J. Dmitri Gallow

Lecture Notes for *An Advanced Introduction to Bayesian Epistemology*

Contents

1

Interpretations of Probability

Consider the following claims about probability:

1. Whether the coin lands heads or tails is a matter of chance
2. It's unlikely to rain
3. A smoker is more likely to get cancer than a non-smoker
4. The chance of rain is 90%

All of these claims are about probability or chance; but they are different *kinds* of claims about probability. (1) says that whether the coin lands heads is *contingent*—it could be true and it could be false. It's not guaranteed to land heads; nor is it guaranteed to not land heads. But it doesn't say anything about the *strength* of this contingency. In contrast, (2) goes further. It doesn't just say that it is contingent whether it rains—it additionally tells us that the contingency is weak; that rain is *unlikely*. Claims like this use a unary predicate like 'likely' or 'unlikely' to describe outcomes or propositions. If we were regimenting these kinds of claims in a formal language, we might think of using a propositional operator like 'Δ', where 'ΔA' says that A is likely. In contrast, claims like (3) make *comparative* claims about one outcome being more, less, or just as likely as another; they invoke a binary relation between outcomes or propositions. These kinds of claims could be regimented in a formal language with a two-place propositional operator like $\succeq$, where '$A \succeq B$' says that A is at least as likely as B. Finally, (4) uses a particular *number* to measure probability. These kinds of claims could be regimented with a *function*, $\mathbb{P}$, from propositions to numbers, where '$\mathbb{P}(A) = x$' says that the probability of A is x. All of these approaches to theorizing about probability have been pursued (and there are interesting relationships between them), but in this course, we're going to focus on the final approach, which is far and away the most prevalent and common. Notice that we may be able to analyze the earlier kinds of claims in terms of the final one. At first glance, (1) will be true iff there's a non-zero probability of the coin landing heads and a non-zero probability that the coin *doesn't* land heads. (2) will be true iff the probability of rain is less than $1/2$. And (3) will be true iff the probability that a smoker gets cancer is greater than the probability that a non-smoker gets cancer.

I want to begin by asking two questions about claims like these:

1. What do these kinds of claims *mean*? What are they *saying*?
2. What determines which probabilities are the right ones?

There are roughly three families of answers to these questions. Those in the first group all say that claims about probabilities are claims about some *objective* quantity and that these quantities may be determined *a priori*. Those in the second group agree that probabilities are objective, but maintain that they can only be determined *a posteriori*. And those in the third group say that probabilities are *subjective*, and can vary from person to person (though they may think that there are more and less rational subjective probabilities for a person to have).

In these notes, we're going to be exploring this final interpretation of probabilities, according to which they are purely subjective, and represent something like the *strength of your belief*. On this view, to say that it's unlikely to rain is just to say that you are more confident that it won't rain than you are that it will rain. To understand why some people were led to this interpretation of probability, it's worthwhile rehearsing some of the other standard views, and the standard objections to those views.

1.1 Objective and A Priori Interpretations

Classical Interpretation According to the Classical Interpretation, to say that the probability that the coin lands heads is one half is just to say that the proportion of 'equally possible' cases in which the coin lands heads is one half. In general, the classical interpretation assigned probabilities *uniformly* over all possible cases. This is known as the "principle of indifference".

Laplace: "The theory of chances consists in reducing all events of the same kind to a certain number of equally possible cases, that is to say, to cases whose existence we are equally uncertain of, and in determining the number of cases favourable to the event whose probability is sought. The ratio of this number to that of all possible cases is the measure of this probability".

The Principle of Indifference If A and B are equally possible, then the probability of A equals the probability of B.

It's worthwhile asking what the classical view meant by 'equally possible'. The most natural interpretation is 'equally *probable*', but that can't be what the principle of indifference is saying, since then the principle would be a tautology.

Objections:

- Bertrand's Paradox (cf. van Fraassen's cube factory).

  - Cubes come from the factory with side lengths between 0 and 2 cm. The side length being between 0 and 1 cm is equally possible as the side length being between 1 and 2 cm. So the probability that the side length is between 0 and 1 cm is one half.

  - Cubes come from the factory with face areas between 0 and 4 cm^2. The face area being in $(0, 1]$ is equally possible as $(1, 2]$, $(2, 3]$, and $(3, 4]$. So the probability that the face area is between 0 and 1 cm^2 is one fourth.

  - But the side length is between 0 and 1 cm iff the face area is between 0 and 1 cm^2. So they must have the same probability.

Insofar as the principle of indifference gives them different probabilities, it has contradicted itself.

- ▹ How is it possible to learn (e.g., the bias of a coin)? No matter how many times the coin is flipped, there remain just two (equally?) possible outcomes.

The classical view was supplemented with an additional rule: the *Rule of Succession*. It says:

$$\text{probability of heads on } N+1\text{st flip} = \frac{\text{number of heads in first } N \text{ flips} + 1}{N+2}$$

Logical Interpretation According to Carnap's *logical* interpretation of probability, probability earn their keep in the role they play in relations of inductive support. These are inductive or probabilistic generalizations of the relation of deductive entailment; and they have entailment as their limiting case.

Given some first-order language with some number of predicates and constants, we construct the set of *state descriptions* by considering all of the possible truth-value assignments to the atomic sentences of the language. For instance, if there are two predicates, F and G, and two constants, a and b, then we have the 16 state descriptions given in the margin.

A *structure description* abstracts away from the identities of the constants, and simply reports how many things have each collection of properties. So a structure description gives a purely qualitative summary of a state description.

The principle of indifference says to give each *state description* equal probability. Carnap, in contrast, says to give each *structure description* equal probability (and to distribute probabilities equally within each structure description).

Example: there are three ravens, a, b, and c, which could either be black, B, or not. The principle of indifference gives us the first *probabilistic truth-table* below; whereas Carnap gives the second:

Ba	Bb	Bc	POI	Carnap
1	1	1	1/8	3/12
1	1	0	1/8	1/12
1	0	1	1/8	1/12
1	0	0	1/8	1/12
0	1	1	1/8	1/12
0	1	0	1/8	1/12
0	0	1	1/8	1/12
0	0	0	1/8	3/12

Carnap then says that some evidence, E, supports a hypothesis, H, iff (and to the extent that)

$$\mathbb{P}(H \mid E) \stackrel{\text{def}}{=} \frac{\mathbb{P}(H\&E)}{\mathbb{P}(E)} > \mathbb{P}(H)$$

State descriptions for F, G and a, b:

1. $Fa\&Fb\&Ga\&Gb$
2. $Fa\&Fb\&Ga\&\neg Gb$
3. $Fa\&Fb\&\neg Ga\&Gb$
4. $Fa\&Fb\&\neg Ga\&\neg Gb$
5. $Fa\&\neg Fb\&Ga\&Gb$
6. $Fa\&\neg Fb\&Ga\&\neg Gb$
7. $Fa\&\neg Fb\&\neg Ga\&Gb$
8. $Fa\&\neg Fb\&\neg Ga\&\neg Gb$
9. $\neg Fa\&Fb\&Ga\&Gb$
10. $\neg Fa\&Fb\&Ga\&\neg Gb$
11. $\neg Fa\&Fb\&\neg Ga\&Gb$
12. $\neg Fa\&Fb\&\neg Ga\&\neg Gb$
13. $\neg Fa\&\neg Fb\&Ga\&Gb$
14. $\neg Fa\&\neg Fb\&Ga\&\neg Gb$
15. $\neg Fa\&\neg Fb\&\neg Ga\&Gb$
16. $\neg Fa\&\neg Fb\&\neg Ga\&\neg Gb$

Structure descriptions for F, G and a, b:

- Two FGs (1)
- Two $F\,\neg G$s (4)
- Two $\neg F\,G$s (13)
- Two $\neg F\,\neg G$s (16)
- One FG and one $F\,\neg G$ (2, 3)
- One FG and one $\neg F\,G$ (5, 9)
- One FG and one $\neg F\,\neg G$ (6, 11)
- One $F\,\neg G$ and one $\neg F\,G$ (7, 10)
- One $F\,\neg G$ and one $\neg F\,\neg G$ (8, 12)
- One $\neg F\,G$ and one $\neg F\,\neg G$ (14, 15)

Notice that Ba confirms $Ba\&Bb\&Bc$.

Objections:

- If there are infinitely many things, any universal generalization will always have probability zero; but it seems like we can confirm the laws of nature even in an infinite universe.
- Carnap's degrees of confirmation are *language-dependent*. Run the procedure with *green*, and you'll confirm "All emeralds are green". Run it with *grue*, and you'll confirm "All emeralds are grue". (cf. Bertrand's paradox)

1.2 *Objective and A Posteriori Interpretations*

Actual Frequentism According to the actual frequentist, to say that the probability that the coin lands heads is one half is just to say that the *actual frequency* of heads landings is one half.

$$\mathbb{P}(\text{a flipped coin lands heads}) = \frac{\#\text{actual flips that land heads}}{\#\text{actual flips}}$$

In general,

$$\mathbb{P}(\text{an } F \text{ is } G) = \frac{\#GF\text{s}}{\#F\text{s}}$$

Note that, on this view, probabilities don't just attach to *outcomes*, but rather to an *outcome* and a *reference class*. I flip a quarter. According to the actual frequentist, 'the probability that the flip lands heads' is ambiguous. It could be referring to the frequency of heads landings amongst *coin* flips, or amongst *quarter* flips, or amongst flips of *American* coins, etc.

Objections:

- Isn't it *possible* for a fair coin to land heads every time it's flipped?
- Couldn't a fair coin be only flipped once? (or never flipped at all?)
- It seems that probabilities can be irrational (for instance, the probability that a radon atom decays in some number of seconds can be e^{-1}); but no actual frequency is irrational.
- Doesn't probability *explain* actual frequencies? It looks like you can explain the fact that about half of the coin flips landed heads by pointing to the fact that the probability of heads was one half. But nothing can explain itself; so this suggests that probabilities aren't actual frequencies.

Hypothetical Frequentism According to hypothetical frequentism, to say that the probability of heads is one half is to say that, were you to flip the coin infinitely many times, the frequency of heads landings would approach one half in the limit.

$$\mathbb{P}(\text{a flipped coin lands heads}) = \frac{\#\text{the first } n \text{ flips that landed heads}}{n}$$

In general,

$$\mathbb{P}(\text{an } F \text{ is } G) = \lim_{n\to\infty} \frac{\#\text{the first } n \text{ } F\text{s that are } G}{n}$$

Objections:

- How do we *order* the Fs? Different orderings will give different limiting frequencies. For instance, we can imagine situations in which ordering the outcomes by their *temporal* locations leads to a different limiting frequency than ordering the outcomes by their *spatial* locations.
- Doesn't this trivialize the Law of Large Numbers?
- Doesn't probability *explain* limiting frequency?

The (Strong) Law of Large Numbers says that, if the coin is fair, then the probability that the frequency of heads landings approaches one half in the limit is 100%.

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{\#\text{first } n \text{ flips which land heads}}{n} = 1/2\right) = 1$$

Long-run Propensitism To say that the probability of heads is one half is to say that the coin-flipping set-up has a *propensity*, or *disposition*, to produce a limiting frequency of one-half.

- A different explanatory challenge: explaining the long-run frequency with the probabilities looks like the 'dormative virtue' explanation of Moliére's physicists

Single Case Propensitism To say that the probability of heads is one half is to say that the coin-flipping set up has a *one half* propensity or disposition to land heads.

- Why think that these propensities satisfy the usual laws of probability?

1.3 *Subjective Interpretations*

Bayesian Interpretation According to the *Bayesians*, probabilities represent somebody's *strength of belief* (or the strengths of belief that they *should have*)

You can determine the person's strength of belief—or *credences*—by considering their dispositions to action—for instance, their betting behavior. (More on this later in the course.)

The basic Bayesian picture is this: at any given time, a person has certain credences. Those credences ought to be *probabilities*.

> *Probabilism:* A rational person's credences will obey the laws of probability.

Moreover, when they learn something new, E, they should adopt a new degree of belief, $\mathbb{P}_E$, so that, for any proposition A:

$$\mathbb{P}_E(A) = \frac{\mathbb{P}(A\&E)}{\mathbb{P}(E)}$$

So long as $\mathbb{P}(E) > 0$, this ratio, $\mathbb{P}(A\&E) \div \mathbb{P}(E)$, is equal to the *conditional probability of A, given E*, often written '$\mathbb{P}(A \mid E)$'.

Conditionalization: A rational person will learn from their evidence by conditioning on it. That is, if their *prior* credence function is $\mathbb{P}$ then their *posterior* credence function, after learning, $\mathbb{P}_E$, will be

$$\mathbb{P}_E(A) = \mathbb{P}(A \mid E)$$

In the rest of these notes, we will be considering what is to be said for and against the Bayesian position; so I won't go through the objections here.

Review Questions

1. What two questions is an interpretation of probability trying to answer? Say how the Classical, Logical, Actual Frequentist, and Hypothetical Frequentist interpretations answer these questions.

2. What does the principle of indifference say? What is Bertrand's Paradox, and why does it pose a problem for the principle of indifference?

3. What is the difference between a *state* description and a *structure* description? Suppose we're going to flip a coin three times. What are the state descriptions, and what are the structure descriptions? Assume that the principle of indifference says that each state description should be given the same probability. Then, is Laplace's *Rule of Succession* compatible with the principle of *Conditionalization*?

4. Why are the probabilities recommended by Carnap's Logical Interpretation language-dependent? Why is this a problem for the interpretation?

5. What are two objections to Actual Frequentism? What are two objections to Hypothetical Frequentism?

2
Theories of Probability

2.1 *Probability as Measure*

The standard mathematical theory of probability treats it as a certain kind of *volume*, or *measure*.[1] Formally, it starts with a set, $\mathcal{W}$. We can think of this set a the set of possible ways for the world to be ('worlds'). It's natural to want to be able to measure, or assign a probability to, *any* collection of worlds from $\mathcal{W}$. But over the 20th century, mathematicians encountered difficulties with doing this (more below). So they instead select a collection of *subsets* of $\mathcal{W}$, $\mathcal{A}$. Think of $\mathcal{A}$ as the set of *propositions*. The pair $(\mathcal{W}, \mathcal{A})$ is known as a *measurable space*.

[1] Kolmogorov, A. N. 1950 [1933]. *Foundations of the theory of probability*. New York: Chelsea Publishing Company.

A *probability function*, $\mathbb{P}$, is any function from the set of propositions, $\mathcal{A}$, to real numbers, satisfying the following constraints:

Non-negativity No probabilities are negative. — For all $A \in \mathcal{A}$, $\mathbb{P}(A) \geqslant 0$.

Normalization The necessary truth has probability 1. — $\mathbb{P}(\mathcal{W}) = 1$.

Finite Additivity If you break a proposition up into finitely many non-overlapping parts, then the probability of the whole is the sum of the probabilities of the parts. — For any $A, B \in \mathcal{A}$, if $AB = \varnothing$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

The set $\mathcal{A}$ is required by Kolmogorov to be at least an *algebra*. An *algebra* is a set of propositions which contains $\mathcal{W}$ itself and is closed under complementation and finite union. — $\mathcal{A}$ is an algebra iff: (a) $\neg A \in \mathcal{A}$ whenever $A \in \mathcal{A}$; and (b) $A \cup B \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$.

If $\mathcal{A}$ is an algebra and $\mathbb{P}$ is non-negative, normalized, and finitely additive, then it will have the following properties as well:

Monotonicity If A entails B, then the probability of B is not less than the probability of A. — If $A \subseteq B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$

Strong Additivity The sum of the probabilities of A and B is equal to the sum of the probabilities of $A \cup B$ and AB. — $\mathbb{P}(A) + \mathbb{P}(B) = \mathbb{P}(A \cup B) + \mathbb{P}(AB)$

We can think of a probability function as giving us a "muddy Venn diagram" (van Fraassen's metaphor).

2.2 *Conditional Probability and Independence*

The *conditional* probability of A, given B, is written $\mathbb{P}(A \mid B)$. This says how likely A is, *on the indicative supposition* that B is true. Most everyone accepts that, so long as $\mathbb{P}(B) > 0$,

Product Rule the probability of AB is equal to the product of the probability of A, given B, and the probability of B.

For all $A, B \in \mathcal{A}$, $\mathbb{P}(AB) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B)$

So long as $\mathbb{P}(B) > 0$, we can derive the conditional probability $\mathbb{P}(A \mid B)$ from the unconditional probabilities $\mathbb{P}(AB)$ and $\mathbb{P}(B)$, since in this case, the product rule implies that $\mathbb{P}(A \mid B) = \mathbb{P}(AB) \div \mathbb{P}(B)$. Some say that, if $\mathbb{P}(B) = 0$, then the conditional probability of A, given B, is undefined—though they give fancy ways of defining conditional probabilities *relative to* a partition. Others say that $\mathbb{P}(A \mid B)$ can be unambiguously well-defined even in cases where $\mathbb{P}(B) = 0$.

