Chapter 10

Accuracy Arguments

The previous two chapters considered Representation Theorem and Dutch Book arguments for probabilism. We criticized both types of argument on the grounds that they begin with premises about practical rationality—premises that restrict a rational agent's attitudes towards acts, or towards sets of bets. Probabilism hopes to establish that the probability axioms are requirements of *theoretical* rationality on an agent's credences, and it's difficult to see how one could move from practical premises to such a theoretical conclusion.

This chapter considers arguments for probabilism that begin with explicitly epistemic premises. The basic idea is that, as a type of representational attitude, credences can be assessed for *accuracy*. We are used to assessing other doxastic attitudes, such as binary beliefs, in terms of their accuracy. A belief in the proposition P is accurate if P is true; disbelief in P is accurate if P is false. A traditional argument moves from such accuracy assessments to a rational constraint on beliefs—in particular, to Chapter 1's Belief Consistency norm requiring an agent's beliefs to be logically consistent. The argument begins by noting that if a set of propositions is logically inconsistent, there is no (logically) possible world in which all those propositions are true. (That's the definition of logical inconsistency.) So if an agent's beliefs are logically inconsistent, she's in a position to know that at least some of them are inaccurate. Moreover, she can know this *a priori*—without knowing any contingent truths. Since an inconsistent set contains falsehoods in *every* possible world, no matter which world is actual her inconsistent belief set misrepresents how things are.[1]

There are plenty of objections one might make to this argument—starting with its assumption that beliefs have a teleological "aim" of being accurate.[2]

But I present the argument here because it offers a good template for the arguments for probabilism to be discussed in this chapter. Whatever concerns you have about the Belief Consistency argument above, keep them in mind as you consider our accuracy arguments for probabilism. Accuracy arguments for probabilism also face an additional challenge not confronted by the accuracy argument for Belief Consistency. When we talk about binary beliefs, it's pretty obvious what it takes for such a doxastic attitude to be accurate. But generalizing the notion of accuracy to credences proves challenging.

When an agent has a nonextreme credence in proposition P, it would be strange to refer to that credence as either accurate or inaccurate full-stop. (It's not as if P has some intermediate *degree* of truth, and a degree of belief in P is accurate just in case those numerical degrees match.) So just as we moved from classificatory to quantitative doxastic attitudes in Chapter 1, we will move from a classificatory to a quantitative concept of accuracy. We will consider various numerical measures that have been proposed over the years for gauging just *how* accurate a particular credence (or set of credences) is. We will begin with historical "calibration" approaches that measure the accuracy of credences by comparing them to frequences. Yet we will fairly quickly reject calibration in favor of the more contemporary "gradational accuracy" approach.

Gradational accuracy uses "scoring rules" to assess the accuracy of credence distributions. Among the many possible scoring rules are a particular class called the "strictly proper scoring rules," which have been favored historically for reasons we will describe. If we rely on strictly proper scoring rules, we can produce an argument for probabilism similar to the Belief Consistency argument above: an agent whose credences violate the probability axioms will be able to see that this decreases those credences' accuracy in every possible world. Yet the resulting argument seems to sneak probabilism into its premises in a question-begging way; we will have to consider whether it can be reformulated to remove this circularity.

Accuracy-based arguments have been offered for a number of Bayesian norms in addition to probabilism, such as the Principal Principle, the Principle of Indifference, etc. (See Further Readings.) We will close this chapter with an argument for Conditionalization based on minimizing expected future inaccuracy. Yet this argument has the same drawback as Dutch Strategy arguments for Conditionalization; it ultimately fails to establish any truly *diachronic* norms.

10.1 Measuring accuracy

10.1.1 Accuracy as calibration

In Section 5.2.1 we briefly considered a putative rational principle for matching one's credence that a particular outcome will occur to the frequency with which that outcome occurs. In that context, the match was supposed to be between one's credence that outcome B will occur and the frequency with which *one's evidence* suggests B occurs. But we might instead assess an agent's credences relative to *actual frequencies* in the world: If events of type A actually produce outcomes of type B with frequency x, an agent's credence that a particular A-event will produce a B-outcome is more accurate the closer it is to x.

Now imagine that an agent managed to be perfectly accurate with respect to the actual frequencies. In that case, she would assign credence 2/3 to outcomes that occurred 2/3 of the time, credence 1/2 to outcomes that occurred 1/2 of the time, etc. Or—flipping this around—propositions to which she assigned credence 2/3 would turn out to be true 2/3 of the time, propositions to which she assigned credence 1/2 would turn out to be true 1/2 of the time, etc. This approach to accuracy—getting the frequencies right, as it were—generates the notion of

Calibration: A credence distribution over a finite set of propositions is perfectly calibrated when, for any x, the set of propositions to which the distribution assigns credence x contains exactly fraction x of truths.

For example, suppose your weather forecaster comes on television every night and reports her degree of confidence that it will rain the next day. You might notice that every time she says she's 20% confident of rain, it rains the next day. In that case she's not a very accurate forecaster. But if it rains on just about 20% of those days, we'd say she's doing her job well. If exactly 20% of the days on which she's 20% confident of rain turn out to have rain (and exactly 30% of the days on which she's 30% confident... etc.), we say the forecaster is **perfectly calibrated**. Calibration is an initially plausible way to gauge accuracy.[3]

I've defined only what it means to be *perfectly* calibrated; measures can be designed to assess comparative degrees of calibration among distributions falling short of the ideal.[4] But all the good and bad features of calibration as accuracy can be understood by thinking just about perfect calibration. First, the good: van Fraassen (1983) and Abner Shimony (1988) both argue

for probabilism by showing that in order for a credence distribution to be embeddable in larger and larger systems with calibration scores approaching perfection, that credence distribution must satisfy the probability axioms. This would be a powerful argument, if it weren't that calibration also has bad features as a measure of accuracy.

Consider two agents, Sam and Diane, who assign the following credence distributions over propositions X_1 through X_4:

	X_1	X_2	X_3	X_4
Sam	1/2	1/2	1/2	1/2
Diane	1	1	1/10	0

Now suppose that propositions X_1 and X_2 are true, while X_3 and X_4 are false. Look at the table and ask yourself whose credences intuitively seem more accurate.[5]

I take it the answer is Diane. Yet Sam's credences are perfectly calibrated—he assigns credence 1/2 to all four propositions, exactly half of which are true—while Diane's credences are not. This is an intuitive flaw with measuring accuracy by calibration.

The same point can be made in a slightly different way by considering the plight of a weather forecaster whose job depends on her forecasts' being perfectly calibrated over a four-day span.[6] Suppose that on each of the first three nights, she expressed a 75% confidence in rain and it rained the next day. Tonight, the final night, she looks at her radar images and sees nary a cloud for hundreds of miles around. She is certain it will not rain tomorrow. Yet she also knows that if she goes on the air and reports a 0% confidence in rain, she will wind up less than perfectly calibrated for the four days. On the other hand, if she reports a 75% confidence in rain tomorrow, she will get a perfect calibration score (its having rained on exactly 75% of the days for which she reported a 75% confidence). Assessing the forecaster's reports according to their calibration encourages her to misrepresent her own credences—and the import of her evidence—on the air.

Perhaps there's something suspicious about assessing an agent's *reports* for accuracy as opposed to her credences themselves. Imagine the forecaster's boss could somehow measure her credences themselves (perhaps with a fancy galvanometer?) and hire or fire her based on their accuracy. Then on the fourth night the forecaster will desperately *wish* she had different credences than the ones she has, in fact, carefully formed on the basis of her available evidence. Assessing the forecaster's credences on the basis of calibration makes her evidence-based credences *unstable*—by the forecaster's own lights, she thinks she would do better (with respect to calibration) if she

assigned different degrees of belief than what she actually does. Yet there's nothing wrong with the forecaster's credences; they're a perfectly rational response to her evidence. The problem is with the calibration method of assessing accuracy; it's making a particular credence distribution look sub-optimal when in fact that distribution is perfectly rationally permissible (if not required!).[7]

One could make various moves here in an attempt to save calibration as a measure of accuracy. For instance, calibration scores are less easily manipulable if we measure them only in the long-run. But then there's a question about assessing credences in non-repeatable events, and soon we're assessing not actual long-run calibration but instead hypothetical calibration in the limit. Before long, we've made all the desperate moves used to prop up the frequency theory of probability (Section 5.1.1), and run into all the same problems.

