi_w) > 0$, so $P_w(\cdot|P^+ = \pi_w)$ is well-defined. WLOG say that $P_w^+(q) > P_w(q|P^+ = \pi_w)$.

³²Ahem. This is a place where I don't know how to avoid the fancy math. Google it, or ask ChatGPT.

Hm. Can I omit reflexivity by just doing the argument with P_v ? Probably need it to show positive probability of denominator.

Let $q' := q \cap \langle P^+ = \pi_w \rangle$. Since P^+ is clear and only certain of truths, we know that $P_w^+(P^+ = \pi_w) = 1$. Thus $P_w^+(q') = P_w^+(q)$. Similarly, $P_w(q|P^+ = \pi_w) = P_w(q'|P^+ = \pi_w)$. So we have that $P_w^+(q') > P_w(q'|P^+ = \pi_w)$. Let's say $P_w^+(q') = x > 0$. Notice that for any $v \notin \langle P^+ = \pi_w \rangle$, $P_v^+(P^+ = \pi_w) = 0 \neq x$ (by clarity and accuracy of P^+ 's certainties), so that $P_v^+(q') = 0$. Hence $\langle P^+(q) = x \rangle$ is true at all and only the worlds in the set $\langle P^+ = \pi_w \rangle$. Thus $P_w(q'|P^+ = \pi_w) = P_w(q'|P^+(q') = x)$. By Reflection, this value must equal x ; but by the above, we know that it doesn't equal x . Contradiction. \square

Fact 4.4.4. If P_w reflects P^+ in a finite probability frame, then P_w is certain that P^+ is clear.

Proof. Suppose—for reductio—that P_w reflects P^+ but leaves open that P^+ is ambiguous. Thus there's some p such that P_w leaves open a possibility in which $\langle P^+(p) = y \rangle$ and $\langle P^+(P^+(p) = y) \rangle < 1$. At any such possibility, P^+ assigns some positive credence x to the hypothesis that $P^+(p)$ isn't equal to y : $\langle P^+(P^+(p) \neq y) = x \rangle$, with $x > 0$. Now substitute this latter claim into Reflection. With q set to $\langle P^+(p) \neq y \rangle$, Reflection would imply that $P_w(P^+(p) \neq y | \langle P^+(P^+(p) \neq y) = x \rangle \& \langle P^+(p) = y \rangle)$ is well-defined and equal to $x > 0$. Thus it follows that

$$P_w(P^+(p) = y | \langle P^+(P^+(p) \neq y) = x \rangle \& \langle P^+(p) = y \rangle) < 1.$$

But that contradicts the ratio formula, which implies that for any proposition r , $P_w(P^+(p) = y | r \& \langle P^+(p) = y \rangle) = 1$. Contradiction. \square