A *partition* is a set of propositions $\mathbf{A}$ such that, for any two $A, B \in \mathbf{A}$, $AB = \varnothing$, and $\bigcup \mathbf{A} = \mathcal{W}$.

Two propositions are said to be *probabilistically independent* iff the probability of their conjunction is the product of their probabilities.

Independence The propositions A and B are probabilistically independent (according to the probability function $\mathbb{P}$) iff

$$\mathbb{P}(AB) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Given our definition of conditional probability, we have that (so long as $\mathbb{P}(A), \mathbb{P}(B) > 0$) A and B are independent iff $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ and $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

2.3 *Regularity*

Some have wanted to impose the following condition on a probability function:

Regularity If A is *possible*, then the probability of A should be positive.
Equivalently: A's probability is zero only if A is impossible.

In cases where the cardinality of $\mathcal{W}$ is finite, this can be achieved using the standard tools. But in cases where $\mathcal{W}$ is uncountably infinite, it cannot.

To appreciate why, we need to start by appreciating a simple fact about cardinality: a countable union of finite sets is itself countable. This means that, if an uncountably infinite set is the union of countably many sets, then at least one of the sets in that union is infinite.

Then, consider the following infinite collection of sets:

$$\begin{aligned} A_2 &= \{w \in \mathcal{W} \mid \mathbb{P}(\{w\}) \geqslant 1/2\} \\ A_3 &= \{w \in \mathcal{W} \mid \mathbb{P}(\{w\}) \geqslant 1/3\} \\ &\vdots \\ A_n &= \{w \in \mathcal{W} \mid \mathbb{P}(\{w\}) \geqslant 1/n\} \\ &\vdots \end{aligned}$$

Notice that $\mathcal{W} = \bigcup_{n=2}^{\infty} A_n$. If $\mathcal{W}$ is uncountably large, then at least one A_n must be infinite. Take one, choose some $m > n$, and let $A_n[m]$ be some finite subset of A_n with m members. Then, by monotonicity

and finite additivity,

$$\mathbb{P}(A_n) \geqslant \mathbb{P}(A_n[m]) \geqslant \underbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}_{m \text{ times}} = \frac{m}{n} > 1$$

If you want probabilities to be regular, then you will need to block some part of the preceding argument. The way this is usually done is by enriching the range of the probability function. Instead of only allowing *real-valued* probabilities, you also allow *infinitesimal* probabilities (probabilities which are less than every real number but still greater than zero).

2.4 *Infinite Additivity*

There is a further constraint which Kolmogorov says "has been found expedient in researches of the most diverse sort":

Countable Additivity If you break a proposition up into a countable infinity of non-overlapping parts, then the probability of the whole is equal to the infinite sum of the probabilities of the parts.

For any $A_1, A_2, \cdots \in \mathcal{A}$, if $A_i A_j = \varnothing$ for each i, j, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

If a probability function is going to be countably additive, then it will have to be defined over an algebra closed under complementation and *countable* union—this is known as a σ-algebra.

$\mathcal{A}$ is a σ-algebra iff: (a) $\neg A \in \mathcal{A}$ whenever $A \in \mathcal{A}$ and (b) $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ whenever $A_1, A_2, \cdots \in \mathcal{A}$.

To appreciate the difference between finite and countable additivity:

Example 1 (The countably infinite fair lottery)**.** *There is a lottery containing a countably infinite number of tickets. The lottery is fair, so each ticket has an equal probability of winning.*

Some (like Bruno de Finetti), think that countably infinite fair lotteries are possible. But countably infinite fair lotteries are incompatible with countable additivity. We can model this example with a set $\mathcal{W} = \{1, 2, 3, 4, \dots\}$, where the number n represents the world in which ticket number n wins. Countable additivity implies that

$$\mathbb{P}(\{1, 2, 3, \dots\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) + \mathbb{P}(\{2\}) + \dots$$

whereas, if the lottery is *fair*, we must have

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{2\}) = \dots$$

Let $\mathbb{P}(\{1\}) = \alpha$. Then, we have that

$$\mathbb{P}(\{1, 2, 3, \dots\}) = \alpha + \alpha + \alpha + \dots$$

If $\alpha = 0$, then $\mathbb{P}(\{1, 2, 3, \dots\}) = 0$, in violation of normalization. But if $\alpha > 0$, then $\mathbb{P}(\{1, 2, 3, \dots\}) = \infty$, in violation of normalization. So there is cannot be a countably infinite fair lottery, if probabilities are countably additive. Since de Finetti thought countably infinite fair lotteries *were* possible, he rejected countable additivity.

Note that almost *nobody* accepts:

Full Additivity If you break a proposition up into *any* number of non-overlapping parts, then the probability of the whole is equal to the sum of the probabilities of the parts.

For any $\mathbf{A} \subseteq \mathcal{A}$, if $AB = \varnothing$ for each $A, B \in \mathbf{A}$, then $\mathbb{P}(\bigcup \mathbf{A}) = \sum_{A \in \mathbf{A}} \mathbb{P}(A) =_{df} \sup\{\sum_{A \in \mathbf{B}} \mathbb{P}(A) \mid \mathbf{B} \subseteq \mathbf{A} \text{ and } \mathbf{B} \text{ is finite}\}$

Almost nobody accepts this, because almost everybody accepts the possibility of *uncountably* infinite fair lotteries.

Example 2 (the uncountably infinite fair lottery)**.** *We have a circular spinner. After being spun, it is equally likely to land anywhere around the circle's circumference.*

However, if we accepted full additivity, then we'd have to reject the possibility of uncountably infinite fair lotteries. The reasoning is the same as in the case of countably additivity. If the spinner is *fair*, then each point on the circle must have an equal probability of being selected. Call that probability, whatever it is, 'α'. If α is greater than zero, then countably additivity implies that there's some countable subset of points from the circle that has a probability of ∞ of being selected. But, if $\alpha = 0$, then full additivity implies that the probability of the spinner landing *somewhere* $\mathbb{P}([0, 1))$, must be zero, in violation of normalization.

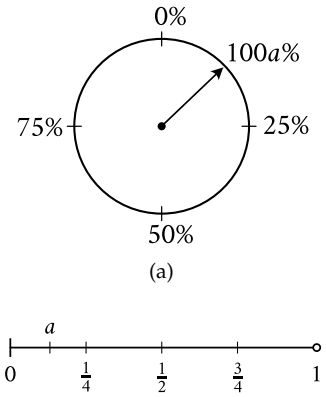


Figure 2.1: In figure 2.1a, the fair spinner. In figure 2.1b, the 'half open' unit interval $[0, 1)$ which we use to model the outcome of the spin.

2.5 *Additivity and Measurability*

Let's think further about the spinner. Here's a general fact, proven by the Italian mathematician Giusepe Vitali: assuming the axiom of choice, there is no probability function for example 2 with the following properties:

(1) Totality: Every set of points in $[0, 1)$ is in $\mathcal{A}$, and so has a probability

(2) Rotation-invariance: For all $A \in \mathcal{A}$, $\mathbb{P}(A) = \mathbb{P}(A \oplus d)$, where $A \oplus d$ is the set A rotated a distance d around the cirlce.

(3) Countable additivity

(The proof of the incompatibility is given in the margin.)

This affords us an argument against countable additivity: every proposition about where the spinner lands should have a rotation-invariant probability. So (1) and (2) should be true. But this implies that countable additivity is false.

Let $\mathbb{Q}$ be the set of rational numbers in $[0, 1)$. $\mathbb{Q} \oplus d = \{q \oplus d \mid q \in \mathbb{Q}\}$. And $\mathcal{R}_\mathbb{Q} = \{\mathbb{Q} \oplus d \mid d \in [0, 1)\}$. $\mathcal{R}_\mathbb{Q}$ is the partition of all the ways of rotating $\mathbb{Q}$ around the circle. For each set in $\mathcal{R}_\mathbb{Q}$, choose a single point v, and let V be the set of points selected. Then, $\mathcal{R}_V = \{V \oplus r \mid r \in \mathbb{Q}\}$ is a *countable* partition of $[0, 1)$. By property (1), each $V \oplus r$ has a probability. By (2), they are all the same probability. By (3), their countable sum is $\mathbb{P}([0, 1)) = 1$. Contradiction.

This is a compelling but unfortunately bad argument. The reason is that (assuming the axiom of choice) there's a conflict between principles like (1) and (2) *on their own*. For there are 'paradoxical' decompositions of a sphere. You can split a sphere up into four disjoint sets, A, B, C, and D, so that B can be rotated into $A \cup B \cup C$ and C can be rotated into $B \cup C \cup D$. (This is known as 'Hausdorff's Paradox', and it's a precursor to the more famous 'Banach Tarski' paradox.) By rotation-invariance, we must have

$$\mathbb{P}(B) = \mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$$

which implies that $\mathbb{P}(A) = \mathbb{P}(C) = 0$. But by another application of

rotation-invariance, we must have

$$\mathbb{P}(C) = \mathbb{P}(B \cup C \cup D) = \mathbb{P}(B) + \mathbb{P}(C) + \mathbb{P}(D)$$

which implies that $\mathbb{P}(B) = \mathbb{P}(D) = 0$. But then $\mathbb{P}(A \cup B \cup C \cup D) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) + \mathbb{P}(D) = 0$. So we cannot have a probaiblity function which both assigns a probability to A, B, C, and D (and their unions) and whose values are rotation-invariant.

But we can make an attempt at a better argument against countable additivity which doesn't appeal to rotation-invariance. Stanislaw Ulam showed that there is no probability function over a measurable space $(\mathcal{W}, \mathcal{A})$ of size $\aleph_1$ such that

(1) Totality: $\mathcal{A}$ contains every subset of $\mathcal{W}$

(2) Non-triviality: there is an uncountable set with positive probability even though every point in the set has probability zero

(3) Countable additivity

So if we assume the continuum hypothesis that $\#[0, 1) = \aleph_1$, then this tells us that, if there's a probability distribution over the landing position of the spinner that gives probability zero to every particular landing spot, then that probability cannot be both countably additive and total (whether it is rotation invariant or not). Some have taken this as an argument for 'unmeasurable' propositions (propositions to which no probability could be assigned). But others have taken it as an argument against countable additivity.

If probabilities are merely finitely additive, then there is no obstacle to assigning (non-trivial) probabilities to *every* proposition.[2]

[2] See K.P.S. Bhaskara Rao, M. Bhaskara Rao. 1983. *Theory of Charges: A Study of Finitely Additive Measures*. New York: Academic Press.

Mathematicians have found rotation-invariance too compelling to give up. They wanted measure (length, area, volume, etc.) to be preserved under rigid rotation and translation. So they decided that not every set of points has a measure. Some sets of points are 'unmeasurable'. Because they understood probability as a particular kind of measure, they decided that not every proposition has a probability.

2.6 Conglomerability

Many have wanted to endorse a principle that's stronger than countable additivity, called 'conglomerability'. According to comglomerability, if $A \in \mathcal{A}$ is any proposition, and $\mathscr{E}$ is any partition, then there will always be some $E_l \in \mathscr{E}$ and some $E_h \in \mathscr{E}$ such that

$$\mathbb{P}(A \mid E_l) \leqslant \mathbb{P}(A) \leqslant \mathbb{P}(A \mid E_h)$$

Another way of expressing the same idea: $\mathbb{P}(A)$ is at most the least upper bound of $\{\mathbb{P}(A \mid E) \mid E \in \mathscr{E}\}$ and at least the greatest lower bound of $\{\mathbb{P}(A \mid E) \mid E \in \mathscr{E}\}$

In other words, it cannot be that, for every $E \in \mathscr{E}$, $\mathbb{P}(A \mid E) < \mathbb{P}(A)$. And it cannot be that, for every $E \in \mathscr{E}$, $\mathbb{P}(A \mid E) > \mathbb{P}(A)$.

Violations of countable additivity lead to violations of conglomerability. (So, by contraposition, conglomerability implies countable additivity.) For an illustrative example, consider the following example.

Example 3 (Two Countably Infinite Fair Lotteries)**.** *There are two inde-*

pendent lotteries, each of which contains a countably infinite number of tickets. Each lottery is fair, so each ticket in each lottery has an equal probability of winning.

We can model this possibility in which ticket n wins the first lottery and ticket m wins the second lottery with an ordered pair, (n, m). Then, we will have $\mathcal{W} = \{(n, m) \mid n, m \in \mathbb{N}\}$. And we can graph these points in the plane, as in figure 2.2.

Since each of these lotteries are *fair* and *independent*, we must have that the probability of each $(n, m) \in \mathcal{W}$ is zero. [Comprehension check: why?]

Let's use 'W_1' for the winning ticket in lottery 1, and 'W_2' for the winning ticket in lottery 2. Then, consider the proposition $W_2 \geqslant W_1$. This is the proposition shown in grey in figure 2.2.

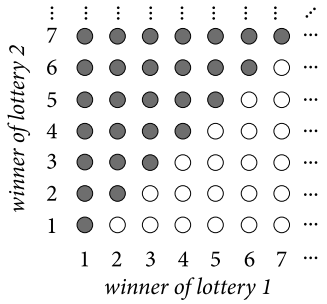


Figure 2.2: Two Countably Infinite Fair Lotteries. In grey, the set of worlds for which the winning ticket in lottery 2 has a number greater than or equal to the winning ticket in lottery 1.

Suppose that the winning ticket in lottery 1 is n, $W_1 = n$. What's your probability that the winning ticket is lottery 2 is $\geqslant n$? Well, there are only finitely many tickets less than n and there are infinitely many tickets $\geqslant n$. If the second lottery is *fair*, then it looks like the probability that the winning ticket in lottery 2 is $\geqslant n$ should be 100%. But this is the answer we will get *no matter which* ticket we suppose won the first lottery. So, for every $n \in \mathbb{N}$,

$$\mathbb{P}(W_2 \geqslant W_1 \mid W_1 = n) = 1$$

But $\{W_1 = n \mid n \in \mathbb{N}\}$ is a partition. So conglomerability implies that

$$\mathbb{P}(W_2 \geqslant W_1) = 1$$

Suppose, on the other hand, that the winning ticket in lottery 2 is n, $W_2 = n$. What's your probability that the winning ticket in lottery 1 is $< n$? Again, there are only finitely many tickets less than n and there are infinitely many tickets $\geqslant n$. If the first lottery is *fair*, then it looks like the probability that the winning ticket in lottery 1 is $< n$ should be 0%. But this is the answer we will get *no matter which* ticket we suppose won the second lottery. So, for every $n \in \mathbb{N}$,

$$\mathbb{P}(W_2 \geqslant W_1 \mid W_2 = n) = 0$$

But $\{W_2 = n \mid n \in \mathbb{N}\}$ is a partition. So conglomerability implies that

$$\mathbb{P}(W_2 \geqslant W_1) = 0$$

But now we've reached a contradiction. If we apply conglomerability to the partition of the outcome of lottery 1, we get one constraint on the probability of $W_2 \geqslant W_1$. And if we apply it to the partition of the outcome of lottery 2, we get a different and incompatible constraint on the probability of $W_2 \geqslant W_1$.

What happened here is fully general. Anytime you have a violation of countable additivity, you'll have a violation of conglomerability.

2.7 Random Variables

If the set of possibilities $\mathcal{W}$ is infinitely large, then we cannot have uniform probabilities defined over every possibility. How, then, do we define a probability function? One common approach is to impose structure on the set $\mathcal{W}$ by parameterizing its members. For instance, if we're interpreting $\mathcal{W}$ as the set of all possible worlds, we might only be interested in how tall Sabeen is. So we can ignore many features of the world, and just assign probabilities to (some) propositions about Sabeen's height.

Sabeen's height is a *variable*. A variable is very much like a *question*, 'how tall is Sabeen?'—it provides us with a partition of possible worlds, which are the possible answers to the question. In addition to giving us this partition, a variable assigns real numbers to each of the possible answers to the question.

Formally, a variable V is a function from $\mathcal{W}$ to $\mathbb{R}$. Given a variable, we can form propositions about the variable's value,

$$\begin{aligned} V = v &\stackrel{\text{def}}{=} \{w \in \mathcal{W} \mid V(w) = v\} \\ V \geqslant v &\stackrel{\text{def}}{=} \{w \in \mathcal{W} \mid V(w) \geqslant v\} \\ V > v &\stackrel{\text{def}}{=} \{w \in \mathcal{W} \mid V(w) > v\} \end{aligned}$$

If every one of these propositions is included in $\mathcal{A}$, then the variable V is said to be *measurable*. A measurable variable is also called a 'random variable'.

Some random variables are *discrete*, meaning that they can take on at most countably many possible real values. For instance, in the infinite fair lottery, we could use the discrete random variable W = which ticket wins. Other random variables are *continuous*, meaning that they can take on any value in some *interval* of the reals. For instance, with the circular spinner, we could use a random variable S = how far around the circle does the spinner stop. S could take on any real value between zero and one.