The correct response here is the same as it was with the frequency theory: Instead of employing a notion that emerges only when events are situated in a larger collective, we find a notion that can be meaningfully applied to single cases considered one at a time (like propensity). Looking back at Sam and Diane, our intuitive judgment that Diane is globally more accurate than Sam arises from local judgments that she was more accurate than him on each individual proposition. If you knew only the truth-value of X_1, you could still have said that Diane was more accurate than Sam on that one proposition. Clearly our accuracy intuitions can be applied piece-wise—to one credence at a time.

10.1.2 Gradational accuracy and scoring rules

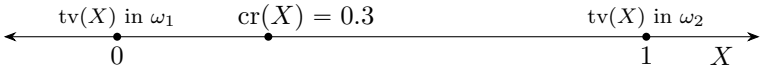
We will now develop a new measure of the accuracy of a credence distribution. It will first measure the accuracy (actually, inaccuracy) of each individual credence assignment—considered one at a time—then combine those measurements into a global inaccuracy score for the entire distribution.

Our guiding idea will be that inaccuracy is distance from truth. To gauge how far an agent's credence $\mathrm{cr}(X)$ in proposition X is from the truth-value of X, we'll need a way to express that truth-value as a number. A natural approach lets 1 stand for truth and 0 stand for falsehood. Just as we have a distribution cr reflecting the agent's credences in propositions, we'll have another distribution tv reflecting the truth-values of those propositions. Distribution tv will assign numerical values to the propositions in $\mathcal{L}$ such that $\mathrm{tv}(X) = 1$ if X is true and $\mathrm{tv}(X) = 0$ if X is false.[8]

Once we have distribution cr representing the agent's credences and dis-

tribution tv representing the truth, we want to measure how far apart these distributions are from each other. (The farther cr is from tv, the more inaccurate the agent's credences are.) Again, our measure of inaccuracy will start by examining one proposition at a time. So we need a function $i(\mathrm{cr}(X), \mathrm{tv}(X))$ that tells us for each individual proposition X how far $\mathrm{cr}(X)$ is from $\mathrm{tv}(X)$.

It's helpful to visualize what we're doing here. Consider the following number line representing values a distribution might assign to X:



Here I've imagined that your credence in X is 0.3. What is the value of $\mathrm{tv}(X)$? That depends what the world is actually like. We can imagine two possible worlds here, ω_1 and ω_2. In ω_1, X is false, so $\mathrm{tv}(X) = 0$. In ω_2, X is true and $\mathrm{tv}(X) = 1$.

Our inaccuracy-measuring function $i(\mathrm{cr}(X), \mathrm{tv}(X))$ is going to gauge how far your credence in X is from its truth-value. A clear *desideratum* on i is **truth-directedness**: i should decrease as $\mathrm{cr}(X)$ approaches $\mathrm{tv}(X)$. Suppose, for instance, that ω_2 is the actual world, so X is true and $\mathrm{tv}(X) = 1$. Then your inaccuracy should decrease as you increase your credence in X (unless your credence surpasses 1!). Yet if ω_1 is the actual world, X is false, and $\mathrm{tv}(X) = 1$, then your inaccuracy will *increase* if your credence in X moves to the right.

Hopefully an obvious candidate for the i-function has suggested itself by now:

$$i(\mathrm{cr}(X), \mathrm{tv}(X)) = |\mathrm{tv}(X) - \mathrm{cr}(X)| \tag{10.1}$$

The absolute value is there to keep this quantity positive, reflecting the idea that we're measuring the *distance* between two points. This measure clearly satisfies truth-directedness. But it's not the only truth-directed, positive inaccuracy measure we could invent. Consider, for instance,

$$i(\mathrm{cr}(X), \mathrm{tv}(X)) = (\mathrm{tv}(X) - \mathrm{cr}(X))^2 \tag{10.2}$$

It's not obvious why one might prefer this squared local inaccuracy measure to the absolute value measure proposed above, especially since the two are ordinally equivalent (given any single truth-value and two credence assignments, both measures will agree on which credence is closer to the truth). We will return to that question in the next section.

Right now, I want to explain how to build a local inaccuracy measure on single credences into a global inaccuracy measure on credence distributions. The answer is simple. Given a credence distribution cr over a finite set of propositions $X_1, X_2, \ldots, X_n$, the global inaccuracy I of that distribution is calculated by summing its inaccuracies for each of the individual propositions. That is,

$$I(\text{cr}, \omega) = i(\text{cr}(X_1), \text{tv}_\omega(X_1)) + i(\text{cr}(X_2), \text{tv}_\omega(X_2)) + \ldots + i(\text{cr}(X_n), \text{tv}_\omega(X_n)) \tag{10.3}$$

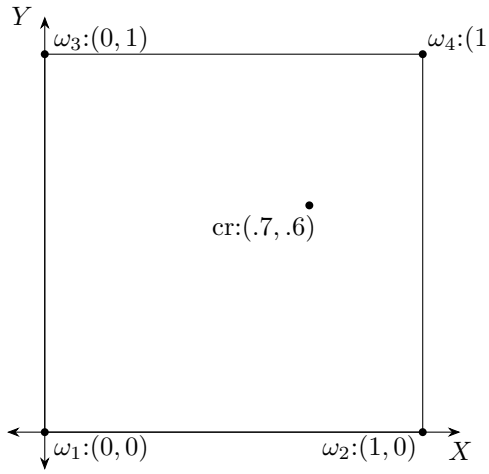
What are the ωs doing in this expression? We will often want to evaluate the inaccuracy of your credence distribution relative to conditions in the actual world. But sometimes we'll wonder how inaccurate your credences would've been if you'd maintained your distribution but lived in a different possible world. In other words, we'll want to evaluate the inaccuracy of a credence distribution cr in an *arbitrary* possible world ω. So we let $\text{tv}_\omega(X_j)$ represent the truth-value of proposition X_j in world ω; $I(\text{cr}, \omega)$ then measures the inaccuracy of cr relative to conditions in that world.[9]

A particular formula $I(\text{cr}, \omega)$ for calculating the global inaccuracy of distribution cr in world ω is called a **scoring rule**. Substituting different measures i of local inaccuracy into the formula for I gives us different scoring rules.[10] To illustrate how different choices of i lead to substantively different scoring rules, we will contrast the scoring rules that result from the absolute-value and squared i-functions of Equations (10.1) and (10.2). (There are infinitely many other candidates we might consider as well, but comparing these two will bring out the crucial contrasts.) I already mentioned that these two local inaccuracy functions are ordinally equivalent. But that doesn't make their respective scoring rules ordinally equivalent—the absolute-value scoring rule and the squaring rule may disagree on which overall credence distribution is more accurate in a given world.

Consider a credence distribution over two propositions X and Y. We visualize such a distribution using a two-dimensional diagram like Figure 10.1. There are now two axes, horizontal for proposition X and vertical for Y. With two (logically unrelated) propositions there are four possible worlds, whose locations I have marked on the diagram. Each world corresponds to an ordered pair of possible values for $\text{tv}(X)$ and $\text{tv}(Y)$. ω_3, for instance, is the world in which X is false and Y is true, so it's located at $(0, 1)$. A credence distribution over the two propositions can also be represented as an ordered pair; I have marked the distribution that assigns $\text{cr}(X) = 0.7$ and $\text{cr}(Y) = 0.6$.

Let's assess the inaccuracy of this cr distribution relative to possible

Figure 10.1: Gradational accuracy for two propositions



world ω_4. That is, we imagine ω_4 is the actual world and ask how inaccurate cr then turns out to be. On the diagram, we are asking how far the point $(0.7, 0.6)$ is from the point $(1, 1)$.

There are multiple ways to measure the distance between two points in space. Two of the most natural correspond to the two scoring rules we're considering. The scoring rule based on the absolute-value inaccuracy measure i is depicted in Figure 10.2. According to this scoring rule, the distance between the credence distribution $(0.7, 0.6)$ and the world $(1, 1)$ is

$$I(\text{cr}, \omega_4) = |1 - 0.7| + |1 - 0.6| = 0.3 + 0.4 = 0.7 \tag{10.4}$$

This scoring rule calculates the distance from $(0.7, 0.6)$ to $(1, 1)$ by totalling up how far you'd have to move horizontally and vertically to get from one point to the other. This is sometimes called the "taxicab" distance; it's how far you'd have to travel to get from one point to the other if you could travel only along a grid of city streets laid out parallel to the axes. I've illustrated such a trip with the short arrows in Figure 10.2.

On the other hand, using the squared local inaccuracy measure i yields a scoring rule known as the Euclidean distance, or Brier score.[11] As depicted in Figure 10.3, this is the distance "as the crow flies" between two points, illustrated by the dark arrow from $(0.7, 0.6)$ to $(1, 1)$. This is what most people naturally think of when the "distance" between two points is discussed.

Figure 10.2: The absolute-value scoring rule

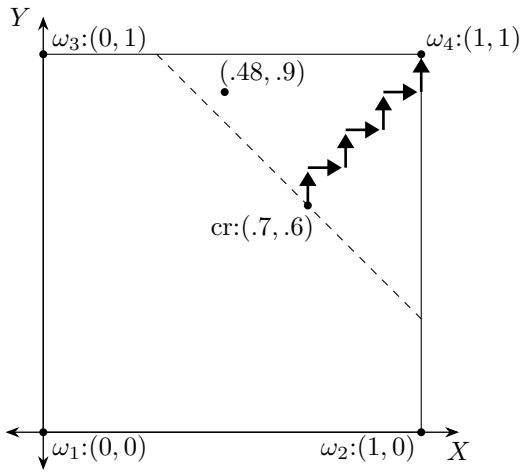
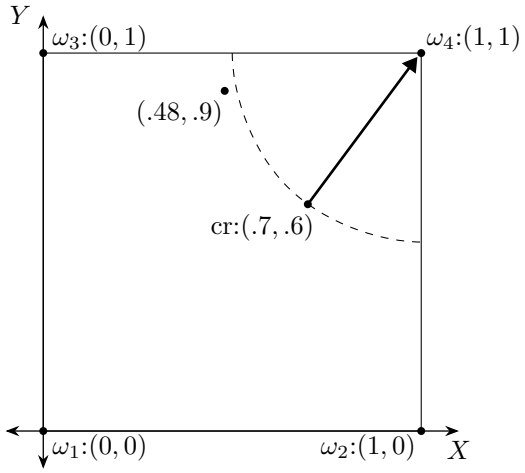


Figure 10.3: The Brier score



Strictly speaking the Brier score is not the distance between two points—as anyone who's studied the Pythagorean Theorem knows, you calculate the hypotenuse of a right triangle by squaring the lengths of the legs, summing, and then *taking the square-root*. We've left off the square-rooting step, because it makes no ordinal difference: if credence distribution cr is farther from the truth than distribution cr′ according to the Brier score, cr will remain farther than cr′ even after those scores are square-rooted. So there is no ordinal difference between the Brier score and the square-rooted Brier. On the other hand, there *is* an important ordinal difference between the Brier score and the absolute value score.

We can see this difference by observing the dashed elements in Figures 10.2 and 10.3. In each figure, the dashes indicate points in the square that are the same "distance" from ω_4 as $(0.7, 0.6)$.[12] The curves are different in the two diagrams because each diagram represents a different measure of "distance". Now consider the credence distribution that assigns 0.48 to X and 0.9 to Y. That distribution lies *inside* the dashed line in Figure 10.2, demonstrating that on the absolute-value score it is closer to ω_4 (and therefore *less* inaccurate relative to that world) than $(0.7, 0.6)$. On the other hand, the same point lies *outside* the dashed curve in Figure 10.3; according to the Brier score that distribution is *more* inaccurate relative to ω_4 than $(0.7, 0.6)$. Depending on which scoring rule we choose, we may change our minds about which of two agents has the more accurate credence distribution in a given world.

10.1.3 Strictly proper scoring rules

So which is the right answer? Is one of the scoring rules we've been consideringthe *correct* measure of inaccuracy? There's a traditional argument for preferring the Brier score to the absolute-value score as a measure of inaccuracy, but to understand it we must first consider *expected* inaccuracies.

Suppose I'm trying to determine the global inaccuracy of our earlier credence distribution that assigns $\mathrm{cr}(X) = 0.7, \mathrm{cr}(Y) = 0.6$. For simplicity's sake, let's suppose I calculate inaccuracies using the absolute-value scoring rule. We've already seen that if the actual world is ω_4, the global inaccuracy of cr will be 0.7 (from Equation (10.4)). Put another way, using the absolute-value score we have $I(\mathrm{cr}, \omega_4) = 0.7$. But ω_4 might not be the actual world. A bit of calculation will reveal these other global inaccuracy scores:

	$I(\text{cr}, \cdot)$	cr'
ω_1	1.3	0.1
ω_2	0.9	0.2
ω_3	1.1	0.3
ω_4	0.7	0.4

Ignore the last column for now; the global inaccuracy of cr in each possible world is reported in the middle column. We can see there that cr is most accurate if ω_4 is the actual world; on the other hand, ω_1 gives cr the highest inaccuracy.

Clearly the global inaccuracy of ω_4 depends on which possible world is actual. But what if I'm uncertain which world is actual? Then I might calculate the value I *expect* for cr's inaccuracy. After all, the inaccuracy of a credence distribution is a numerical quantity, and just like any numerical quantity I may calculate my expectation for its value. Let's suppose *my* credences are the distribution cr' shown in the final column of the table above. So I am least confident that world ω_1 is actual, and most confident that world ω_4 is actual. My expectation for the global inaccuracy of cr is

$$\begin{aligned}&\text{EI}(\text{cr}) = \\ &I(\text{cr}, \omega_1) \cdot \text{cr}'(\omega_1) + I(\text{cr}, \omega_2) \cdot \text{cr}'(\omega_2) + I(\text{cr}, \omega_3) \cdot \text{cr}'(\omega_3) + I(\text{cr}, \omega_4) \cdot \text{cr}'(\omega_4) \\ &= 1.3 \cdot 0.1 + 0.9 \cdot 0.2 + 1.1 \cdot 0.3 + 0.7 \cdot 0.4 = 0.92\end{aligned} \tag{10.5}$$

For each world, I calculate how inaccurate cr would be in that world and multiply by my credence cr' that that world is actual.[13] I then sum the results across all four worlds. Notice that because I'm more confident in, say, worlds ω_2 and ω_4 than I am in worlds ω_1 and ω_3, the former worlds have more influence on my expected inaccuracy calculation.