Given a continuous random variable, V, we can define a probability function over propositions of the form $V \geqslant v$ with what's called a *probability density function*, or 'p.d.f.', f_V. The value of this function, $f_V(v)$, doesn't tell us how probable it is that $V = v$. We know that the probability of $V = v$ is zero. Instead, it tells us how *dense* the probability is at $V = v$. Think about it like this: take a small region around $V = v$ with width 2ϵ, and consider the ratio $\mathbb{P}(v - \epsilon \leqslant V \leqslant v + \epsilon) \div 2\epsilon$. And consider what happens to this ratio as ϵ gets smaller and smaller. This is the *density* of the probability of $V = v$.

In general, we find the probability that the random variable takes a value between a and b by taking the area under the curve f_V in between a and b,

$$\mathbb{P}(a \leqslant V \leqslant b) = \int_a^b f_V(v)\, \mathrm{d}v$$

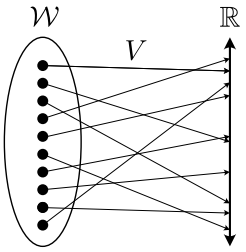
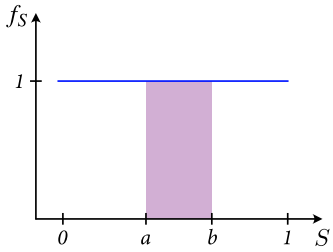
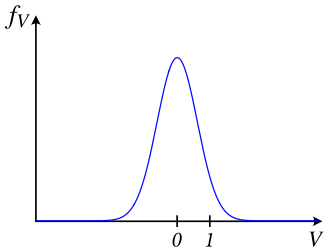


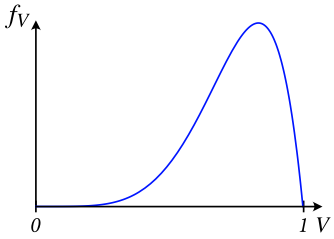
Figure 2.3: A variable maps the worlds in $\mathcal{W}$ to real numbers, $\mathbb{R}$.



(a) A uniform probability density function over S.



(b) A normal (0,1) distribution over V.



(c) A beta (6, 2) distribution over V.

Figure 2.4: Probability density functions. The probability that a random variable takes on a value between a and b is given by the area under the p.d.f. between a and b.

Given a discrete random variable, V, we can define the variable's *expected value*, $\mathbb{E}[V]$, as follows:

$$\mathbb{E}[V] \stackrel{\text{def}}{=} \sum_{v} v \cdot \mathbb{P}(V = v)$$

And, given a continuous random variable, V, with p.d.f. f_V, we may define the variable's *expected value*, $\mathbb{E}[V]$, as

$$\mathbb{E}[V] \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} v \cdot f_V(v) \, \mathrm{d}v$$

Review Questions

1. There are three core axioms of probability theory—what are they? In addition, we saw three other, more controversial probability principles: regularity, countable additivity, and conglomerability. What do these additional principles say? Explain why, if you think that infinite fair lotteries are possible, you will reject both countable additivity and conglomerability.

2. Is countable additivity compatible with the principle of indifference? Why or why not?

3. Explain what an 'unmeasurable' proposition is, and explain why mathematicians think that there are unmeasurable propositions.

4. What is a random variable? What is a probability density function, and how is it different from a probability function? What is the expected value of a random variable?

3
Subjective and Objective Probability

3.1 Two Kinds of Probability

On the Bayesian interpretation, probabilities represent the *degrees of belief* or *credences* of some (actual or idealized) epistemic agent. They are representational states. On this interpretation, to say "rain is more likely than snow" is just to say "I'm more confident that it will rain than I am that it will snow", and to say that "The probability of rain is 90%" is just to say that you are 90% sure that it will rain".

But quantum mechanics *seems* to involve probabilities that aren't just the degrees of belief of any particular agent. Lewis wants to accommodate the kinds of objective probabilities found in (some interpretations of) quantum mechanics. He is a Bayesian, but he's a pluralist Bayesian. He thinks that credences are probabilities, but he doesn't think that they are the *only* kind of probability out there. He believes that there are *both* subjective probabilities or credences *and* objective probabilities or chances.

Terminology: within philosophy, it has become customary to reserve the word 'chance' for objective probabilities.

In *A Subjectivist's Guide to Objective Chance*, Lewis is uncertain how to understand these kinds of objective chances. He is opposed to 'non-Humean' whatsits. So he ultimately wants to analyze all talk of chance in terms of the 'Humean mosaic' of the spatiotemporal arrangement of local, perfectly natural properties and relations. Nonetheless, Lewis believes that there is a connection between these two probabilities.

Lewis accepted the thesis of Humean Supervenience: that all facts supervene on "the spatiotemporal arrangement of local qualities throughout all of history, past, present and present and future."

3.2 The Principal Principle

Lewis thinks that you can have credences about the objective chances, just as you can have credences about any other facts about the world. And, moreover, he thinks that your credences about the chances constrain your credences about other matters.

3.2.1 Questionnaire

Q1: Later today, I'll toss a fair coin which has a 50% chance of landing heads and a 50% chance of landing tails. To what degree should you believe that it will land heads?

A1: 50%

Q2: You have the evidence that the center of mass of the coin is displaced, that 90 of the past 100 flips have landed tails, and that duplicates of this coin land landed tails in about 90% of tosses. Yet you remain certain that the coin has a 50% chance of landing heads. To what degree should you believe that it will land heads?

A2: 50%

Q3: The coin has already been tossed. You remain certain that it had (at the time of the toss) a 50% chance of landing heads. You saw it land tails. To what degree should you believe that it landed heads?

A3: Nearly 100%.

Lewis wants a general principle that accounts for these judgments. He calls it "the principal principle", since he believes it captures everything we know about objective chance.

The Principal Principle If $\mathbb{C}_0$ is a reasonable initial (or *ur-prior*) credence function, A is any proposition, $\langle C\, h_t(A) = x\rangle$ is the proposition that the time t chance of A is x, and E is any proposition compatible with $\langle C\, h_t(A) = x\rangle$ that is *admissible* at time t, then

$$\mathbb{C}_0(A \mid \langle C\, h_t(A) = x\rangle\ E) = x$$

Notice that Lewis doesn't require that A is in the domain of the credence function $\mathbb{C}_0$, nor that $\mathbb{C}_0(E) > 0$. That's because Lewis accepted Regularity (he thought credences could be infinitesimal), and he thought that they should be defined over any proposition.

3.2.2 *Reasonable Initial Credence*

In the basic, off-the-shelf Bayesian model, we said that your credences change, over time, by conditioning them on what you've learned. If you start out with the credence function $\mathbb{C}$, learn E, and then end up with the new credence function $\mathbb{C}_E$, $\mathbb{C}$ is called your 'prior' credence function, and $\mathbb{C}_E$ is called your 'posterior' credence function. And conditionalization said that your posterior should be your prior, conditioned on your newly acquired evidence, E. That is: $\mathbb{C}_E(A)$ should be $\mathbb{C}(A \mid E)$ (which is just $\mathbb{C}(AE) \div \mathbb{C}(E)$, so long as $\mathbb{C}(E) > 0$.)

$$\underbrace{\mathbb{C}}_{\textit{prior}} \xrightarrow{\textit{learn } E} \underbrace{\mathbb{C}_E}_{\textit{posterior}}$$

Lewis has a slightly different picture in mind. To appreciate this picture, notice that your prior *itself* is the result of a previous learning experiences; and the prior of *that* previous learning experience was the result of a previous learning experience before *that*, and so on and so forth.

$$\cdots \mathbb{C} \xrightarrow{\textit{learn } H} \mathbb{C}_H \xrightarrow{\textit{learn } G} \mathbb{C}_{HG} \xrightarrow{\textit{learn } F} \mathbb{C}_{HGF} \xrightarrow{\textit{learn } E} \mathbb{C}_{HGFE}$$

If you are rational, then each of these priors were reasonable to hold, given the evidence you had acquired up to that point. Lewis thinks that we can extend further back, and ask about which opinions it would've been reasonable to hold in the absence of *any* evidence. Now, by the time you were in a position to start having degrees of belief, *you* already had lots of evidence. But we can imagine a *superbaby* who comes into existence with the ability to form credences, but who has yet to undergo any experiences or acquire any evidence. Lewis thinks that

there would be credences that would be unreasonable for superbaby to adopt; and there are credences that would be reasonable for superbaby to adopt. He calls these reasonable 'initial' or 'ur-prior' credences.

Then, Lewis accepts a slightly different learning norm (in addition to conditionalization):

> *Ur-prior Conditionalization* For all times t, if E is your total evidence at time t and $\mathbb{C}_t$ is a reasonable credence function for you to have at t, then there is a reasonable initial credence function $\mathbb{C}_0$ such that, for any proposition A, $\mathbb{C}_t(A) = \mathbb{C}_0(A \mid E)$

If we think that there is just a single reasonable initial credence, then ur-prior conditionalization will entail conditionalization. But if there is more than a single reasonable initial credence, then the two norms are logically independent. Suppose that there are two reasonable initial credences, $\mathbb{C}_0$ and $\mathbb{C}'_0$. And suppose that, on even days, your credence is $\mathbb{C}_0$ conditioned on your total evidence; whereas, on odd days, your credence is $\mathbb{C}'_0$ conditioned on your total evidence. Then, you'll satisfy ur-prior conditionalization but not conditionalization. In the other direction, suppose that there's some *unreasonble* initial credence $\mathbb{C}''_0$ such that whenever your total evidence is E, your credences are given by $\mathbb{C}''$ conditioned on E. (And suppose you never lose evidence.) Then, you'll satisfy conditionalization but not ur-prior conditionalization.

Together with ur-prior conditionalization, the Principal Principle entails:

> *The Current Principle* If your total evidence at time t is time t admissible, and $\langle \mathcal{C}h_t(A) = x\rangle$ is compatible with your total evidence at t, then a reasonable time t credence in any proposition A, given that the time t chance of A is x, is x.
>
> $$\mathbb{C}_t(A \mid \langle \mathcal{C}h_t(A) = x\rangle) = x$$

3.2.3 *The Proposition $\langle \mathcal{C}h_t(A) = x\rangle$*

'$\mathcal{C}h_t$' is a definite description: 'the time t chance function'. Like other definite descriptions ('the first postmaster general', 'the morning star'), it picks out different things in different possible worlds. Similarly, '$\mathcal{C}h_t(A)$' is the definite description 'the time t chance of A'. So $\mathcal{C}h_t(A)$ is what we called last class a *random variable*. It's a function from worlds in $\mathcal{W}$ to real numbers between zero and one. The value this random variable takes on a world w—which I'll write '$\mathcal{C}h_{w,t}(A)$'—is the objective chance that A has at the world w and the time t.

In contrast, '$\mathbb{C}_0$' is not a definite description. It is a name for a *particular* probability function. It picks out the very same probability function in every possible world.

As a matter of notation, I'll write functions in blackboard boldface (like $\mathbb{P}$, $\mathbb{C}$, or $\mathbb{C}h_t$) when I'm talking about a particular probability function; and I'll write functions in calligraphic font (like $\mathcal{P}$, $\mathcal{C}$, $\mathcal{E}$, or $\mathcal{C}h_t$) when I'm talking about a definite description for a probability function.

Lewis held that—at least putting aside *de se* propositions which we'll come back to later in the course—your credences were defined over sets of worlds, just as in the standard Kolmogorov theory of probability. So, for him, the proposition $\langle \mathcal{C}h_t(A) = x\rangle$ is just the set of worlds in which the objective chance of A is x at time t,

$$\langle \mathcal{C}h_t(A) = x\rangle = \{w \in \mathcal{W} \mid \mathcal{C}h_{w,t}(A) = x\}$$

In the proposition $\langle \mathcal{C}h_t(A) = x\rangle$, we are allowed to replace 'x' with any real number between 0 and 1. But we are not allowed to replace 'x' with a definite description. If we could, then the principal principle would give terrible advice. Suppose we could replace 'x' with the definite description '$\mathcal{C}h_t(A)$'. Then, the principle would imply that $\mathbb{C}_0(A \mid \langle \mathcal{C}h_t(A) = \mathcal{C}h_t(A)\rangle) = \mathcal{C}h_t(A)$. Since $\langle \mathcal{C}h_t(A) = \mathcal{C}h_t(A)\rangle$ is necessarily true, this reduces to $\mathbb{C}_0(A) = \mathcal{C}h_t(A)$. But then, the princi-

For any probability function $\mathbb{P}$, and any propositions $A, B \in \mathcal{A}$, if $\mathbb{P}(B) = 1$, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. [Can you say why?]

ple would require you to know *a priori* the time t chance of any proposition.

Lewis doesn't assume that every proposition has a chance—for instance, perhaps there's no objective chance that murder is wrong. So it could be that, for some proposition, A, $\langle Ch_t(A) = x\rangle = \varnothing$. [What does the principle say in that case?]

Lewis has explicitly assumed that chances are time-dependent. As time passes, the objective chance of an outcome can go up or down. Suppose, for instance, that you enter a maze and at every turn, you flip a coin to decide whether to turn left or right. As you move through the maze, the chance that you find your way out may change. At the start of the maze, half of the paths lead to an exit. Since each path is equally likely to be taken, the chance of exit is one half. However, after some bad luck initially, you end up at a point in the maze where only one third of the paths ahead lead to exit. At that point, the chance of exit has fallen to one third.

3.2.4 *Admissibility*

The clause about E being *admissible* for the time t is needed to make sure that Lewis can give what he takes to be the right answers to questions 2 and 3 from his questionnaire. For question 2: he wants to be able to say that you should continue setting your credences to the chances, given any amount of information about times *before* t. So he wants information about the chances to screen off any other information about the past. So any amount of information about times before t must be admissible. For question 3: he wants to be able to say that, once you've gotten information about times *after* t, you no longer have to set your credences to the chances.

A proposition E is said to 'screen off' A from B iff A and B are independent, conditional on E. That is, E screens off A from B iff $\mathbb{P}(A \mid EB) = \mathbb{P}(A \mid E)$.

Lewis doesn't provide necessary and sufficient conditions for E being time t admissible. As I read him (following Chris Meacham) this is mainly because he is worried about cases involving time travel into the past. However, if we place these situations to the side, then he accepts the following two sufficient conditions for admissibility:

See Meacham. 2010. "Two Mistakes Regarding the Principal Principle", in *the British Journal for the Philosophy of Science*. 61: 407–431.

> *Admissibility* E is time t admissible if either (1) E is *about* times before t; or (2) E is a counterfactual conditional saying how chance counterfactually depends upon history.
>
> Boolean combinations of admissible propositions are themselves admissible.

To understand these sufficient conditions, some notation: let '$H_{w,t}$' be the *history* of w up to time t. It is the set of worlds which perfectly agree with w up to the time t. And let $\mathcal{H}_t = \{H_{w,t} \mid w \in \mathcal{W}\}$ be the partition of worlds by their time t histories. When Lewis says that E is *about* times before t, he means that it is a union of cells from the partition $\mathcal{H}_t$.

Lewis gives a more careful definition of *aboutness* elsewhere that he appeals to here. (See his *Relevant Implication*.) Quickly: a *subject matter*, $\mathcal{M}$, is a partition of the set of possible worlds, the cells of which correspond to ways things might be with respect to that subject matter. And a proposition is *about* $\mathcal{M}$ iff it does not distinguish between worlds within a cell of $\mathcal{M}$. That is: for any $M \in \mathcal{M}$ and any two $w, w' \in M$, either both or neither of w and w' are included in the proposition. It follows that the proposition is a union of cells from $\mathcal{M}$.

When Lewis says that E is a conditional saying how chance depends upon history, he means that it is a conditional of the form $H_{w,t^*} \boxrightarrow$

$\langle C\, h_{t^*}(A) = x\rangle$ or $H_{w,t^*} \boxrightarrow \langle C\, h_{t^*} = \mathbb{P}\rangle$, where '$\mathbb{P}$' is a particular probability function, so that $\langle C\, h_{t^*} = \mathbb{P}\rangle$ is the set of worlds w such that $C\, h_{w,t^*} = \mathbb{P}$. Importantly, we needn't have $t^* = t$ in order for these propositions to be admissible at the time t. Information about how *future* chance counterfactually depends upon the *future* can be admissible.

3.3 *Reformulation*

Lewis uses the sufficient conditions on admissibility to offer a reformulation of the principal principle. Let's start with the idea of a *time t theory of chance* for a world, which we ca denote '$T_{w,t}$' ('T' for *t*heory). A time t theory of chance for a world says how, at that world, the chances counterfactually depend upon the history of the world up to the time t. That is, for each $H_t \in \mathcal{H}_t$, $T_{w,t}$ gives us a counterfactual of the form $H_t \boxrightarrow \langle C\, h_t = \mathbb{P}\rangle$ which is true at w.