In general, if my credence distribution is cr' and the finite set of worlds under consideration is $\omega_1, \omega_2, \ldots, \omega_n$, I can calculate my expected inaccuracy for any distribution cr as follows:

$$\text{EI}(\text{cr}) = I(\text{cr}, \omega_1) \cdot \text{cr}'(\omega_1) + I(\text{cr}, \omega_2) \cdot \text{cr}'(\omega_2) + \ldots + I(\text{cr}, \omega_n) \cdot \text{cr}'(\omega_n) \tag{10.6}$$

This equation generalizes the expected inaccuracy calculation of Equation (10.5) above.[14]

Warning: We are discussing the epistemic goal of *minimizing* expected inaccuracy, where inaccuracy is a feature of credence distributions that can be measured numerically. Some authors prefer to

discuss credence distributions' **epistemic utility**, a numerical measure of epistemic value whose expectation rational agents *maximize*. Perhaps there are many aspects of a credence distribution that make it epistemically valuable or disvaluable. But many authors work under the assumption that accuracy is the sole determiner of a distribution's epistemic value, in which case that value can be calculated directly from the distribution's inaccuracy. (The simplest way is to let the epistemic utility of distribution cr in world ω equal $1 - I(\text{cr}, \omega)$.) When reading about accuracy arguments, be sure to notice whether the author asks agents to *minimize inaccuracy* or *maximize utility*. On either approach, the best credence is the one closest to the pin (in this case, the distribution tv). But with inaccuracy, as in golf, lowest score wins.

Equation 10.6 allows me to calculate my expected inaccuracy for any credence distribution, probabilistic or otherwise. If I wanted, I could even calculate an expectation for the global inaccuracy of my *own* credence distribution. (To do so, I simply replace cr with cr' throughout the equation.) But this is a fraught calculation to make. When I calculate my expected inaccuracy for my own current credences and compare them to the inaccuracy I expect for someone else's credences, I might find that I expect that other distribution to be less inaccurate than my own. We will say that one credence distribution **defeats another in expectation** if the latter assigns a lower expected inaccuracy to the former than it does to itself.

When an agent's credences are defeated in expectation by another distribution, she faces the same kind of "instability" we encountered with our weather forecaster—except this time the instability is coming from expected accuracy considerations rather than calibration measurements. If an agent's credence distribution leads her to expect that some other distribution is more accurate than her own, then (as far as minimizing expected inaccuracy goes) she will wish she had that other distribution instead of her own. The following norm rules out this kind of instability for rational credence distributions:

Permissibles Not Defeated: If an agent's credence distribution is rationally permissible, and she measures inaccuracy with an acceptable scoring rule, she will not expect any other distribution to be more accurate than her own.

Sometimes in epistemology we find a doxastic position that takes itself to

be *better* than other views; this doesn't mean it's a correct or even rational position. If you believe there's a material world, you will take your own beliefs to be more accurate than those of the skeptic, but that doesn't entail that you're doing better than the skeptic in any significant sense. Yet if a particular doxastic position takes itself to be *worse* than other views, this seems to be a serious flaw. James M. Joyce writes,

> If, relative to a person's own credences, some alternative system of beliefs has a lower expected epistemic [inaccuracy], then, by her own estimation, that system is preferable from the epistemic perspective. This puts her in an untenable doxastic situation. She has a *prima facie* epistemic reason, grounded in her beliefs, to think that she should not be relying on those very beliefs. This is a probabilistic version of Moore's paradox. Just as a rational person cannot fully believe "X but I don't believe X," so a person cannot rationally hold a set of credences that require her to estimate that some other set has higher epistemic utility. [This] person is... in this pathological position: her beliefs undermine themselves. (2009, p. 277)

Permissibles Not Defeated rules out such pathological distributions.

Yet Permissibles Not Defeated can also be used to rule out certain scoring rules. In some cases a credence distribution that is perfectly rationally permissible will look unstable if we assess it using the wrong scoring rule. Suppose I tell you I'm about to roll a fair die. You entertain six propositions, one for each possible outcome of the roll, and let's imagine that you assign each of those propositions a credence of 1/6. (In other words, $\mathrm{cr}(1) = \mathrm{cr}(2) = \mathrm{cr}(3) = \mathrm{cr}(4) = \mathrm{cr}(5) = \mathrm{cr}(6) = 1/6$.) I submit that this is at least a rationally *permissible* distribution in your situation.

But let's see what happens if, besides having this perfectly permissible credence distribution, you also use the absolute-value scoring rule to assess accuracy. You entertain six possible worlds—let's call them ω_1 through ω_6, with the subscripts indicating how the roll comes out in a given world. In world ω_1, the roll comes out 1, so $\mathrm{tv}(1) = 1$ while $\mathrm{cr}(1) = 1/6$, so by the absolute-value rule your local inaccuracy for the proposition that the roll comes out 1 is $|\mathrm{tv}(1) - \mathrm{cr}(1)| = |1 - 1/6| = 5/6$. On the other hand, your local inaccuracy in world ω_1 for the proposition that the roll comes out 2 is $|\mathrm{tv}(2) - \mathrm{cr}(2)| = |0 - 1/6| = 1/6$. In ω_1 you'll have the same inaccuracy for each of the other four propositions about how the roll might've come out, so your total, global inaccuracy in ω_1 will be $5/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 10/6$. A bit of reflection shows that you'll have the same global inaccuracy

score in each of the 6 possible worlds you entertain, so by Equation (10.6) your expected global inaccuracy (for your *own* credence distribution) will be 10/6.

Now consider your crazy friend Ned, who in light of the same evidence assigns credence 0 to each one of the six roll-outcome propositions. (So $\mathrm{cr}_N(1) = \mathrm{cr}_N(2) = \mathrm{cr}_N(3) = \mathrm{cr}_N(4) = \mathrm{cr}_N(5) = \mathrm{cr}_N(6) = 0$.) Let's calculate how inaccurate you expect Ned to be. In ω_1, $\mathrm{tv}(1) = 1$ while $\mathrm{cr}_N(1) = 0$, so Ned's inaccuracy for the proposition that the roll comes out 1 is $|1-0| = 1$. Yet in that world $\mathrm{tv}(2) = 0$ and $\mathrm{cr}_N(2) = 0$, so Ned's inaccuracy for the proposition that the roll comes out 2 is 0. Similarly for propositions 3, 4, 5, and 6. So Ned's total inaccuracy in world ω_1 is 1. A bit of reflection shows that Ned's total inaccuracy in each of the six worlds will be 1, so your expected global inaccuracy for him will be 1 as well. By your own lights, you expect crazy Ned to be more accurate than you are!

So if we calculate inaccuracy using the absolute-value rule, your credence distribution turns out to be defeated in expectation by Ned's. Yet Ned's distribution isn't *better* than yours in any epistemic sense—in fact, the Principal Principle would say that your distribution is rationally required while his is rationally forbidden! Something has gone wrong, and it isn't the credences you assigned. Instead, it's the scoring rule you used to assess your credences and Ned's. The absolute-value scoring rule is often thought to be an unacceptable scoring rule, precisely on the grounds that it can make distributions like crazy Ned's look better than your own. If we had calculated the expected inaccuracies using the Brier score instead, we would not have obtained this result. (See Exercise **??**.)

In this example the absolute-value rule and the Brier score are representative of two much broader classes of scoring rules, distinguished by the following definition:

Strictly Proper Scoring Rule: An agent with a probabilistic credence distribution who uses a strictly proper scoring rule will take herself to defeat in expectation every other distribution.

That is, if an agent assigns credences over a set of propositions that satisfy the probability axioms, and she assesses accuracy using a strictly proper scoring rule, she will always calculate her own distribution to have a *lower* expected inaccuracy than that of any other distribution over the same set of propositions. The absolute-value scoring rule is not strictly proper. In the example above, your credence distribution over possible roll outcomes satisfies the probability axioms. Yet if you use the absolute-value rule, there

is another distribution (Ned's) that you expect to be more accurate than your own.

The Brier score, on the other hand, is a strictly proper scoring rule. An agent with a probabilistic distribution who uses the Brier score will never take another distribution to defeat her own in expectation. In fact, the news is even better than that: A probabilistic agent who uses the Brier score will always expect herself to do *better* with respect to accuracy than any other distribution she considers. The Brier score is not the only scoring rule with this feature. Just for the sake of illustration, here's another local inaccuracy function that generates a strictly proper scoring rule:

$$i(\mathrm{cr}(X), \mathrm{tv}(X)) = -\log(1 - |\mathrm{tv}(X) - \mathrm{cr}(X)|) \tag{10.7}$$

The distinction between strictly proper scoring rules and other scoring rules is sometimes explained in terms of **credence elicitation**. Suppose we're going to pay a weather forecaster's pay in proportion to the accuracy of her reports. If we assess her accuracy using an improper scoring rule (like the absolute-value score), and if her credences satisfy the probability axioms, then there will be cases in which *by her own lights* she expects to be more accurate if she reports something other than her own credences. So we'll have incentivized her to make on-air reports that don't reflect what she truly thinks will happen. On the other hand, if we want to encourage probabilistic agents to report the credences they actually assign, a good strategy is to reward or punish them based on a strictly proper scoring rule.[15]

We now know enough about strictly proper scoring rules to argue that they are acceptable measures of gradational accuracy while improper rules are not. Strictly proper scoring rules have traditionally been favored because of something like the following argument:

Argument that Only Strictly Proper Scoring Rules are Acceptable

(Premise 1) On an acceptable scoring rule, no rationally permissible credence distribution can be defeated in expectation.

(Premise 2) Credence distributions that satisfy the probability axioms are rationally permissible.

(Definition) If a scoring rule is improper, it allows credence distributions satisfying the probability axioms to be defeated in expectation.

(Conclusion) Only strictly proper scoring rules are acceptable.

We will return to the details of this argument later; just to locate it in the context of the present discussion: Premise 1 is a rewording of Permissibles Not Defeated. The step I've called "Definition" comes from the definition of a strictly proper scoring rule. The argument concludes that we should always use strictly proper scoring rules in assessing the accuracy of credence distributions.

10.2 Joyce's accuracy argument for probabilism

It's bad enough when an agent expects another agent's distribution to be more accurate than her own—she may be more accurate than her rival in some worlds, less accurate in others, but on balance she expects to lose out. But it's even worse when an agent discovers that another distribution **accuracy dominates** her own. One credence distribution accuracy dominates another when the first is more accurate than the second in *every* possible world. Being defeated in expectation by another distribution is kind of like having a twin sister who takes all the same classes as you but has a better GPA. Being accuracy dominated is like that twin sister's getting a better grade than you *in every single class*.[16]

We already said that relative to an acceptable scoring rule no rationally permissible credence distribution should ever be defeated in expectation (that was our Permissibles Not Defeated rule). That rule straightforwardly entails

Permissibles Not Dominated: If an agent's credence distribution is rationally permissible, and she measures inaccuracy with an acceptable scoring rule, no other distribution will be more accurate than her own in every possible world.

If a distribution accuracy dominates the agent's, it will also have a lower *expected* inaccuracy than her distribution (because it will have a lower accuracy in each possible world). So being accuracy dominated is one way (a particularly extreme way) of being defeated in expectation. Permissibles Not Defeated says that permissible credence distributions are never defeated in expectation; this entails that they are also never dominated. So Permissibles Not Dominated is simply a consequence of what we've already assumed.

Repurposing a theorem of de Finetti's (1974), Joyce (1998) demonstrated the

Gradational Accuracy Theorem: Given a credence distribution cr over

a finite set of propositions $X_1, X_2, \ldots, X_n$, if we measure inaccuracy $I(\text{cr}, \omega)$ by the Brier score then:

- If cr does *not* satisfy the probability axioms, there exists a probabilistic distribution cr$'$ such that $I(\text{cr}', \omega) < I(\text{cr}, \omega)$ in *every* logically possible world ω; and
- If cr *does* satisfy the probability axioms, no such cr$'$ exists.

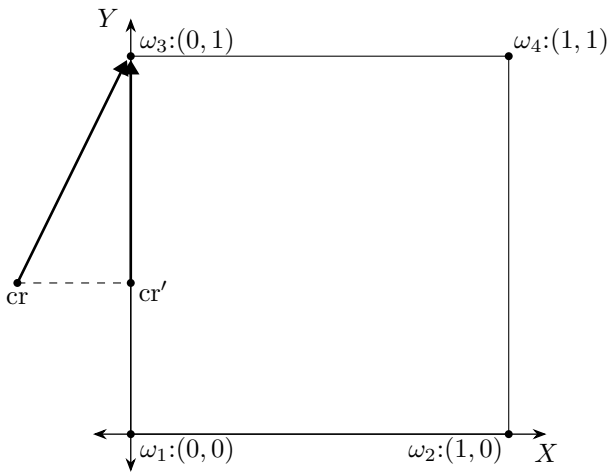
The Gradational Accuracy Theorem has two parts. The first part says that if an agent has a non-probabilistic credence distribution cr, we will be able to find a probabilistic distribution cr$'$ that accuracy dominates cr. No matter what the world is like, distribution cr$'$ is guaranteed to be less inaccurate than cr. So the agent with distribution cr can be certain that, come what may, she is leaving a certain amount of accuracy on the table by assigning cr rather than cr$'$. There's a cost in accuracy, independent of what you think the world is like and therefore discernible *a priori*, to assigning a non-probabilistic credence distribution—much as there's a guaranteed accuracy cost to assigning logically inconsistent beliefs. On the other hand (and this is the second part of the theorem), if an agent's credence distribution is probabilistic then no distribution is more accurate in every possible world. This seems a strong advantage of probabilistic credence distributions.[17]

Predd et al. (2009) showed that a similar Gradational Accuracy Theorem can be proven for any strictly proper scoring rule (not just the Brier score). Proving the second part of the theorem is difficult, but I will illustrate how the first part can be proven for the Brier Score. There are three probability axioms—Non-Negativity, Normality, and Finite Additivity—so we need to show how violating each one leaves a distribution susceptible to accuracy domination. We'll take them one at a time, in order.

Suppose credence distribution cr violates Non-Negativity by assigning some proposition a negative credence. In Figure 10.4 I've imagined that cr assigns credences to two propositions, X and Y, bearing no special logical relations to each other. cr violates Non-Negativity by assigning $\text{cr}(X) < 0$. (The value of $\text{cr}(Y)$ is irrelevant to the argument, but I've supposed it lies between 0 and 1.) We introduce probabilistic cr$'$ such that $\text{cr}'(Y) = \text{cr}(Y)$ but $\text{cr}'(X) = 0$; cr$'$ is the closest point on the Y-axis to distribution cr. Clearly cr$'$ is closer to ω_2 and ω_4 than cr is, so by the Brier score (reflecting distance as the crow flies) cr$'$ is less inaccurate than cr relative to both ω_2 and ω_4.

What if ω_3 is the actual world? I've indicated the distances from cr and cr$'$ to ω_3 with arrows. Because cr$'$ is the closest point on the Y-axis to cr,

Figure 10.4: Violating Non-Negativity



the points cr, cr′, and ω_3 form a right triangle. The arrow from cr to ω_3 is the hypotenuse of that triangle, while the arrow from cr′ to ω_3 is a leg. So the latter must be shorter, and cr′ is less inaccurate by the Brier score relative to ω_3. A parallel argument shows that cr′ is less inaccurate relative to ω_1. So cr′ is less inaccurate than cr relative to each possible world.

That takes care of Non-Negativity.[18] The accuracy argument against violating Normality is depicted in Figure 10.5. Suppose X is a tautology and cr assigns it some value other than 1. Since X is a tautology, there are no logically possible worlds in which it is false, so we need consider only the possible worlds marked as ω_2 and ω_4 in the diagram. We construct cr′ such that $\text{cr}'(Y) = \text{cr}(Y)$ and $\text{cr}'(X) = 1$. cr′ is closer than cr to ω_4 because the arrow from cr to ω_4 is the hypotenuse of a right triangle of which the arrow from cr′ to ω_4 is one leg. A similar argument shows that cr′ is closer than cr to ω_2, demonstrating that cr′ is less inaccurate than cr in every logically possible world.