It's worth pausing here for a second—why does Lewis get to assume that there *is* a true counterfactual like this? There's a famous debate between Lewis and Stalnaker about the principle known as 'counterfactual excluded middle': $(A \boxrightarrow C) \vee (A \boxrightarrow \neg C)$. Stalnaker accepted it, but Lewis rejected it. A quantified version of the principle says that $A \boxrightarrow (\exists x)Fx \vdash (\exists x)(A \boxrightarrow Fx)$. Lewis denied that this was valid. He gave examples like the following: if I were taller than 6 feet, then there's some height I'd be. But it's not the case that there's some height h such that, if I were taller than 6 feet, my height would be h. It's not that I'd be 6'1', since a world where I'm a half inch taller than 6 feet would be more similar to the actual world; and it's not that I'd be a half inch taller than 6 feet, since a world where I'm a quarter inch taller than 6 feet would be more similar still. But here, Lewis seems happy to say *not only* that, if the history were H_t, then there would be some $\mathbb{P}$ such that $\mathbb{P}$ is the objective chance function, $H_t \boxrightarrow (\exists \mathbb{P})\langle C\, h_t = \mathbb{P}\rangle$, but *moreover*, that there's some function $\mathbb{P}$ such that, were the history H_t, $\mathbb{P}$ would be the objective chance function, $(\exists \mathbb{P})(H_t \boxrightarrow \langle C\, h_t = \mathbb{P}\rangle)$. This is a 'special case' of counterfactual excluded middle that Lewis was willing to accept.[1]

[1] See in particular the discussion in section 10, page 27, of his *Causal Decision Theory*.

So, given any world w and any time t, the time t theory of chance for world w will be a conjunction of counterfactuals specifying how the time t objective chance function counterfactually depends upon the history of the world up until time t,

$$T_{w,t} = \bigcap_{H_t \in \mathcal{H}_t} H_t \boxrightarrow \langle C\, h_t = \mathbb{P}\rangle$$

For each conjunct, we choose the probability function $\mathbb{P}$ which makes that conjunct true at the world w.

The *complete* theory of chance for a world w is just the conjunction of the time t theory of chance for w, for every time t.

$$T_w = \bigcap_t T_{w,t} = \bigcap_t \bigcap_{H_t \in \mathcal{H}_t} H_t \boxrightarrow \langle C\, h_t = \mathbb{P}\rangle$$

The complete theory of chance for a world w is admissible at every time. And the history of the world up until the time t is admissible at t. So the principal principle implies that, whenever $H_{w,t}T_w$ is compatible with $\langle C\,h_t(A) = x\rangle$, we will have

$$\mathbb{C}_0(A \mid \langle C\,h_t(A) = x\rangle H_{w,t}T_w) = x$$

Moreover, whenever $H_{w,t}T_w$ is compatible with $\langle C\,h_t(A) = x\rangle$, it will *entail* $\langle C\,h_t(A) = x\rangle$. For the antecedent $H_{w,t}$, together with the conditional $H_{w,t} \mathrel{\square\!\!\rightarrow} \langle C\,h_t = \mathbb{P}\rangle$ will entail $\langle C\,h_t = \mathbb{P}\rangle$. And if this is compatible with $\langle C\,h_t(A) = x\rangle$, then it must be that $\mathbb{P}(A) = x$. So $\langle C\,h_t = \mathbb{P}\rangle$ must entail $\langle C\,h_t(A) = x\rangle$. So $H_{w,t}T_w$ entails something which entails that $\langle C\,h_t = \mathbb{P}\rangle$—so it entails it.

Whenever one proposition, E, entails another, F, $EF = E$. So $\langle C\,h_t(A) = x\rangle H_{w,t}T_w = H_{w,t}T_w$. And if $H_{w,t}T_w$ entails $\langle C\,h_t(A) = x\rangle$, then $C_{w,t}(A) = x$. So the principal principle implies that, for any w and any t,

$$\mathbb{C}_0(A \mid H_{w,t}T_w) = C\,h_{w,t}(A)$$

We thus have:

> *The Principal Principle Reformulated* If $\mathbb{C}_0$ is a reasonable initial credence function, w is any world, t any time, and A any proposition in the domain of $C\,h_{w,t}$, then
>
> $$C\,h_{w,t}(A) = \mathbb{C}_0(A \mid H_{w,t}T_w)$$

Does the reformulated principal principle imply the original? Lewis suggests that it *almost* entails the original—but not quite.

To see how the reverse entailment might hold, notice that $\{H_{w,t}T_w \mid w \in \mathcal{W}\}$ is a partition, and that (1) the proposition $\langle C\,h_t(A) = x\rangle$ is a union of cells from this partition; and (2) any proposition which meets Lewis's sufficient condition for admissibility will be a union of cells of the partition.

If this were a finite partition, it would follow straightaway that $\mathbb{C}_0(A \mid \langle C\,h_t(A) = x\rangle E) = x$. The reason is that any finitely additive probability will be conglomerable over a finite partition. That is: if $\mathbb{C}_0$ is a finitely additive probability, $\mathscr{E}$ is a finite partition of the proposition F,[2] and $\mathbb{C}_0(A \mid E) = x$ for every $E \in \mathscr{E}$, then $\mathbb{C}_0(A \mid F) = x$, too. The conjunction $\langle C\,h_t(A) = x\rangle E$ is a disjoint union of propositions of the form $H_{w,t}T_w$ (with t fixed and w variable), and for every one of these propositions, we will have $\mathbb{C}_0(A \mid H_{w,t}T_w) = x$. So, if $\{H_{w,t}T_w \mid w \in \mathcal{W}\}$ were a finite partition, we would have that $\mathbb{C}_0(A \mid \langle C\,h_t(A) = x\rangle E) = x$.

So we can *almost* recover the original principal principle from the reformulation. But we cannot quite get there—for two reasons. Firstly, as we saw last class, conglomerability needn't hold if $\mathcal{E}$ is an *infinite* partition. And as Lewis says, "indeed we would expect the history-theory partition to be infinite". Secondly, even if we had conglomerability over the history-theory partition, we would only have derived

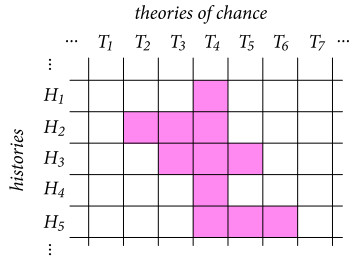


Figure 3.1: The partition $\{H_{w,t}T_w \mid w \in \mathcal{W}\}$. Any proposition which meets Lewis's sufficient conditions for admissibility will be a union of the cells of this partition (like the shaded proposition above).

[2] $\mathscr{E}$ is a partition *of* F iff $\bigcup \mathscr{E} = F$ and any two propositions in $\mathscr{E}$ are disjoint.

the original principle *for the time t admissible propositions which are covered by Lewis's sufficient conditions*. We wouldn't have derived it for other admissible propositions (and, officially, Lewis is neutral about whether there are other admissible propositions).

3.4 Consequences of the Principal Principle

Lewis thinks that the principal principle "capture[s] all we know about chance". Even though it only concerns the *connection* between chance and reasonable (initial) credence, we can use it to show various things about objective chance. In particular, Lewis highlights the following consequences of the principal principle:

1. First consequence: the past is no longer chancy–the objective chance of any historical proposition is always one. To appreciate this consequence, suppose (for *reductio*) that there was some world w such that $Ch_{w,t}(H_{w,t}) = x$, for some x less than one. Then, $H_{w,t}$ is compatible with $\langle Ch_t(H_{w,t}) = x\rangle$. Moreover, $H_{w,t}$ is admissible. So the principal principle implies that $\mathbb{C}_0(H_t \mid H_t\langle Ch_t(H) = x\rangle) = x$. But it follows from the definition of conditional probability that $\mathbb{C}_0(H_t \mid H_t\langle Ch_t(H) = x\rangle) = 1$. Contradiction. So there is no world w such that $Ch_{w,t}(H_{w,t}) = x$, for some x less than one. So, for every world, w, $Ch_{w,t}(H_{w,t}) = 1$.
2. Second consequence: the chances are probabilities. This follows from the reformulated version of the principal principle, together with the observation that a conditional probability function is itself a probability function. [Why?] I want to emphasize that this consequence is slightly deeper than Lewis lets on. We saw last class that there is much controversy about which principles probability satisfies—is it regular? countably additive? conglomerable? According to the principal principle, once we've answered these questions for *reasonable initial credence*, we have thereby also answered them for *objective chance*.

   According to Lewis, reasonable initial credence is regular. So the objective chances are regular in a restricted sense: at the world w and the time t, they give positive probability to any possibility compatible with the history of w at time t and the theory of chance for w. That is: because Lewis thinks that credences can be infinitesimal, he thereby thinks that objective chances can be infinitesimal, too.
3. Third consequence: future chances come from past chances by conditioning on the intervening history. Take two times, t_1 and t_2. Let I be a complete description of everything that happens in between t_1 and t_2. Then, $H_{w,t_2} = H_{w,t_1}I$. By the reformulated principal principle,

$$Ch_{w,t_2}(A) = \mathbb{C}_0(A \mid H_{w,t_2}T_w) = \mathbb{C}_0(A \mid H_{w,t_1}IT_w) = Ch_{w,t_1}(A \mid I)$$

4. Fourth consequence: observed frequencies can provide evidence about the chances. For instance: suppose that we are going to flip a fair coin nine times, and we are considering two chance hypotheses: firstly, that the coin flips have an independent 1/3rd chance of landing heads

($H_{1/3}$). Secondly, that the coin flips have an independent 2/3rds chance of landing heads ($H_{2/3}$). We observe that 3 of the ten flips land heads. Call this evidence 'E'. Let 1 be the time before the flips have taken place, and let 2 be the time after the flips have taken place. Then, using *the current principle* and the odds ratio form of Bayes' rule,[3]

$$\begin{aligned}\frac{\mathbb{C}_2(H_{1/3})}{\mathbb{C}_2(H_{2/3})} &= \frac{\mathbb{C}_1(E \mid H_{1/3})}{\mathbb{C}_1(E \mid H_{2/3})} \cdot \frac{\mathbb{C}_1(H_{1/3})}{\mathbb{C}_1(H_{2/3})} \\ &= \frac{\left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^6}{\left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^6} \cdot \frac{\mathbb{C}_1(H_{1/3})}{\mathbb{C}_1(H_{2/3})} \\ &= \frac{\left(\frac{2}{3}\right)^3}{\left(\frac{1}{3}\right)^3} \cdot \frac{\mathbb{C}_1(H_{1/3})}{\mathbb{C}_1(H_{2/3})} \\ &= 8 \cdot \frac{\mathbb{C}_1(H_{1/3})}{\mathbb{C}_1(H_{2/3})}\end{aligned}$$

which means that this observation has made $H_{1/3}$ eight times more likely than $H_{2/3}$ than it was before the observation was made.

[3] This says that

$$\frac{\mathbb{P}(A \mid E)}{\mathbb{P}(B \mid E)} = \frac{\mathbb{P}(E \mid A)}{\mathbb{P}(E \mid B)} \cdot \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

This is really crucial: for Bayesians, it is the principal principle which allows the observation of frequencies to raise their credences in hypotheses about the objective chances. Chances that make the frequencies more likely are confirmed. Chances that make the frequencies less likely are disconfirmed.

3.5 *Another Reformulation*

The third consequence tells us that the objective chance function evolves through time by conditioning it on the intervening history. We can think of the objective chances like an agent who learns what has happened throughout history up to the present moment. As time rolls on, it learns everything about what has transpired.

But then, just as we thought about an ur-prior *credence* function, we can also think about the ur-prior *chance* function at a world, w: $Ch_{w,0}$. This is the function which, when conditioned on the proposition $H_{w,t}$, gives $Ch_{w,t}$. We could use a Lewisian theory of chance to determine such a function; if $H_t \boxrightarrow \langle Ch_t = \mathbb{P}\rangle$ is true at w, then $Ch_{w,0}(- \mid H_t) = \mathbb{P}(-)$.[4] And we could use such a function to determine a Lewisian theory of chance. If $Ch_{w,0}(- \mid H_t) = \mathbb{P}$, then we could stipulate that $H_t \boxrightarrow \langle Ch_t = \mathbb{P}\rangle$ is true at w. Going back-and-forth between claims about ur-chance and claims about a Lewisian theory of chance in this way relies upon some assumptions about the relationship between counterfactuals and chance that we should probably interrogate further—and that one of you could profitably interrogate further for a term paper. But let's assume that we can think about the theory of chance in this way. Then, we could give an alternative reformulation of the Principal Principle:

[4] When I write '$Ch_{w,0}(- \mid H_t)$', I'm just referring to a function that you hand a proposition, A, and it hands you back the value of $Ch_{w,0}(A \mid H_t)$.

The Principal Principle Reformulated Again If $\mathbb{C}_0$ is a reasonable initial credence function, A is any proposition, and $\langle C\, h_0 = \mathbb{P}\rangle$ says that the ur-prior chances are given by $\mathbb{P}$, then

$$\mathbb{C}_0(A \mid \langle C\, h_0 = \mathbb{P}\rangle) = \mathbb{P}(A)$$

3.6 The Principal Principle and Humean Supervenience

The Principal Principle tells us many things about chance. In fact, it tells us enough to rule out several interpretations of chance—including Lewis's own.

3.6.1 The Best Systems Account

Distinguish two views about the metaphysics of objective chance: *Humeans* hold that facts about chance supervene on "the spatiotemporal arrangement of local qualities throughout all of history, past, present, and future". In other words: once you tell me exactly what happens throughout the entire history of the universe, you will have told me all that there is to tell about objective chance. No two worlds differ in *chance* without differing in chance *outcomes*.

For instance: the actual frequentist is a Humean about objective chance. Once you tell me everything that's happened, you've told me enough to know what the actual frequencies are, which is enough to tell me what the objective chances are.

Lewis wasn't an actual frequentist, but he was a Humean. His views about chance were integrated with his views about laws of nature—for him, the objective chances were just a particular kind of law of nature. His view about laws is called the 'Best Systems Analysis' of laws. To understand the view, imagine going to God and asking him to tell you what the world is like. He starts off by saying "ok, well, at time t_0, particle p is at location x with momentum m, and...". You interrupt, and tell God: "Ok, look, I don't have time to get all of the details—could you give me an executive summary?" At that point, God tells you "Ok, well, every particle's acceleration is equal to the resultant forces acting on it divided by its mass, and ..." and then God lists off the things we think of as laws of nature. On Lewis's view, all that it is for something to be a law is for it to be a highly informative and simple 'executive summary' of the truths at a world.

More carefully, consider all of the possible deductive, axiomatic systems (some collection of axioms closed under deductive consequence). Some of these systems will be highly informative (strong). Others will be particularly simple, with a small number of short axioms. Strength and simplicity are both virtues of an axiomatic system, and these virtues compete. As the system becomes simplier, it typically becomes less informative; and as it becomes more informative, it typically becomes less simple. Consider the axiomatic system which strikes the *best balance* of simplicity and strength. According to Lewis, the generaliza-

tions which follow from that 'best' axiomatic system are the laws of nature.

At least—that's the story about deterministic laws. But what about *chancy* laws (like the ones found in some interpretations of quantum mechanics)? According to Lewis, sometimes the best way to strike a balance between simplicity and strength is to give, not a universal generalization, but rather a *chance law*. For instance, if about half of the flips of the coin land heads and about a half of the flips of the coin land tails, then a good 'executive summary' would tell you that there's a one half chance of the coin landing heads on any particular flip. In the case of chance laws, informativeness is to be understood in terms of 'fit' (how likely the chance laws make the actual world). It is a virtue of a chance law that it makes the actual world's history likely. But this virtue trades off against the virtue of simplicity. The actual chance laws are the ones that strike the best balance of simplicity and fit.

3.6.2 *The 'Bug' in the Principal Principle*

The bombshell: Lewis's theory of chance seems to be incompatible with his principal principle.

Suppose that the only thing that will happen throughout the history of the universe is that we will flip a coin a hundred times. We can model this case with a set of possibilities, $\mathcal{W}$, where each $w \in \mathcal{W}$ is a sequence of 'H's and 'T's with 100 entries.

Consider two worlds: w_f and w_u. At w_f, we get 50 heads landings and about 50 tails landings, and there's no discernible pattern in these landings. For instance, it's not $HTHTHTHT\ldots$. At w_u, the coin lands heads on each and every flip. That is, $w_u = HHHHH\ldots H$ (a sequence of 100 'H's).

The chance law which says each coin has a 50% independent chance of landing heads will fit w_f better than any other; and it is simpler than other competitor chance laws. So, at w_f, the objective chance of the coin landing heads on any flip is one half, and the outcome of different flips are probabilistically independent of each other.

On the other hand, the chance law which says that the coin has a 100% chance of landing heads on each flip fits w_u better than any other. And it is very simple. So, at w_u, the objective chance of the coin landing heads on any—and, therefore, every—flip is 100%.

At w_f, the objective chance of the coin landing heads on each and every flip is $(1/2)^{100}$. So

$$Ch_{w_f,0}(w_u) = (1/2)^{100}$$

This means that $\langle Ch_0(w_u) = (1/2)^{100}\rangle$ is not empty. It at least contains w_f.