Explaining how to accuracy-dominate a Finite Additivity violator requires a three-dimensional argument sufficiently complex that I will leave it for an endnote.[19] But we can show in two dimensions what happens if you violate one of the rules that follows from Finite Additivity, namely our Negation rule. Suppose your credence distribution assigns cr-values to two propositions X and Y such that Y is the negation of X. If you violate Negation, you'll have $\text{cr}(X) \neq 1 - \text{cr}(Y)$.

Figure 10.5: Violating Normality

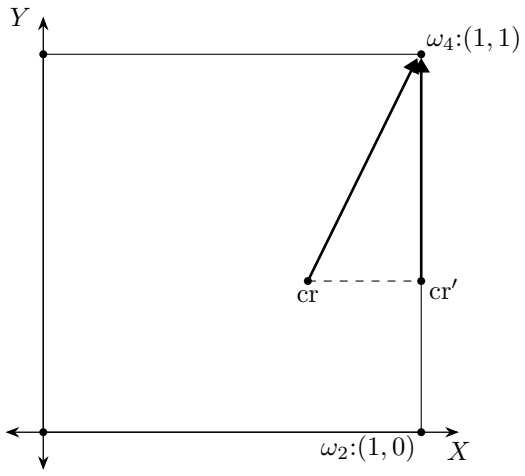
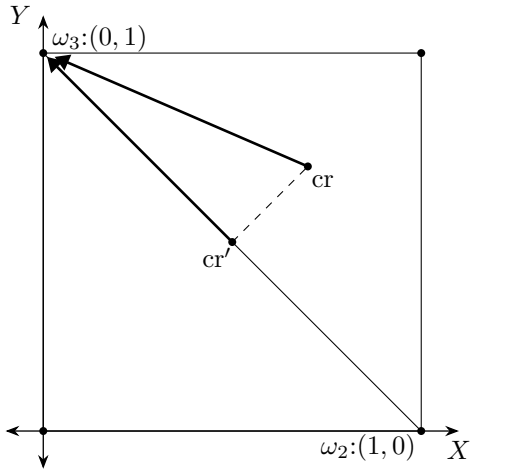


Figure 10.6: Violating Negation



I've depicted only ω_2 and ω_3 in Figure 10.6 because only those two worlds are logically possible (since X and Y must have opposite truth-values). The diagonal line connecting ω_2 and ω_3 has the equation $Y = 1 - X$; it contains all the credence distributions satisfying Negation. If cr violates Negation, it will fail to lie on this line. Then we can accuracy-dominate cr with the point closest to cr lying on that diagonal line (call that point cr$'$). Once more, we've created a right triangle with cr, cr$'$, and world ω_3. The arrow representing the distance from cr to ω_3 is the hypotenuse of this triangle, while the arrow from cr$'$ to ω_3 is its leg. So cr$'$ has the shorter distance, and if ω_3 is the actual world cr$'$ will be less inaccurate than cr according to the Brier score. A parallel argument applies to ω_2, so cr$'$ is less inaccurate than cr in each of the two logically possible worlds.[20]

Joyce (1998, 2009) leverages the advantage of probabilistic credence distributions displayed by the Gradational Accuracy Theorem into an argument for probabilism:

Gradational Accuracy Argument for Probabilism

(Premise) On an acceptable scoring rule, no rationally permissible credence distribution can be accuracy dominated.

(Result) Only strictly proper scoring rules are acceptable.

(Theorem) On a strictly proper scoring rule, any non-probabilistic credence distribution can be accuracy dominated.

(Conclusion) All rationally permissible credence distributions satisfy the probability axioms.

In this argument, "Premise" is Permissibles Not Dominated. "Result" comes from the previous section's argument that only strictly proper scoring rules are acceptable. "Theorem" is the Gradational Accuracy Theorem. And the conclusion of this argument is Probabilism.

10.2.1 An objection to this argument

Unlike representation theorem and Dutch Book arguments, the Gradational Accuracy Argument for Probabilism has nothing to do with an agent's decision-theoretic preferences over practical acts. It clearly pertains to the *theoretical* rationality of credences assigned in pursuit of an *epistemic* goal: accuracy. (This is why Joyce's original 1998 paper was titled "A Nonpragmatic Vindication of Probabilism".) So the argument is not susceptible to one of the complaints we made against those alternatives.

It may, however, be susceptible to our Linearity In, Linearity Out concern. As I've framed the Gradational Accuracy Argument, its second step rules out improper scoring rules. As I noted above, one can argue for this step via our Argument that Only Strictly Proper Scoring Rules are Acceptable. Yet Premise 2 of that latter argument states that "Credence distributions that satisfy the probability axioms are rationally permissible." At least as I've reconstructed the arguments, they run like this: Start by assuming that probabilistic credence distributions are rationally permissible. Use that assumption to restrict our attention to strictly proper scoring rules. Having restricted ourselves to strictly proper scoring rules, use them to argue that probabilism is rationally required.

Isn't this somewhat circular? Admittedly, the premise we started with (that probabilistic credences are *permitted*) is weaker than the conclusion with which we ended (that probabilistic credences are *required*). Still, we're assuming something about the rationality of probabilism in order to prove something about the rationality of probabilism. Sounds like Linearity In, Linearity Out to me.

In offering various accuracy-based arguments for probabilism, Joyce has been well aware of this concern. Originally in his (1998) he selected the Brier score over the absolute-value score not on grounds of propriety but instead on the grounds that Brier evinced a number of appealing formal properties. Joyce showed that any scoring rule displaying those properties would leave non-probabilistic credence distributions accuracy-dominated, allowing us to run a version of the Gradational Accuracy Argument for Probabilism. Maher (2002), however, argued that these properties were implausible as requirements on rationally-acceptable scoring rules, and defended the absolute-value score. So Joyce (2009) shifted tactics, and ran his argument based on our Premise 2.[21]

Premise 2 is stated somewhat ambiguously in the Argument that Only Strictly Proper Scoring Rules are Acceptable—what exactly does it have to say for that argument to work? We *don't* need to maintain that in any situation, any probabilistic credence distribution would be rationally permissible to assign. Only an extreme Subjective Bayesian (Section 5.1.2) who maintained the Regularity Principle would agree to such a premise; it would be rejected by anyone who believed in further synchronic constraints beyond the probability axioms (such as the Principal Principle) or believed that an agent could receive evidence that rationally required her to be certain of particular propositions. What we *do* need is the claim that for any language $\mathcal{L}$ and any probabilistic credence distribution over $\mathcal{L}$, there exists some situation in which it would be rationally permissible for an agent to assign that

distribution over that language.

Given this premise, the argument against improper scoring rules runs as follows: Take an arbitrary improper scoring rule, and suppose for *reductio* that it's acceptable for determining rational permissibility. By the definition of impropriety, some probabilistic distribution will be defeated in expectation on that rule. This type of defeat is recognizable *a priori*, and is independent of the particulars of an agent's situation. So now go to the situation (whose existence is guaranteed by our premise) in which that distribution is rationally permissible. By supposition the distribution is defeated in expectation on the (acceptable) improper scoring rule, so by Expected Inaccuracy Permissiblity it is not rationally permissible. But we said the distribution in question was rationally permissible in this situation, so we have a contradiction.

Now that we've understood what Premise 2 says, how might we defend it in a fashion that doesn't beg the question in an argument for probabilism? Joyce argues that for any probabilistic credence distribution, we could imagine a situation in which an agent is rationally certain that those credence values reflect the objective chances of the propositions in question. By the Principal Principle, the relevant credence distribution would then be rationally required. Hájek (2009a) responds that many languages over which probabilistic distributions are assigned contain propositions that couldn't possibly *have* objective chances. (For instance, propositions about the physical laws that give rise to objective chances.) For such a proposition, it wouldn't be rational for an agent to be certain that her credence equalled the objective chance, so Joyce's argument wouldn't establish the rational permissibility of the probabilistic distribution in question.