However, $\langle Ch_0(w_u) = (1/2)^{100}\rangle$ does *not* contain w_u. For w_u is *incompatible* with the objective chance of w_u being $(1/2)^{100}$. At w_u, the objective chance of w_u is 100%. In general, if A and B are incompat-

ible (and the probability of B is non-zero), then the probability of A, conditional on B, must be zero.

So

$$\mathbb{C}_0(w_u \mid \langle C\, h_0(w_u) = (1/2)^{100}\rangle) = (1/2)^{100} \quad \text{by the Principal Principle, but}$$
$$\mathbb{C}_0(w_u \mid \langle C\, h_0(w_u) = (1/2)^{100}\rangle) = 0 \quad \text{by probabilism}$$

Contradiction. So by assuming Lewis's theory of chance and his principal principle, we arrived at a contradiction. So the principal principle is incompatible with his theory of objective chance.

More generally, the problem is that, according to Humean theories of chance, chance can be *modest*—the chance function can fail to know that *it is* the objective chance function.[5] That is, we can have a world w and a time t such that

$$Ch_{w,t}(\langle C\, h_t = Ch_{w,t}\rangle) < 1$$

But, by the second reformulation of the principal principle, we must have

$$\mathbb{C}_0(\langle C\, h_t = Ch_{w,t}\rangle \mid \langle C\, h_t = Ch_{w,t}\rangle) = Ch_{w,t}(\langle C\, h_t = Ch_{w,t}\rangle)$$

By the definition of conditional probability, the right hand side must be 1 whenever it is defined. So the principal principle requires the left hand side to be 1 also. Which means that the principal principle requires objective chance to be *immodest*—it must be certain that it is the objective chance function.

[5] The terminology of 'modesty' and 'immodesty' comes from thinking of the objective chance function as a kind of expert. If an expert is *modest*, then they are not certain that they are the expert. And if the expert is *immodest*, then they are certain that they are the expert.

3.7 *The New Principle*

One reaction to this is to reject Lewis's theory of chance. However, that's not the reaction of Lewis himself. He instead took up the suggestion of Michael Thau and Ned Hall that the 'old' principle be revised.[6]

According to Hall, we should think of the principal principle as an instance of a more general kind of principle: a principle of *expert deference*. In general, an expert in some domain is someone whose probabilities about the propositions in that domain you regard as more trustworthy than your own. And the objective chance function has a particularly wide domain of expertise. But Hall thinks we should distinguish two different kinds of expert: what he calls a *database* expert and an *analyst* expert. A database expert is a function you should trust and defer to because it knows so much more than you. For a database expert, $\mathcal{E}$, you should satisfy the kind of constraint imposed by the original principal principle, and set $\mathbb{C}(A \mid \langle \mathcal{E} = \mathbb{E}\rangle) = \mathbb{E}(A)$. In contrast, an *analyst* expert is a function you should trust and defer to *not* because it has more information than you, but rather because it is better at evaluating the evidential bearing of some propositions on another—it is better at *analyzing* evidence, even though it may not have as much evidence as you do.

[6] See Lewis's 'Humean Supervenience Debugged', Thau's "Undermining and Admissibility", and Hall's 'Correcting the Guide to Objective Chance', and 'Two Mistakes about Credence and Chance'.

When it comes to an analyst expert, you shouldn't just defer to them straightaway—you may have some relevant evidence the analyst expert lacks. Instead, you should bring the expert up to speed by conditioning their probability function on any information you are taking for granted, and only defer to them *then*. If the analyst expert is modest—if they don't know that they *are* the expert—then you will also have to bring them up to speed on *this* information.

In general, for an analyst expert, if E is your total evidence, then you should have

$$\mathbb{C}(A \mid \langle \mathcal{E} = \mathbb{E} \rangle) = \mathbb{E}(A \mid \langle \mathcal{E} = \mathbb{E} \rangle E)$$

Or, appealing to ur-prior conditionalization, you should have, for any proposition E,

$$\mathbb{C}_0(A \mid \langle \mathcal{E} = \mathbb{E} \rangle E) = \mathbb{E}(A \mid \langle \mathcal{E} = \mathbb{E} \rangle E)$$

Then, Hall contends that objective chance is a modest analyst expert. So, we should have that, for any E,

$$\mathbb{C}_0(A \mid \langle C\, h_t = \mathbb{P} \rangle E) = \mathbb{P}(A \mid \langle C\, h_t = \mathbb{P} \rangle E)$$

In particular, take any world w, any time t, let $E = H_{w,t}T_w$, and consider the probability function $C\,h_{w,t}$. Then,

$$\mathbb{C}_0(A \mid \langle C\, h_t = C\,h_{w,t} \rangle H_{w,t}T_w) = C\,h_{w,t}(A \mid \langle C\, h_t = C\,h_{w,t} \rangle H_{w,t}T_w)$$

The conjunction $H_{w,t}T_w$ entails (by *modus ponens*) that $\langle C\,h_t = C\,h_{w,t} \rangle$. So we could re-write this as

$$\mathbb{C}_0(A \mid \langle H_{w,t}T_w) = C\,h_{w,t}(A \mid H_{w,t}T_w)$$

Assuming that the past has chance 1, this reduces to

$$\mathbb{C}_0(A \mid \langle H_{w,t}T_w \rangle) = C\,h_{w,t}(A \mid T_w)$$

The New Principle If $\mathbb{C}_0$ is a reasonble initial credence function, w is any world, t any time, and A is any proposition in the domain of $C\,h_{w,t}$, then

$$\mathbb{C}_0(A \mid H_{w,t}T_w) = C\,h_{w,t}(A \mid T_w)$$

Review Questions

1. What does Lewis's *Principal Principle* say? Explain why the principal principle is incompatible with Humean views about objective chance, using the Actual Frequentist as an example. What is the *New Principle*, and why is it, unlike the original principal principle, compatible with Actual Frequentism?

2. In my statement of the principal principle, what is '$C\,h_t$', and why is it written in a calligraphic font?

3. What is the complete *theory of chance* for a world, and how can we use it to reformulate Lewis's *principal principle*?

4. Challenge: Lewis's original principal principle contained an *admissibility* clause, but the reformulation does not. Why not?

4
The Philosophy of Statistics

Within the theory of *probability*, we build probabilistic models and use them to calculate the chances of various outcomes or observations. When you're doing probability theory, you are assuming that you know precisely what the underlying probabilistic model is, and you are deducing the consequences of that model.

Example 4 (Spinning a Biased Coin—A Binomial Process)**.** *We have a coin whose bias is* $1/3$*—meaning that, if it is spun,then the chance that it lands on heads is* $1/3$*. The outcome of each spin is probabilistically independent of the outcome of every other spin. We will spin the coin ten times. What is the chance that it lands heads at least* 8 *times*?

The bias of a coin *flip* doesn't vary with the mass distribution of the coin. It pretty much always has a chance of about 51% that it lands on the side that was facing up when the coin was flipped. But coin *spins* can have substantial biases one way or the other. You can test this for yourself: get a few coins of the same denomination and spin them each. You'll probably discover that one of them gets way more heads than the other.

We can model this case with a set $\mathcal{W}$ of sequences of Hs and Ts. A sequence $HHTHHTHHTH$ represents a possibility in which the coin landed heads on the first and second spin, tails on the third spin, heads on the fourth spin, and so on. And we can parametrize the space $\mathcal{W}$ with a (discrete) random variable, X, which counts the number of heads landings.

Then, it will turn out that, for values of x between 0 and 10,

$$\mathbb{C}h(X = x) = \binom{10}{x} \quad (1/3)^x \cdot (2/3)^{10-x}$$

$\binom{10}{x}$ is pronounced '10 choose x', and it counts the number of ways of choosing x things from a set of size 10. In general, n choose k is $n!/k!(n-k)!$.

The chance of each possible value of X is shown in figure 4.2. Adding up the values of $\mathbb{C}h(X = 8)$, $\mathbb{C}h(X = 9)$, and $\mathbb{C}h(X = 10)$, we get that

$$\mathbb{C}h(X \geqslant 8) = 67/19{,}683 = 0.00340395$$

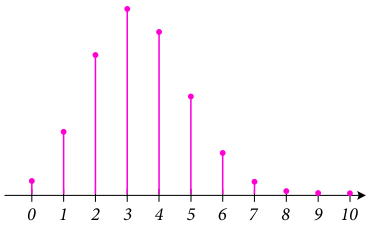


Figure 4.1: A probability mass function for the random variable X. The height of the line above x gives $\mathbb{C}h(X = x)$.

Example 5 (Weight Loss Drug—A Normal Distribution)**.** *Suppose that we divide people into two groups*: *a* control *group and a* test *group. The control group is given a placebo and the test group is given a new weight loss drug that is completely ineffective. Taking the weight loss drug does not affect someone's weight at all. Nonetheless, people's weights fluctuate randomly. If* D *is the difference between the average weight change in the control group and the average weight change in the test group, then* D *has a* standard normal *probability density function. (As in figure 4.2.) What is the chance that the test group lost on average* 1.7 *more pounds than the control group*?

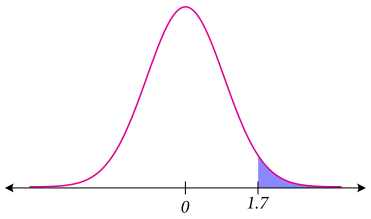
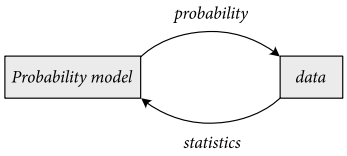


Figure 4.2: A probability density function for the random variable D. The area of the blue region gives $\mathbb{C}h(D \geqslant 1.7)$.

Using a table for the standard normal distribution, you can find that $\mathbb{C}h(D \geqslant 1.7) \approx 0.045$.

In *probability*, you start from some modeling assumptions—like, for instance, that the coin has a 1/3rd bias towards heads or that the difference between weight lost in the test and control group has a standard normal distribution—and you *deduce* consequences about the probabilities of various outcomes.

When you're making a *statistical inference*, you are trying to go the other way around. You are looking at the *outcome* of some experiment and trying to infer something about the probabilities that generated that outcome.



4.1 Fischer's Test of Significance

One standard method of statistical inference is known as a 'significance test'. The simplest version of significance testing was devised by Ronald Fischer.

The basic idea behind the test of significance is this: you formulate a probabilistic hypothesis, H, about the value of some random variable. You then conduct an experiment, gather data, and if, according to the chance hypothesis, the probability that you'd get data like that is sufficiently low, then you *reject* the hypothesis H.[1]

Example 6 (Testing the Bias of a Coin)**.** *We have a coin of unknown bias, b. The outcome of each spin is known to be probabilistically independent of the outcome of every other spin. H is the chance hypothesis that the coin is 1/3rd biased towards heads*: $H : B = 1/3$. *We spin the coin ten times and find that* 8 *of the spins land heads.*

As we calculated above, the probability that we'd see *that* many heads landings (or more) is around 0.3%. That's very improbable. So, if the chance hypothesis H were true, then it's very improbable that we'd have seen as many heads as we in fact saw (or more). So we should reject the hypothesis H.

Example 7 (Testing the Effectiveness of a Weight Loss Drug)**.** *We divide people into two groups*: *a control group and a test group. The control group is given a placebo and the test group is given a new weight loss drug. H is the hypothesis that the weight loss drug is ineffective, and that the variable D* (*the difference between the average weight change in the control group and in the test group*) *has a standard normal distribution. We observe an average difference of* 1.7 *pounds.*

As we calculated above, the probability that we'd see *that much* of a weight difference (or greater) is about 4.5%. That's very improbable. So we should reject the chance hypothesis H, and conclude that the weight loss drug has *some* effect.

4.1.1 Fischerian Significance Testing and Popperian Falsificationism

According to a traditional view in the philosophy of science, what distinguishes science from non-science or pseudo-science is that scientists seek evidence that *supports* and *confirms* their hypotheses; and

[1] How low is 'sufficiently low'? Fisher offhandedly suggested that, if the hypothesis made the evidence less than 5% likely, then the data were 'significant': "It is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not". This '0.05' cut off for statistical significance has been wildly influential within statistics. If the hypothesis made the evidence 5.1% likely, this is considered unpublishable. Researchers have incentives to fudge their data to make it over the 5% line. You can see in the data that there are far more test results just over the 5% significance level than you'd expect, indicating that researchers follow this incentive.

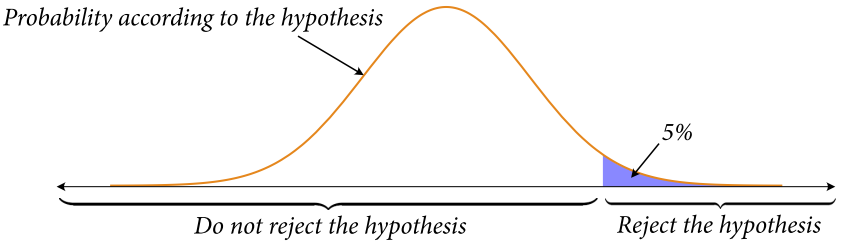


Figure 4.3: Fischer's test of significance (a 'one tailed' test). In orange, the probability density function over potential obervations, according to the chance hypothesis of interest. The values in the blue region have a probability of 5%. If you observe a value in that region, you reject the hypothesis. Else, you do not reject the hypothesis.

the kinds of hypotheses scientists accept are those with supporting and confirmatory evidence; whereas non-science does not have supporting or confirmatory evidence. On this view, scientists start with a hypothesis, derive empirical predictions from it, and then check to see whether the empirical predictions are borne out. If they are, then the hypothesis is *confirmed*. If they are not, then the hypothesis is refuted.

Logic of Confirmation	*Logic of Falsification*
If H, then E	If H, then E
E	not-E
$\therefore H$	$\therefore$ not-H

According to Sir Karl Popper, the distinguishing mark of science is not verification, but rather *falsification*. What makes a theory scientific is that it is falsifiable; non-scientific or pseudo-scientific theories are not falsifiable. Popper thought that his views in the philosophy of science solved Hume's problem of induction, since the logic of theory *confirmation* is inductive; but the logic of theory *rejection* or *falsification* is purely deductive. So Popper's 'solution' to the problem of induction was that science simply does not utilize induction.

One serious philosophical objection to Popper's views were that science very often tests *probabilistic* hypotheses, which cannot be *deductively* falsified. For instance, in example 3, if we choose to reject the probabilistic hypothesis that the coin has a bias of $B = 1/2$ on the basis of the evidence that it landed heads up 8 out of 10 times, this inference is not a deductively valid one. The observed evidence is *consistent* with the coin having a bias of $1/2$.

Philosophers of science have almost unanimously rejected Popper's views—though there are a few hold-outs—and his views about chance hypotheses in particular is seen as a rather weak point in his theory. Nonetheless, Ronald Fischer's significance tests were heavily influenced by Popper's views. However, Fischer didn't share Popper's views about chance hypotheses being unscientific. Instead, Fischer thought that we could reject a hypothesis if the evidence was sufficiently *unlikely*, conditional on that hypothesis (even if the evidence was strictly speaking compatible with the hypothesis). So Fischer agreed with Popper that probabilistic hypotheses could not be *confirmed* by evidence—no amount of evidence would make it reasonable to *accept* a probabilistic hypothesis. But he thought that nonetheless, they could be *falsified* by evidence—there was evidence that would make it reasonable to *re-*

Popper's own view was that chance hypotheses were not scientific. He claimed that

> Probability statements, in so far as they are not falsifiable, are metaphysical and without empirical significance; and in so far as they are used as empirical statements they are used as falsifiable statements.

Here, his idea is that a probabilistic hypothesis like $H : B = 1/2$ is unscientific, precisely because it is compatible with any evidence whatsoever. But a more contentful hypothesis like 'the bias of the coin is $1/2$ *and it will not land heads up more than 80% of the time*' was scientific, because that hypothesis is flatly incompatible with some evidence, and could therefore be falsified by it.

ject the hypothesis.

Fischer's significance tests follow the following logic:[2]

> *Logic of Probabilistic Falsification*
> If H, then it is unlikely that E
> E
> $\therefore$ it unlikely that H

[2] This is the inference Sober calls 'probabilistic modus tollens'

For instance, in example 3: if the bias of the coin were $1/2$, then it is unlikely that the coin would land heads on at least 8 out of 10 spins. The coin did land heads on at least 8 out of 10 spins. So it is unlikely that the bias of the coin is $1/2$, and we should reject this hypothesis.

In example 4: If the weight loss drug were ineffective, then it is unlikely that the test group would lose 1.7 more pounds than the control group on average. The test group did lose 1.7 more pounds than the control group, on average, so it is unlikely that the weight loss drug is ineffective, and we should reject this hypothesis.

4.2 *Bayesian Critiques of Fischer's Test of Significance*

Bayesians think that Fischer's proposed significance test has committed at least three probabilistic fallacies. Let's take them in turn.

4.2.1 *The Principle of Total Evidence*

Notice that, in a significance test, we're not really asking about how likely we would be to get some evidence, if the hypothesis were true. In the case of a continuous random variable, we know that the probability of getting the data we actually got was *zero*—same as any other data we could have received. Instead, we're asking about how likely we'd be to get evidence *at least as far from the mean* as the data we got (or perhaps, at least as far from the mean *in this direction* as the data we got). That's a strange feature of the inference. Why are we not looking at the *strongest* thing we've learned? And if we're going to be looking at something weaker than what we learned, then why not instead ask how likely we'd be to get evidence at least as *close* to the mean as the data we in fact got?

Bayesians are committed to the following principle:

Principle of Total Evidence Any statistical inference should be based on *all* of the evidence you've received.

The Fisherian significance test violates the principle of total evidence by basing its inference on something logically weaker than the total evidence.

4.2.2 *The Base Rate Fallacy*

Here's a probability puzzle that a number of doctors get wrong:

Example 8. *There is a symptom-less disease that* 10% *of the population has. We have a test that is* 90% *reliable at detecting whether or not someone has the*

disease. That is: *the 'false positive' and 'false negative' rates are both* 10%. *Conditional on you having the disease, the probability that the test gives a 'false negative' and says you don't have it is* 10%. *And, conditional on you not having the disease, the probability that the test gives a 'false positive' and says you do have it is* 10%.

You get the test and it comes back positive. What is the probability that you have the disease?

Many people—including may doctors—answer '90%'. But this is incorrect. In fact, the answer is '50%'.

	Neg	*Pos*
Healthy	81	9
Sick	1	9

Figure 4.4: Suppose that there are 100 people, and 10 of them have the disease. Then, 9 of the 10 sick people will get a true positive, and 1 of the sick people will get a false negative. And 81 of the 90 healthy people will get a true negative, whereas 9 of the healthy people will get a false positive. When you learn that you got a positive test result, you learn that you are in the right-hand column.

The fallacy that people make when they say that you are 90% likely to have the disease is called the 'base rate fallacy'—so-called because you are ignoring the base rate of the disease in the population at large. If the disease is antecedently unlikely (as in this case), then even if the test is particularly reliable, a positive result needn't raise the probability that you have the disease above 50%.

In general, we need to be careful to distinguish the probability of a *hypothesis* (like, e.g., that you are sick) given the *evidence* (like, e.g., a positive test result) from the probability of the *evidence* given the *hypothesis*.

$$\underbrace{\mathbb{P}(H \mid E)}_{\textit{posterior}} \qquad \underbrace{\mathbb{P}(E \mid H)}_{\textit{likelihood}}$$

Recall that the quantity on the left is called the 'posterior' probability of the hypothesis, given the evidence. The quantity on the right is called the 'likelihood'—it is how likely the hypothesis makes the evidence. According to Bayesians, the quantity on the left is what we should be concerned with when we're asking whether or not to accept a hypothesis. According to the Bayesian, you can measure the degree to which a piece of evidence confirms a hypothesis by comparing your posterior credence $\mathbb{C}(H \mid E)$ to the prior credence $\mathbb{C}(H)$.

Bayesian Theory of Confirmation If your posterior credence is greater than your prior, $\mathbb{C}(H \mid E) > \mathbb{C}(H)$, then E has confirmed H. If the posterior is less than the prior, $\mathbb{C}(H \mid E) < \mathbb{C}(H)$, then E has disconfirmed H. If the posterior is equal to the prior, $\mathbb{C}(H \mid E) = \mathbb{C}(H)$, then E has neither confirmed nor disconfirmed H.

The Bayesian thinks that likelihoods have an important role to play in inductive inference. But they do not think that they are the entire story. The likelihoods relate the prior to the posterior *via Bayes' Theorem*:

$$\underbrace{\mathbb{P}(H \mid E)}_{\textit{posterior}} = \frac{\mathbb{P}(E \mid H)}{\mathbb{P}(E)} \cdot \underbrace{\mathbb{P}(H)}_{\textit{prior}}$$

Bayes' theorem is just that—a theorem. But together with the Bayesian's theory of confirmation (which is *not* a theorem, but a substantive philosophical assumption), and the principal principle, the theorem tells us something about the nature of confirmation. Roughly, what it says is that theories that make better predictions get confirmed; and theories that make worse predictions get disconfirmed. Think about it like this:

the ratio $\mathbb{P}(E \mid H) \div \mathbb{P}(E)$ gives a measure of how well the hypothesis H did predicting the evidence. If the ratio is positive, then H made E more likely than it was antecedently. In that case, $\mathbb{P}(H \mid E)$ will be greater than $\mathbb{P}(H)$, and E will *confirm* H.

The 'base rate fallacy' is effectively the fallacy of ignoring the prior probability of the hypothesis. In example 5, it is ignoring the prior probability that you have the disease. But it seems that the same fallacy is taking place in Fischerian significance testing.

Defenders of significance testing often respond to this allegation by appealing to a frequentist interpretation of probability—for this reason, these statistical techniques are often known as 'frequentist' statistics. Their response is that the case of significance testing is importantly different from the medical test in example 5. For, while there is a non-trivial frequency for the disease in the population, there is no non-trivial frequency for a probabilistic hypothesis. Either the hypothesis is true, in which case its probability is 1, or the hypothesis is false, in which case its probability is 0. And these probabilities won't change when you condition on E.

Frequentists draw a sharp distinction between the probability of the evidence *according to* the hypothesis and the probability of the evidence *conditional on* the hypothesis. To emphasize this difference, they use different notation for the latter.

$\mathbb{P}(E; H)$ = the probability of E *according to* H

$\mathbb{P}(E \mid H)$ = the probability of E *conditional on* H

At base, the disagreement has to do with what kinds of probabilities there are, and what role they have to play in the practice of science. Both Bayesians and Frequentists will draw a distinction between the *chance* of E according to H and the *chance* of E conditional on H. And both Bayesians and Frequentists can agree that the *chance* of the true chance hypothesis is 1 and the *chance* of the false chance hypothesis is 0. What they disagree about is whether there is another kind of probability out there where the chance of E according to H is equal to the chance of E conditional on H. Bayesians think that there is: a reasonable *credence* function is a probability like this.

$\mathbb{C}h_H(E)$ = the chance of E *according to* H

$\mathbb{C}(E \mid H)$ = a reasonable credence in E *conditional on* H

Because they accept something like Lewis's principal principle, Bayesians think that—at least in the absence of inadmissible evidence—these two quantities should align. Your credence in E, conditional on H, should be the probability of E *according to* H. That is,

$$\mathbb{C}(E \mid H) = \mathbb{C}h_H(E)$$

For instance, in example 3, Bayesians think that your credence in $X \geqslant 8$, conditional on $H : b = 1/3$, should be around 0.3%. And, in example 4, they think that your credence that $D \geqslant 1.7$, conditional on the null hypothesis that the drug is ineffective, should be around 4.5%.

4.2.3 *Neyman-Pearson Significance Test*

Jerzy Neyman and Egon Pearson developed a slightly different version of Fischer's test of significance. Rather than focusing on a *single* probabilistic hypothesis, Neyman and Pearson were interested in

making inferences in cases where there was some *collection* of hypotheses you are deciding between. In the simplest case, there could be just two hypotheses. These two hypotheses are called the 'null' hypothesis and the 'alternative' hypothesis, and written 'H_0' and 'H_1'

Neyman and Pearson thought that statistical inference was a matter of managing two types of *error*: the error of rejecting the null when the null is actually true (a 'type I error'), and the error of failing to reject the null when the null is actually false (a 'type II error'). The probability of a type I error is standardly called 'α' and the probability of a type II error is standardly called 'β'. Neyman and Pearson advised you to formulate a rule for deciding between the two hypotheses that appropriately balanced these two probabilities.

	H_0	H_1
accept H_0	no error	type II error
accept H_1	type I error	no error

$\mathbb{P}(\text{type I error}) = \alpha$

$\mathbb{P}(\text{type II error}) = \beta$

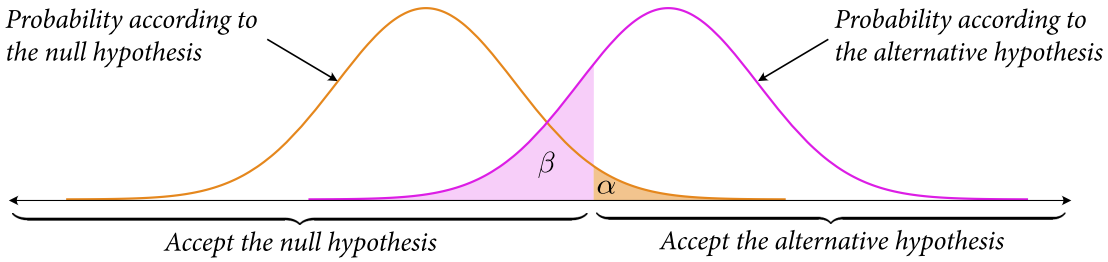


Figure 4.5: Neyman-Pearson significance testing. The axis is the values of the random variable whose value we are observing in our experiment. The orange curve is the probability density function over values of the random variable according to the null hypothesis. and the purple curve is the probability density function over values of this random variable according to the alternative hypothesis. α is the probability of a type I error (accepting the alternative when the null is true), and β is the probability of a type II error (accepting the null when the alternative is true). $1 - \beta$ is called the 'power' of the test.

The *power* of a Neyman-Pearson significance test is defined to be $1 - \beta$. That is: the power of the test is the probability the alternative hypothesis gives to you accepting the alternative hypothesis. The higher the power, the better the test's ability to detect the truth of the alternative hypothesis. So Neyman and Pearson advise you to select the test with the *greatest possible* power. Notice that this allows them to answer an objection raised to Fischer: why select a region with probability α *in the tail* of the null distribution? Why not instead select a region with probability α in the middle of the null distribution? Neyman and Pearson justify their choice of rejection region on the grounds that it has the greatest possible power (amongst those with a certain fixed probability of a type I error). In other words: Neyman and Pearson advise you to first select an α, and thereupon choose a test which minimizes β. This will generally lead you to select a rejection region in the tails of the null distribution.

4.2.4 *Lindley's Paradox*

Example 9. *We have a coin of unknown bias. We do know, however, that it is either fair* ($B = 1/2$)*, or else it is biased* 4/5*ths towards heads* ($B = 4/5$)*. We spin the coin* 100 *times and observe that it lands heads* 60 *times.*

We have two hypotheses: the 'null' hypothesis that the coin is unbiased, $H_0 : B = 1/2$, and the 'alternative' hypothesis that the coin is 4/5ths biased towards heads, $H_1 : B = 4/5$. Let X be the random variable which counts the number of heads landings. Then, figure 4.6 gives the probability distributions for the values of X, given the null hypothesis (in blue) and given the alternative hypothesis (in orange).

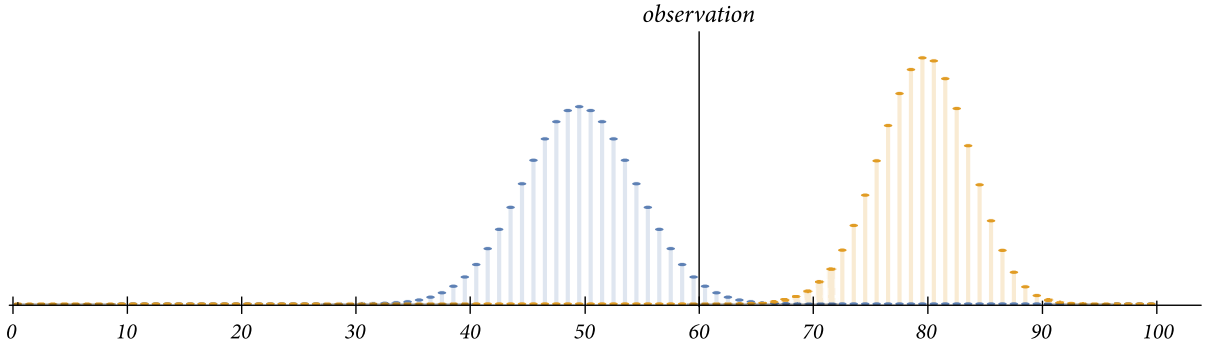


Figure 4.6: A probability mass function for the random variable 'number of heads landings', both according the null hypothesis $H_0 : B = 1/2$ (in blue), and according the alternative hypothesis $H_1 : B = 4/5$ (in orange).

How likely is it that we would have observed at least 60 heads, if the coin were unbiased? That is: how likely is our evidence (or, rather, a logical weakening of it), $X \geqslant 60$, on the assumption that the null hypothesis is true? It's about 2.8% likely. Since 2.8% is less than 5%, the null hypothesis has be refuted by the significance test. Following Fisher's logic, the hypothesis H_0 should be rejected, and we should conclude that the coin has a bias of $4/5$.

Similarly, if we select an α of 5%, Neyman and Pearson will advise us to reject H_0 if we see 59 or more of the spins land heads. So, on a Neyman-Pearson significance test, too, we will accept that the coin has a bias of $4/5$. Notice that the *power* of this test is extremely high—it is nearly 100%.

But given the Bayesian's principle of total evidence and the Bayesian theory of confirmation, the null hypothesis is *very well* confirmed by this evidence. Suppose that we started out 50% confident that the coin was fair and 50% confident that the coin was biased. Then, the probability that the coin is fair (H_0), conditional on our total evidence (E), will be well over 99.9%.[3]

$$\begin{aligned}\mathbb{P}(H_0 \mid E) &= \frac{\mathbb{P}(E \mid H_0) \cdot \mathbb{P}(H_0)}{\mathbb{P}(E \mid H_0) \cdot \mathbb{P}(H_0) + \mathbb{P}(E \mid H_1) \cdot \mathbb{P}(H_1)} \\ &= \frac{\left[(1/2)^{60} \cdot (1/2)^{40}\right] \cdot 1/2}{\left[(1/2)^{60} \cdot (1/2)^{40}\right] \cdot 1/2 \;+\; \left[(4/5)^{60} \cdot (1/5)^{40}\right] \cdot 1/2} \\ &\approx 0.999786\end{aligned}$$

Intuitively, what's going on here is that, even though the evidence was *very* unlikely according to the null hypothesis, it was *even less* likely according to the alternative hypothesis. So while H_0 didn't do a very good job predicting this evidence, it did a *much better* job predicting the evidence than H_1 did.

$$\mathbb{P}(X \geqslant 60; H_0) = \sum_{x=60}^{100} \binom{100}{x} (1/2)^x \cdot (1/2)^{100-x} \approx 0.028$$

$$\mathbb{P}(X \geqslant 59; H_1) = \sum_{x=60}^{100} \binom{100}{x} (4/5)^x \cdot (1/5)^{100-x} \approx 0.999999555$$

[3] In this example, our total evidence will include the precise order in which the coin landed heads and tails; so our evidence E includes more than just the information that $X = 60$. Nonetheless, this additional information won't make any difference to the posterior probability of the hypotheses, so Bayesians could decide to simply condition on $X = 60$, even though this isn't strictly speaking the *total* evidence. Random variables like this are called 'sufficient statistics', because, once you know the value of the random variable, that's enough—you don't need to know anything else about the data, since that additional information won't affect the posterior probabilities.

4.3 *Bayesian Statistics*

The Bayesian takes a very different approach to statistics. In a sense, we have already seen the core Bayesian idea: it is encoded in the two Bayesian norms of the *principal principle* and *conditionalization*. You take a prior which satisfies the principal principle, and condition it on

your *total* evidence. If a hypothesis's probability goes up, then that hypothesis has been confirmed. If the hypothesis's probability goes down, then that hypothesis has been disconfirmed.

The Bayesian wants to carefully distinguish two different kinds of probability functions: the *objective chance* function which is generating the data, $\mathcal{C}h$, and your *subjective credence* function, $\mathbb{C}$. Your subjective credence is just a fixed probability function, but the objective chance function is unknown—it is something about which you have credences.

As a reminder, I am using the calligraphic notation '$\mathcal{C}h$' for definite descriptions of probability function, and I am using blackboard boldface '$\mathbb{P}$' or '$\mathbb{C}$' for a particular (rigidly designated) probability function.

Even though the Bayesian statistician will treat the objective chance function as unknown, they will often assume that it comes from a particular *family* of probability distributions.

4.3.1 *Families of Probability Distributions*

Think back to examples 3 and 4. In those cases, we knew *something* about the probability distribution over the variables X ("how many times the coin lands heads") and D ("the difference in average weight change between the test and the control group"). In the case of the coin, we knew that the outcomes of successive spins were independent and that each time the coin was spun, it had the same probability of landing heads. This meant that we knew the random variable X had a probability distribution in the so-called 'binomial' family.

Binomial Distribution A (discrete) random variable X, which takes on integer values between 0 and n, has a *binomial* distribution iff there's some number $B \in [0, 1]$ such that

$$\mathcal{C}h(X = x) = \binom{n}{x} \cdot B^x \cdot (1 - B)^{n-x}$$

If we know that X has a binomial distribution, then we can completely characterize a chance hypothesis in terms of a single number: the number B. (For instance, in example 3: we can characterize the chance hypothesis in terms of the bias of the coin.)