Yet there's a much more general way of attacking the assumption that any probabilistic credence distribution could be rationally permissible under the right conditions. Recall our characters Mr. Prob, Mr. Bold, and Mr. Weak. Mr. Prob satisfies the probability axioms, while Mr. Bold violates Finite Additivity by having his credence in each proposition be the square-root of Mr. Prob's credence in that proposition. Mr. Bold happily assigns a higher credence to every uncertain proposition than Mr. Prob does. In arguing for probabilism, we look to establish that Mr. Bold's (and Mr. Weak's) credences are rationally forbidden. If we could establish that rational credences match the numerical values of known frequencies or objective chances, then Mr. Bold's distribution could be ruled out immediately, because frequencies and chances must be additive.[22] But part of Mr. Bold's boldness is that even when he and Mr. Prob are both certain that a particular proposition has a particular nonextreme chance, he's willing to assign

that proposition a higher credence than its chance value. Mr. Bold is willing to be even more confident of a given experimental outcome than its numerical chance!

The accuracy argument for probabilism requires Mr. Bold to be a bit more aggressive than we've noted in the past. Before Mr. Bold might have allowed that both his approach and Mr. Prob's are rationally permissible. But now we've seen that if Mr. Bold grants that probabilistic distributions are generally *permissible*, this premise underwrites an argument that probabilistic distributions are always *required*. So Mr. Bold must now maintain that his kind of response to evidence about chances—setting credences greater than the chance values themselves—is the only rationally permissible response. While we might intuitively think this approach is crazy, the accuracy-based *argument* for probabilism is question-begging against it.

10.2.2 Do we really need Finite Additivity?

[HERE I WILL DISCUSS THE LINDLEY PAPER]

10.3 An accuracy argument for Conditionalization

Arguing for probabilism on non-circular accuracy-based grounds turns out to be difficult. But if you've already accepted probabilism, a remarkable accuracy-based argument for updating by Conditionalization becomes available. The relevant result was proven by Hilary Greaves and David Wallace (2006).[23] We start by restricting our attention to strictly proper scoring rules. Doing so is non-circular in this context, because we imagine that we've already accepted probabilism as rationally required. This lets us appeal to the credence-elicitation features of proper rules, as in the Argument that Only Strictly Proper Scoring Rules are Acceptable.

Greaves and Wallace think of Conditionalization as a *plan* one could adopt for how to change one's credences in response to one's future evidence. Imagine we have an agent at time t_i with probabilistic credence distribution cr_i, who is certain she will gain some evidence before t_j. Imagine also that there's a finite partition of propositions $E_1, E_2, \ldots, E_n$ in $\mathcal{L}$ such that the agent is certain the evidence gained will be a member of that partition. The agent can then form a plan for how she intends to update—she says to herself, "If I get evidence E_1, I'll update my credences to such-and-such"; "If I get evidence E_2, I'll update my credences to so-and-so"; etc. In other words, an updating plan is a function from members of the evidence partition to cr_j

distributions she would assign in response to receiving that evidence. Conditionalization is the plan that directs an agent receiving partition member E_m as evidence between t_i and t_j to set $\mathrm{cr}_j(\cdot) = \mathrm{cr}_i(\cdot \mid E_m)$.

Greaves and Wallace next show how the agent can calculate the expected inaccuracy of each available plan[24] from her point of view at t_1. The calculation proceeds in six steps:

1. Pick a possible world ω to which the agent assigns non-zero credence at t_i.

2. Figure out which member of the partition $E_1, E_2, \ldots, E_n$ the agent will receive as evidence between t_i and t_j if ω turns out to be the actual world. (This will always be possible because possible worlds are maximally specified.) We'll call that piece of evidence E.

3. Take the updating plan being evaluated and figure out what credence distribution it recommends to the agent if she receives evidence E between t_i and t_j. This is the credence distribution the agent will assign at t_j if ω is the actual world and she follows the plan in question. We'll call that distribution cr_j.

4. Whichever scoring rule we've chosen (among the strictly proper scoring rules), use it to determine the inaccuracy of cr_j if ω is the actual world. (In other words, calculate $I(\mathrm{cr}_j, \omega)$.)

5. Multiply that inaccuracy value by the agent's t_i credence that ω is the actual world. (In other words, calculate $\mathrm{cr}_i(\omega) \cdot I(\mathrm{cr}_j, \omega)$.)

6. Repeat this process for each world to which the agent assigns positive credence at t_i, then sum the results.

This calculation has the t_i agent evaluate an updating plan by determining what cr_j distribution that plan would recommend in a particular possible world. She assesses the recommended distibution's accuracy in that world, weighting the result by her confidence that the world in question will obtain. Repeating this process for each possible world and summing the results, she develops an overall expectation of how accurate her t_j credences will be if she implements the plan.

Greaves and Wallace go on to prove the following theorem:

Accuracy Updating Theorem: Given any strictly proper scoring rule, probabilistic distribution cr_i, and evidential partition in $\mathcal{L}$, a t_i agent who calculates expected inaccuracies as described above will

> find Conditionalization more accurate than any updating plan that recommends different credences at t_j.

The Accuracy Updating Theorem demonstrates that from her vantage point at t_i, an agent with probabilistic credences and a strictly proper scoring rule will expect to be most accurate at t_j if she updates by Conditionalization. Given a principle something like Permissibles Not Defeated for updating plans, we can use this result to argue that no updating plan deviating from Conditionalization is rationally acceptable.

But does this argument show that the agent is rationally required to *update* by Conditionalization between t_i and t_j? If she's interested in minimizing expected inaccuracy, then at t_i she should certainly *plan* to update by conditionalizing—of all the updating plans available to the agent at t_i, she expects Conditionalization to be most accurate. Yet being required to make a particular plan is different from being required to implement it. At t_j the agent may remember what she planned at t_i, but why should the t_j agent do what her t_i self thought best? Among other things, the t_j agent has more evidence than her t_i self did.

This is the same problem we identified in Chapter 9 for diachronic Dutch Strategy arguments. The Accuracy Updating Theorem establishes a *synchronic* point about which policy an accuracy-concerned t_i agent will hope her t_j self applies. But absent a substantive premise that agents are rationally required later to honor their earlier plans, we cannot move from this *synchronic* point to a genuinely *diachronic* norm like Conditionalization.

Notes

[1]In discussions about internalism versus externalism in epistemology (whether it be access internalism, mentalism, or some other version), rationality is often taken to be an internalist normative category. So how can we argue to rational consistency norms from an externalist standard like accuracy? The key is the *a priori* nature of the considerations presented to the agent; without knowing anything empirical about the world, an agent with logically inconsistent beliefs is in a position to know that at least some of the propositions she believes are false. She need not ascertain any facts about reliability or anything else not internally available in order to come to this conclusion.

[2]See the Further Readings for criticisms of accuracy arguments on the grounds that they are teleological.

[3]There's also been some interesting empirical research on how well-calibrated agents' credences are in the real world. A robust finding is that people tend to be overconfident in their opinions—only, say, 70% of the propositions to which they assign credence 0.9 turn out to be true. For a survey of the literature see (Lichtenstein, Fischoff, and Phillips 1982).

[4]See (Murphy 1973)—and notice that the author is himself a meteorologist! Like so many notions in probabilism, the idea of calibration as accuracy was hinted at in Ramsey. In the latter half of his (1931), Ramsey asks what it would be for credences "to be consistent not merely with one another but also with the facts." (p. 93) He later writes, "Granting that [an agent] is going to think always in the same way about all yellow toadstools, we can ask what degree of confidence it would be best for him to have that they are unwholesome. And the answer is that it will in general be best for his degree of belief that a yellow toadstool is unwholesome to be equal to the proportion of yellow toadstools which are in fact unwholesome." (p. 97)

[5]This example is taken from (Joyce 1998).

[6]It's sweeps week.

[7]Here's another, related problem for calibration as a measure of accuracy: an agent who assigns credences over a partition of n propositions can guarantee herself a perfect calibration score (in every possible world!) by always assigning each proposition a credence of $1/n$. Depending on how you feel about the Principle of Indifference (Section 5.3), this might be a reasonable assignment when the agent has no evidence relevant to the members of the partition. But even if she receives ample evidence favoring one partition member over others (perhaps she's rolling a die that she *knows* to be biased), calibration will continue to consider her perfect if she continues to assign $1/n$ to each member.

[8]Compare the practice in statistics of treating a proposition as a dichotomous random variable with value 1 if true and 0 if false.

[9]Notice that we're keeping the numerical values of the distribution cr constant as we measure inaccuracies relative to different possible worlds. $I(\text{cr}, \omega)$ doesn't somehow measure the inaccuracy in world ω of the credence distribution the agent *would* have had in that world. Instead, given a particular credence distribution cr of interest to us, we will use $I(\text{cr}, \omega)$ to measure how inaccurate *that very numerical distribution* is relative to each of a number of distinct possible worlds.

[10]While different i-functions yield different scoring rules, notice that our definition of I in terms of i causes all scoring rules to have certain features in common. First, while $I(\text{cr}, \omega)$ is in some sense a global measure of the inaccuracy of cr in world ω, it doesn't take into account any wholistic or interactive features among the individual credences cr assigns. The inaccuracy of cr with respect to proposition X_j is calculated strictly in terms of the truth-value of X_j; the results are then summed across the X_j. So no possible interactions or comparisons between the cr-values assigned to distinct X_j are taken into account. Second, the i-value of each X_j contributes equally to the sum $I(\text{cr}, \omega)$. One might think that in certain circumstances some X_j are much more important to be accurate about than others. That would suggest weighting the disparate i-values before summing, in contrast to the strict equaniminty imposed by I.

[11]Named after George Brier—another meteorologist!—who discussed it in his (1950).

[12]The dashed elements are like contour lines on a topographical map. There, every dashed point on a given contour line lies at the same altitude. Here, every dashed point has the same inaccuracy relative to world ω_4.

[13]Strictly speaking ω_1 is a world, not a proposition, so cr' doesn't assign it a value. Here I'm employing the convention that "$\text{cr}'(\omega_1)$" is the credence that distribution cr' assigns to the proposition that ω_1 is the actual world.

[14]Readers familiar with decision theory (perhaps from Chapter 7) may notice that the expected-inaccuracy calculation of Equation (10.6) strongly resembles Savage's formula for calculating expected utilities. Here a "state" is a possible world ω_i that might be actual, an "act" is assigning a particular credence distribution cr, and an "outcome" is

the inaccuracy that results if ω_i is actual and one assigns cr. Savage's expected utility formula was abandoned by Jeffrey because it yielded implausible results when states and acts were not independent. Might we have a similar concern about Equation (10.6)? What if the act of assigning a particular credence distribution is not independent of the state that a particular one of the ω_i obtains? Should we move to a Jeffrey-style expected inaccuracy calculation, and perhaps from there to some analogue of Causal Decision Theory? As of this writing, this question is only just beginning to be explored in the accuracy literature, in articles such as (Greaves 2013) and (Konek and Levinstein ms).

[15](Joyce 2009, p. 266) reports that "The term 'scoring rule' comes from economics, where values of $[I]$ are seen as imposing penalties for making inaccurate probabilistic predictions."

By the way, one sometimes sees "strictly proper" scoring rules distinguished from "proper" scoring rules. Relative to a proper scoring rule, any probabilistic distribution expects itself to do *at least as well as* other distributions with respect to accuracy. On a strictly proper score, a probabilistic distribution will expect itself to do *better than* every other distribution. If we evaluate the probabilistic weather forecaster using a proper scoring rule rather than a strictly proper one, she will have no *incentive* to report a distribution other than her own, but she sometimes won't see any *harm* in doing so. With a strictly proper rule, she will always see an *advantage* in reporting her credences accurately. Though distinguishing propriety from strict propriety would require slightly rewording some of the arguments that follow, I'm mainly interested in the general structure of those arguments. So I won't bother with the relevant subtleties here.

[16]Like the distinction between propriety and strict propriety, a distinction is sometimes drawn between "strong" and "weak" accuracy domination. cr' strongly dominates cr just in case cr' is less inaccurate than cr in every possible world. cr' weakly dominates cr if cr' does at least as well as cr with respect to accuracy in every world and better than cr in at least one world. All of my references to "accuracy dominance" in what follows will be references to strong accuracy dominance. For discussion of proving the results of this section using the concept of weak accuracy dominance, see (Schervish, Seidenfeld, and Kadane 2009).

[17]The second part of the Gradational Accuracy Theorem stands to the first part much as the Converse Dutch Book Theorem stands to the Dutch Book Theorem (Chapter 9).

[18]Notice that a similar argument could be made for any cr lying outside the square defined by ω_1, ω_2, ω_3, and ω_4. So this argument also shows how to accuracy dominate a distribution that violates our Maximum rule.

Now one might wonder why we *need* an argument that credence-values below 0 or above 1 are irrational—didn't we stipulate our scale for measuring degrees of belief such that no value could ever fall outside that range? On some ways of understanding credence, arguments for Non-Negativity are indeed superfluous. But one might define credences purely in terms of their role in generating preferences (as discussed in Chapter 8) or in sanctioning bets (as discussed in Chapter 9), in which case there would be no immediate reason why a credence couldn't take on a value below zero.

[19]Suppose you assign credences to three propositions X, Y, and Z such that X and Y are mutually exclusive and $Z \dashv\models X \vee Y$. We establish X-, Y-, and Z-axes, then notice that only three points in this space represent logically possible worlds: $(0, 0, 0)$, $(1, 0, 1)$, and $(0, 1, 1)$. The distributions in this space satisfying Finite Additivity all lie on the plane passing through those three points. If your credence distribution cr violates Finite Additivity, it will not lie on that plane. We can accuracy-dominate it with distribution cr' that is the closest point to cr lying on the plane. If you pick any one of the three logically

possible worlds (call it ω), it will form a right triangle with cr and cr$'$, with the segment from cr to ω as the hypotenuse and the segment from cr$'$ to ω as a leg. That makes cr$'$ closer than cr to ω.

[20]To give the reader a sense of how the second part of the Gradational Accuracy Theorem is proven, I will now argue that no point lying inside the box in Figure 10.6 and on the illustrated diagonal may be accuracy dominated with respect to worlds ω_2 and ω_3. In other words, I'll show how satisfying Negation wards off accuracy domination (assuming one measures inaccuracy by the Brier score).

Start with distribution cr$'$ in Figure 10.6, which lies on the diagonal and therefore satisfies Negation. Imagine drawing two circles through cr$'$, one centered on ω_2 and the other centered on ω_3. To improve upon the accuracy of cr$'$ in ω_2, one would have to choose a distribution closer to ω_2 than cr$'$—in other words, a distribution lying inside the circle centered on ω_2. To improve upon the accuracy of cr$'$ in ω_3, one would have to choose a distribution lying inside the circle centered on ω_3. But since cr$'$ lies on the line connecting ω_2 and ω_3, those circles are tangent to each other at cr$'$, so there is no point lying inside *both* circles. Thus no distribution is more accurate than cr$'$ in both ω_2 and ω_3.

[21]Joyce's argument in his (2009) is structurally different from the arguments I've presented. He uses some mathematical results to establish probabilism from Coherent Admissibility, the principle that no acceptable scoring rule allows any probabilistic distribution to be accuracy-dominated. Joyce then argues for Coherent Admissibility on the grounds of our Premise 2 and Permissibles Not Dominated. So his argument proceeds without assuming Permissibles Not Defeated. Still, the crucial assumption I'm attacking (Premise 2) underlies both Joyce's approach and the one I've presented.

[22]See note 5 in Chapter 5.

[23]For an alternative accuracy-based approach to updating, see (Leitgeb and Pettigrew 2010a,b).

[24][ENDNOTE ABOUT G AND W'S NOTION OF AN "AVAILABLE PLAN"]