In general, if you know that a random variable has a probability distribution from a given family, then you will be able to completely specify the probability distribution over that random variable with a small number of *parameter* values. In the case of the binomial distribution, there is just one parameter value. For another common probability family,

Normal Distribution A (continuous) random variable X, which can take on any real number as its value, has a *normal* distribution iff there's some pair of real number (μ, σ), such that the probability density function for X, f_X, is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

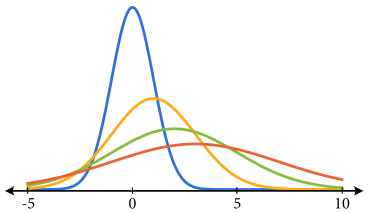


Figure 4.7: In blue, the normal distribution with mean 0 and standard deviation 1, $\mathcal{N}(0, 1)$. In orange, $\mathcal{N}(1, 2)$. In green, $\mathcal{N}(2, 3)$. In red, $\mathcal{N}(3, 4)$.

μ is the *mean* of the distribution, and σ is the *standard deviation* of the distribution. A normal distribution with mean μ and standard deviation σ is written '$\mathcal{N}(\mu, \sigma)$'. Some sample normal distributions are shown in figure 4.7.

Beta Distribution A (continuous) random variable X, which can take on any number between 0 and 1, has a *beta* distribution iff there's some pair of real number (α, β), such that the probability density function for X, f_X, is given by

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \cdot x^{\alpha - 1} \cdot (1 - x)^{\beta - 1}$$

Just as the normal distribution is characterized by two parameters—the mean and the standard deviation—the beta distribution is characterized by two parameters, α and β. (Note: these 'α's and 'β's are *different* from the 'α's and 'β's from Neyman-Pearson significance testing.) Also, note that the $\mathrm{Beta}(1, 1)$ distribution is just the *uniform* distribution over the unit interval $[0, 1]$. (See figure 4.9.)

There are many other families of probability distributions, but for our purposes, we only need to familiarize ourselves with these three.

4.3.2 *Conjugate Priors and Laplace's Rule of Succession*

Suppose you know that the objective *chance* function is inside the binomial family. For instance, suppose you know that the coin has a certain bias, B, and that the outcome of different spins are independent of each other.

As the Bayesian approaches statistical inference, inference is all a matter of taking a prior credence distribution over the values of the parameter B (the bias of the coin), and then updating that prior credence distribution on the observed data. In this case, we can think about B as a random variable—it takes on different values in different epistemically possible worlds. This random variable takes on values in between 0 and 1 (inclusive). And we can suppose that your subjective credences are distributed over all propositions of the form $B \geqslant b$.

We assume that the objective chance distribution falls in some family or other. As we've seen, this objective chance distribution can be characterized by some number of (epistemic) random variables which parameterize the chance distributions in that family. For mathematical convenience, Bayesian statisticians look for a prior distribution over the random variable B with the following property: whichever probability family the prior starts out in, the posterior will *remain* in that same family. The prior credence distribution with this property is said to be *conjugate* to the objective chance distribution. For instance, the beta distribution is conjugate to the binomial distribution. Moreover, it's conjugate in a particularly lovely way.

It's not important, but if you're curious, '$\Gamma(x)$' is the gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$

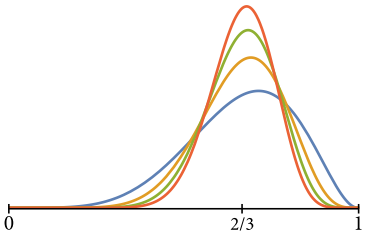


Figure 4.8: In blue, the beta distribution with parameter values $\alpha = 6$ and $\beta = 3$, $\mathrm{Beta}(6, 3)$. In orange, $\mathrm{Beta}(10, 5)$. In green, $\mathrm{Beta}(14, 7)$. In red, $\mathrm{Beta}(18, 9)$.

If you know that the outcome of successive coin spins are independent with constant bias B, and your prior subjective credence distribution over B is the uniform distribution—that is, the $\text{Beta}(1,1)$ distribution—then your posterior distribution after observing h heads and t tails will be the $\text{Beta}(h+1, t+1)$ distribution.

There's another fact worth dwelling on for a minute. The *expected value* of any random variable with a $\text{Beta}(\alpha, \beta)$ distribution is $\alpha/(\alpha+\beta)$. So suppose you begin with the uniform distribution over potential biases of the coin, and you spin it a number of times n and observe h heads and $n-h$ tails. Then, you will end up with the $\text{Beta}(h+1, (n-h)+1)$ posterior distribution over B (the bias of the coin), and the expected value of B will be

$$\mathbb{E}[B] = \frac{h+1}{h+1+(n-h)+1} = \frac{h+1}{n+2}$$

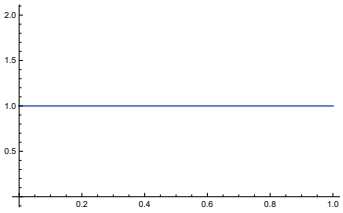
And since your probability that the next coin lands heads will equal to your expectation of the coin's bias, we have that your credence that the coin will land heads on the $n+1$st spin will be equal to the number of heads you've observed in the first n spins plus 1, divided by $n+2$—which is just Laplace's *rule of succession*!

While the Beta (1, 1) distribution is a *uniform* distribution, not all conjugate priors will be uniform in this way. For instance, suppose (as in example 4) you know that a random variable, D, has a normal objective chance distribution, with an unknown mean but a known standard deviation of 1. That is, the objective chance distribution over D is given by $\mathcal{N}(\mu, 1)$, for some unknown value μ. Then, the conjugate subjective prior (over potential values of μ) will be a normal distribution. Suppose (just for illustration) that you start off with a standard normal ($\mathcal{N}(0,1)$) prior distribution over the values of D. And then you observe an average difference of d between the weight changes in the test and the control group. Then, your posterior distribution over values of m will change to $\mathcal{N}(d/2, 1/2)$. (See figure 4.10).

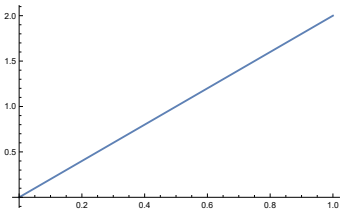
4.3.3 *Bayesian Inference*

The output of a Bayesian inference will just be a probability distribution like the ones shown in figure 4.9. This is a *subjective* probability distribution over the parameter values characterizing the *objective* chance function. For instance, in our running coin spinning example, the output of the Bayesian inference will be a probability distribution over the unit interval. The interpretation is that, if you started out with the given prior, then you should end up with the posterior the Bayesian provides.

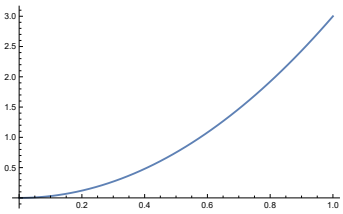
A general recipe is to start with some assumptions about the family of the objective chance distribution, then find a subjective probability distribution over the parameter values characterizing that objective chance distribution which is conjugate to the objective chance distribution's family, and then update the prior on the observed data by



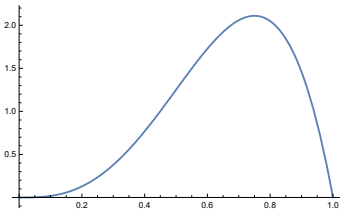
(a) before any outcomes; $\text{Beta}(1,1)$



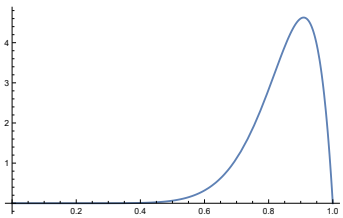
(b) after 1 head; $\text{Beta}(2,1)$



(c) after 2 heads; $\text{Beta}(3,1)$



(d) after 3 heads and 1 tail; $\text{Beta}(4,2)$



(e) after 10 heads and 1 tail; $\text{Beta}(11,2)$

Figure 4.9: Evolution of your credence distribution over B as you learn more and more outcomes.

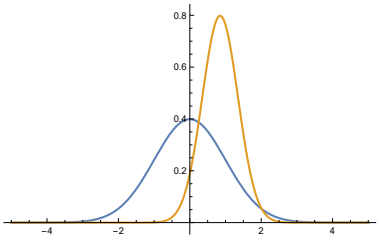


Figure 4.10: In blue, the prior (standard normal) distribution over potential values of D's mean. In orange, the posterior $\mathcal{N}(1.7/2, 1/2)$ distribution that you get from conditioning the prior on the observation of an average weight difference of 1.7.

conditionalization. At the end of this process, you get a posterior credence distribution over objective parameters.

You can use these probability distributions to construct what's known as a 'credible interval' (which is very different from the 'confidence intervals' used by frequentists). A 99% *credible interval* is an interval which is 99% likely to contain the true parameter value (given whatever prior distribution).

The way to understand a '99% confidence interval' is to understand the method by which it was constructed. It is constructed according to a method such that, if you were to construct intervals according to that method over and over again, in the long run, 99% of those intervals would contain the true value. This doesn't mean that *in this particular case*, it is 99% likely that the interval contains the true parameter value.

4.4 The Problem of the Priors

One of the most central criticisms of Bayesian methods is their reliance on prior probability distributions. According to Frequentists, we can subject a chance hypothesis to objective test without having to form any prior opinions about the hypothesis. But for Bayesians, we must begin with a subjective prior probability distribution over hypotheses. This reliance on a prior gives rise to two related objections to Bayesian inference.

4.4.1 Subjectivism

One aspect of the problem of the priors is that the priors seem to be irredeemably *subjective*. The methods of science call for objective, or at least, *intersubjective* methods for evaluating hypotheses. Science is, after all, a collaborative undertaking. But Bayesians cannot give us this, for their priors merely represent subjective degrees of belief, which could potentially vary from scientist to scientist. Deborah Mayo puts the point like this:

> In science, it seems, we want to know what the data are saying, quite apart from the opinions we start out with.[4]

[4] Mayo, *Error and the Growth of Scientific Knowledge*, page 76

At this point, it's worth pointing out that there are a variety of different forms of Bayesianism out there, and that some are more subjective than others. In fact, there's a continuum of different kinds of Bayesianism, depending upon how stringent they take the requirements of rationality to be.[5] On one side are what I'll call the 'radical subjectivist' Bayesians:

[5] I. J. Good famously joked that there are more versions of Bayesianism than there are Bayesians.

Radical Subjectivism In the absence of evidence, *any* probability function is a reasonable credence function.

The radical subjectivist thinks that there are but two norms of epistemic rationality: probabilism and conditionalization. So long as you abide by these two norms, you will be epistemically rational.

A slightly less radical version of subjectivism additionally accepts a principle like Lewis's *Principal Principle* (or, perhaps, the New Principle)—but *that's it*.

Subjectivism In the absence of evidence, any probability function which satisfies the principal principle is a reasonable credence function.

Subjectivism is slightly more constraining than radical subjectivism, but it is still a *very* liberal epistemology. According to the subjectivist,

you could assign arbitrarily high prior credence to the hypothesis that all emeralds are green, or to the hypothesis that there is a giant tea kettle on the other side of Jupiter.

More objective are those who think that there are rational requirements on credences, but who think that nonetheless, there is a range of permissible opinion.

Permissive Objectivism Not just any probability function satisfying the principal principle is a reasonable prior credence. But the requirements of rationality are not so demanding that they pin down exactly one reasonable prior.

Permissively Objective Bayesians think that reasonable people can disagree, but unlike the subjectivists, they don't think that any disagreement is reasonable (or rather, they don't think that any disagreement between people who satisfy the principal principle is reasonable).

On the other extreme, there are so-called 'Objective Bayesians':

Impermissive Objectivism In the absence of evidence, there is exactly one reasonable prior probability function.

Impermissive objectivists think that there is a 'one true prior', deviation from which is irrational. They often also endorse the stronger thesis that it is determinate what the one true prior is—they will formally characterize it. Given the way Bayesian statistics goes in practice, you might expect the one true prior to be a conjugate distribution. But philosophers tend to regard the use of conjugate priors as a mere arithmetic convenience with little epistemic significance. Most impermissive objectivists endorse the *principle of indifference* which we encountered earlier in the course—or a version of it known as the 'principle of maximum entropy'. Many Bayesians have therefore taken the Bertrand paradoxes as a reason to reject impermissive objectivism. So at least many traditional Bayesians have been fairly subjected to the charge of subjectivism.

But Bayesians have a response to the charge—they appeal to various 'convergence' or 'washing out' theorems.[6] To appreciate these theorems and their relevance to the charge of subjectivity, suppose that we have two subjectivist Bayesians, Alice and Bob. Alice and Bob both have credence functions that satisfy the axioms of probability (in particular, they have countably additive credence functions). Alice and Bob both assign probability zero to all the same propositions about the bias of the coin. And both Alice and Bob update their credences by conditionalization. Alice and Bob are going to spin a coin repeatedly and learn how it landed. Both Alice and Bob (and the coin) are immortal, and they will do this forever. Then, the convergence theorem says that, in the limit as the number of spins go to infinity, both Alice and Bob will (with probability 1) converge in their opinions—*moreover*, in the limit as the number of spins goes to infinity, they will both (with probability 1) invest credence 1 in the *true* bias of the coin.

The upshot: Bayesians may insist that, while the priors may be infected with subjectivity, this subjectivity has less and less impact on

[6] If you want to see the technical details, you should look for Doob's *martingale convergence theorem*. See also the exposition on page 144 of Earman's *Bayes or Bust?*.

the posterior the more and more evidence you receive.

It's not clear how persuasive this response is. As Keynes pointed out, in the long run, we're all dead. All scientific practice takes place in the short run, with finite bodies of evidence. And in the short run, Alice and Bob's opinions can be as disparate as we like. Suppose, for instance, that Alice and Bob have observed 199 spins of the coin land heads and only 1 land tails. Alice's prior over biases of the coin was the uniform $\mathrm{Beta}(1, 1)$, whereas Bob's was the massively *non*-uniform $\mathrm{Beta}(1, 1000)$. After updating on the evidence, Alice is nearly certain that the coin heavily biased heads, but Bob is nearly certain that the coin is heavily biased tails. (See figure 4.11.)

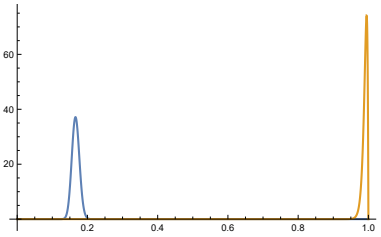


Figure 4.11: In blue, $\mathrm{Beta}(200, 1001)$. In orange, $\mathrm{Beta}(200, 2)$.

This point generalizes: for *any finite* body of data, E, and any prior credence function $\mathbb{C}$ such that the posterior $\mathbb{C}(H \mid E) \geqslant 0.99$, we can find *another* prior credence function $\mathbb{C}'$ such that $\mathbb{C}'(H \mid E) \leqslant 0.01$. So it looks like we're stuck with the subjectivity of priors in the short run. Of course, this is less concerning if we think that it would be irrational to have something like $\mathrm{Beta}(1, 1000)$ as your prior. So the convergence theorems give more comfort to the permissive objectivist Bayesians.

4.4.2 *The Catchall and the Problem of New Hypotheses*

So there seems to be subjectivity in the prior probability of a *particular* hypothesis. But Bayesian inference about a particular hypothesis doesn't just rely upon your prior probability for *that one* hypothesis. It relies upon a prior probability distribution over a *partition* of hypotheses, $\{H_0, H_1, \ldots, H_N\}$. To appreciate this, consider Bayes' theorem:

$$\mathbb{C}(H_0 \mid E) = \frac{\mathbb{C}(E \mid H_0) \cdot \mathbb{C}(H_0)}{\mathbb{C}(E)}$$

How do we calculate the prior probability of the *evidence*? Well, given a partition of hypotheses $\{H_0, H_1, \ldots, H_N\}$ (which will include the hypothesis of interest H_0), the law of total probability allows us to decompose $\mathbb{C}(E)$ into a weighted sum of likelihoods, $\mathbb{C}(E) = \sum_{i=0}^{N} \mathbb{C}(E \mid H_i) \cdot \mathbb{C}(H_i)$.

$$\mathbb{C}(H_0 \mid E) = \frac{\mathbb{C}(E \mid H_0) \cdot \mathbb{C}(H_0)}{\sum_{i=0}^{N} \mathbb{C}(E \mid H_i) \cdot \mathbb{C}(H_i)}$$

Assume that each hypothesis determines a unique probability for E, and assume that you have no inadmissible information. Then, the principal principle (and conditionalization) will tell us that the likelihood must equal this chance, $\mathbb{C}(E \mid H) = Ch_H(E)$. So we will have

$$\mathbb{C}(H_0 \mid E) = \frac{Ch_{H_0}(E) \cdot \mathbb{C}(H_0)}{\sum_{i=1}^{N} Ch_{H_i}(E) \cdot \mathbb{C}(H_i)}$$

All of the quantities $Ch_{H_i}(E)$ are intersubjective. So the remaining subjectivity in Bayesianism can be reduced down to the subjectivity of the prior probability distribution over the partition of possible hypotheses, $\mathbb{C}(H_0), \mathbb{C}(H_1), \ldots, \mathbb{C}(H_N)$.

Now, when we're conducting a scientific test or gathering evidence, we don't always know all of the possible hypotheses. For one instance:

when we're spinning the coin, it could be the the random variable *number of times the coin lands heads* has a binomial distribution; but it could have some other distribution altogether. Maybe the outcomes are not independent. Maybe the bias of the coin changes over time. We have some idea how things go if we're *certain* that the coin spins follow a binomial distribution—but what if they don't? For another: the perihelion of Mercury's orbit (the point in the orbit where it is closest to the sun) moves. This *precession* of Mercury's perihelion was difficult to account for with Newton's theory of gravitation. And it was much easier to account for with Einstein's theory of general relativity. But when the precession of Mercury's perihelion was discovered, nobody knew about Einstein's theory—they didn't even have the necessary concepts to formulate the theory, since non-Euclidean geometry had not even been discovered yet. But since we need the possible hypotheses to be a *partition*, we can't just leave Einstein's theory out.

The Bayesian will say that, if you have the N well-defined hypotheses H_0, H_1, ... H_N, all of which are incompatible but which don't cover all of the possibilities, you should also have a 'none of the above' hypothesis, $\neg(H_0 \vee H_1 \vee \cdots \vee H_N)$ that you carry around with you. This 'none-of-the-above' hypothesis is often called the 'catchall' hypothesis.

The catchall hypothesis raises two different problems. First problem: it is unclear what the likelihoods should be for the catchall hypothesis. Suppose that random variable *how many times the coin lands heads* doesn't have a binomial distribution. Then, how likely is it that you'd see 8 out of 10 spins land heads? Or suppose that none of the current theories of gravity are true. Then, how likely is it that Mercury's perihelion would precess? It looks like any answer to these questions is just going to be arbitrary and incredibly subjective.

Second problem: what do you do when a new hypothesis arrives on the scene? At this point, the algebra over which your credences are defined will have to be *expanded*. For instance, when scientists acquired the concepts necessary to entertain Einstein's theory of general relativity, they suddenly had a new hypothesis to have credences about. This process is called 'awareness growth'. While Bayesians have firm views about how your credences should change when you gain evidence (conditionalization), they have less firm views about how your credences should change when your awareness grows. There are proposals, but there's nothing like an orthodox view.

4.4.3 *The Problem of Old Evidence*

Clark Glymour raised a third problem for the Bayesian's theory of confirmation. When Einstein's theory of relativity was being formulated, the precession of the perihelion of Mercury was *already* well-known. It was *old* evidence. Assuming that scientists are good Bayesian agents, they have already *conditioned* on this evidence. So their credence in the precession of the perihelion of Mercury was already 100%. So their credence in any hypothesis, conditional on this evidence, must be

See Glymour's *Why I Am Not a Bayesian*. One response to Glymour comes from Daniel Garber's "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory". Garber contends that what scientists learned that confirmed the theory of relativity, T, *wasn't* the old evidence, E. Instead, he suggests that it is the new (purely a priori) information that the theory T *implies* E (together with our background knowledge), $T \vdash E$. Now, standard Bayesianism requires that you assign probability 1 to any *a priori* truth. But Garber shows how to relax this assumption, and allow that $\mathbb{C}(T \vdash E) < 1$. He then shows that, even if you have $\mathbb{C}(E) = 1$, you can still have $\mathbb{C}(T \mid T \vdash E) > \mathbb{C}(T)$. So the purely *a priori* information that Einstein's theory implied the precession of the perihelion of Mercury could be taken to confirm Einsteins' theory, even though the precession of the perihelion was old news.

equal to their unconditional credence in that same hypothesis. That's because

$$\text{if } \mathbb{C}(E) = 1, \quad \text{then} \quad \mathbb{C}(H \mid E) = \mathbb{C}(H)$$

So, given the Bayesian's theory of confirmation, it follows that the precession of the perihelion of Mercury *did not confirm* Einstein's general theory of relativity.

4.5 Likelihoodism

Bayesians and Frequentists aren't the only views in the philosophy of statistics. There's another character known as the 'likelihoodist'. Likelihoodism doesn't say anything about which hypotheses we should accept or reject. Instead, it only talks about which hypotheses the evidence favors over which other hypotheses.

Likelihoodism (Qualitative) The evidence E favors hypothesis H_1 over hypothesis H_2 iff the likelihood of E given H_1 is greater than the likelihood of E given H_2, $\mathbb{P}(E \mid H_1) > \mathbb{P}(E \mid H_2)$.

Likelihoodism (Quantitative) The degree to which E favors H_1 over H_2 is given by the likelihood ratio $\mathbb{P}(E \mid H_1) \div \mathbb{P}(E \mid H_2)$

Assuming your credences satisfy the principal principle, there will be no difference between the objective chance assigned by E by the hypothesis H and your credence in E conditional on H. So we could replace '$\mathbb{P}(E \mid H)$' with '$\mathbb{C}(E \mid H)$' or with '$\mathbb{P}(E; H)$', or whatever.

Note two things about likelihoodism: first, it is not a theory about what we should accept. It is only a theory about what the evidence *favors*, or *supports*. Secondly, it is not a theory about absolute favoring; it is instead only a theory about comparative favoring. The thesis doesn't tell you anything about whether some piece of evidence supports a hypothesis *full stop*. Instead, it only tells you something about whether the evidence supports one hypothesis *over* another.

On Sober's telling, likelihoodism is strictly weaker than Bayesianism. For Bayesianism, together with the following assumption, implies the qualitative version of likelihoodism.

Favoring E favors H_1 over H_2 iff conditioning on E raises the ratio $\mathbb{P}(H_1) \div \mathbb{P}(H_2)$—that is, iff

$$\frac{\mathbb{P}(H_1 \mid E)}{\mathbb{P}(H_2 \mid E)} > \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}$$

To appreciate that Bayesianism and Favoring together imply Likelihoodism, consider the following consequence of Bayes' theorem:

$$\frac{\mathbb{P}(H_1 \mid E)}{\mathbb{P}(H_2 \mid E)} = \frac{\mathbb{P}(E \mid H_1)}{\mathbb{P}(E \mid H_2)} \cdot \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}$$

Therefore, if conditioning on E raises the ratio $\mathbb{P}(H_1) \div \mathbb{P}(H_2)$, then the fraction $\mathbb{P}(E \mid H_1) \div \mathbb{P}(E \mid H_2)$ must be greater than one. So it

must be that the likelihood of E, conditional on H_1, is greater than the likelihood of E, conditional on H_2.

The sense of 'favoring' used in Likelihoodism is different from its use in natural language. For instance, Sober gives the following example: you hear loud noises in the attic. This evidence is *certain* conditional on the hypothesis that there are gremlins bowling in the attic. But it is only somewhat likely on the hypothesis that there is a possum in the attic. But it sounds strange, to say the least, that this evidence favors the hypothesis that there are gremlins bowling in the attic over the hypothesis that there is a possum in the attic.

Sober suggests that we understand 'favoring' as a term of art. He doesn't say very much to help us glom onto this term of art, but if we are Bayesians, then we may understand him as simply *defining* the word 'favoring' via the biconditional above: E favors H_1 over H_2 iff conditioning on E raises the ratio $\mathbb{P}(H_1) \div \mathbb{P}(H_2)$.

Likelihoodism avoids the subjectivity of Bayesianism by doing away with the priors. It also avoids concerns about the catchall by not requiring a partition of possible hypotheses—it only requires two well-defined hypotheses to compare. (Does it avoid the problem of old evidence?)

However, likelihoodism also tells us much less than either Frequentism or Bayesianism. Frequentism tells us when we can accept a hypothesis; and Bayesianism tells us how confident to be in various hypotheses (given a prior). But the likelihoodist doesn't do either of these things. They only say something about which hypotheses the evidence favors over which others. Richard Royall distinguished three different questions:

1. What does the evidence say?
2. What should you believe?
3. What should you do?

Bayesians and Frequentists both attempt to answer questions (2) and (3). But the likelihoodist is only attempting to answer question (1)—at least, they are attempting to say something relevant to the question. As we've seen, Bayesians can accept the likelihoodist's answer to question (1), given their stipulative use of the term 'favor'. So it can seem that the disagreement between Bayesians and Likelihoodists isn't so much epistemic as practical—they are disagreeing about how science is to be conducted, and what statisticians should concern themselves with. The Bayesian thinks that they should be concerned with the probability of various hypotheses; whereas the likelihoodist thinks that they should only be concerned with which hypotheses the evidence favors over which others.

Bayesians typically approach (3) through the lens of expected utility theory; Neyman and Pearson addressed it through their significance tests, where it was assumed that various courses of action might depend upon which hypothesis was accepted, and their choice of α and β were meant to be informed decision-theoretically in terms of the badness of type I and type II errors.

Review Questions

1. Suppose that you want to know whether zinc reduces the duration of the common cold. Describe, in broad outline, how you would

investigate this question using Fischer's *Test of Significance*. What objections would a Bayesian raise to this test?

2. Suppose you know that a coin either has a bias of 50% towards heads or a bias of 90% towards heads, and you wish to know which. Describe how you would investigate this question using Neyman & Pearson's Significance Test.

3. What is 'Lindley's Paradox', and how could it be used to argue against Neyman & Pearson's Significance Test?

4. Suppose you know that a coin either has a bias of 50% towards heads or a bias of 90% towards heads, and you wish to know which. Describe how you would investigate this question using Bayesian statistics. What's an objection that someone could raise to this procedure?

5. Explain what the likelihoodist says about evidence. Suppose you know that a coin either has a bias of 50% towards heads or a bias of 90% towards heads, and you wish to know which. Would a likelihoodist tell you how to settle this question? Why or why not?

5
Arguments for Probabilism

5.1 Probabilism and Its Detractors

Recall, the Bayesian interprets (at least some) probabilities as a rational person's *degrees of belief*, or *credences*. Because of this, they are committed to at least the following rational norm:

Probabilism A rational person's credences will obey the laws of probability

That is, if you are rational and if $\mathbb{C} : \mathcal{A} \to \mathbb{R}$ is your credence function (where $\mathcal{A} \subseteq \mathcal{P}(\mathcal{W})$ is an algebra), then we will at least have that

Non-negativity None of your credences are negative — For all $A \in \mathcal{A}$, $\mathbb{C}(A) \geqslant 0$

Normalization Your credence in any necessary truth is 100%. — $\mathbb{C}(\mathcal{W}) = 1$

Finite Additivity The sum of your credences in two incompatible propositions is equal to your credence in their union. — For any $A, B \in \mathcal{A}$, if $AB = \varnothing$, then $\mathbb{C}(A \cup B) = \mathbb{C}(A) + \mathbb{C}(B)$.

And perhaps we will also want your credences to satisfy further rationality constraints, like

Countable Additivity The sum of your credences in countably many disjoint propositions is equal to your credence in their union. — For any $A_1, A_2, \cdots \in \mathcal{A}$, if $A_i A_j = \varnothing$ for each i and j, then $\mathbb{C}(A_1 \cup A_2 \cup \cdots) = \mathbb{C}(A_1) + \mathbb{C}(A_2) + \cdots$.

Conglomerability If Π is any partition (each cell of which is included in $\mathcal{A}$) and A is any proposition from $\mathcal{A}$, then there will always be some $P_l \in \Pi$ and some $P_h \in \Pi$ such that your credence in A no less than your conditional credence in A, given P_l, and no greater than your conditional credence in A, given P_h. — $\inf_{P \in \Pi} \mathbb{C}(A \mid P) \leqslant \mathbb{C}(A) \leqslant \sup_{P \in \Pi} \mathbb{C}(A \mid P)$

Throughout, when I say that your credences are 'probabilistic', I'm only talking about non-negativity, normalization, and finite additivity. (We'll come back to countable additivity later on.)

It's worth recognizing that probabilism is non-trivial and indeed has been denied. For instance, Arthur Dempster and Glenn Schafer gave a theory of what they called 'belief functions' which is non-probabilistic. (I'll just call it a 'Dempster-Schafer function', and I'll write it 'bel'.) A Dempster-Schafer function is any function from propositions in $\mathcal{A}$ to real numbers that satisfy the following axioms:

Zero Normalization $bel(\varnothing) = 0$

Unit Normalization $bel(\mathcal{W}) = 1$

Superadditivity For any $A_1, A_2, \ldots, A_n \in \mathcal{A}$, if $A_i A_j = \varnothing$ for each $i, j \leqslant n$, then

$$bel\left(\bigcup_{i=1}^{n} A_i\right) \geqslant \sum_{j=1}^{n} \sum_{\substack{J \subseteq \{1,2,\ldots,n\}: \\ \#J = j}} (-1)^{j+1} bel\left(\cap_{j \in J} A_j\right)$$

Here's another way of thinking about a Dempster-Schafer function: take the algebra of propositions $\mathcal{A}$ and assign non-negative numbers to these propositions such that the number given to $\varnothing$ is zero and such that the numbers sum up to 1.

This kind of function is called a *mass* function; and any mass function will determine a Dempster-Shafer function bel, via the identity $bel(A) = \sum_{B:B \subseteq A} mass(B)$. For instance, in the example from the margin, $bel(\{1\}) = 0$, $bel(\{1,2\}) = 0.5$, $bel(\{2,3\}) = 0.3$, and $bel(\{1,2,3\}) = 1$.

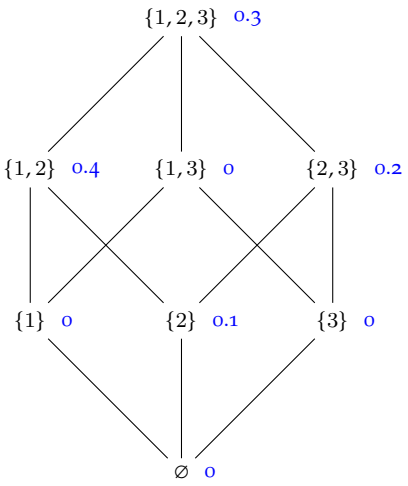


Figure 5.1: A sample mass function.

To interpret what's going on here: think about 1, 2, and 3 as three different horses in a horse race. It could be that you have some evidence that 3 will lose, some evidence that 1 will lose, and some evidence that 2 will win. In that case, when I ask how confident you are that 3 will lose ($\{1,2\}$), you'll just add up the evidence you have that 3 will lose, and the evidence you have that 2 will win, getting that you're one half confident that 3 loses, $bel(\{1,2\}) = 0.5$. And when I ask how confident you are the 3 will win, you'll notice that you don't have any evidence suggesting that 3 will win, so you'll have no confidence that 3 wins, $bel(\{3\}) = 0$. (In the case of a finite algebra, if a mass function only gives positive values to the singleton propositions, then it will determine a probability distribution.)

Bayesians are committed to thinking that this approach to degrees of belief is fundamentally mistaken—but a bunch of smart people have taken it very seriously. What do Bayesians have to say about why they are wrong? Here, we're going to consider three different arguments for the conclusion that your credences should be probabilities (and *not*, for instance, Dempster-Shafer functions).

5.1.1 *Representation Theorems*

The first justification of probabilism piggybacks on a justification of the norm of *expected utility maximization*. So it's worth spending a bit of time thinking about what this norm says. The norm of expected utility maximization says that you should prefer acts with higher expected utilities and disprefer acts with lower expected utilities.

Example 10. *Before you are two boxes. The first one contains a bowl of jellybeans. The contents of the second one are hidden from you. Yesterday I rolled a fair three sided die. If (but only if) the die landed on 3, I put a chocolate ice cream cone in the second box. The utility you attach to a chocolate ice*

cream cone is 4 utiles and the utility you attach to a jelly bean is 1 utile. The utility of getting neither a jellybean nor a chocolate ice cream cone of 0 utiles. You can take at most one box.

You can think of utilities as something like the strength of your desires. When we say that the utility of the ice cream cone is 4, the utility of the jellybeans is 1, and the utility of neither is 0, we are saying that the degree to which you want the ice cream more than the jellybeans is three times larger than the degree to which you want the jellybeans more than nothing.

In this example, you have three acts available to you: you can take no boxes, you can take the first box, or you can take the second box. The utilities of the first two options are known: taking no boxes gets you a guaranteed 0 utiles; and taking the first box gets you a guaranteed 1 utile. So you should prefer taking the first box to taking no box:

$$\text{no box} \prec \text{box 1}$$

But the utility you'll get from the second box is unknown. If the die lands 1 or 2, then $\mathcal{U}(\text{box 2}) = 0$; whereas, if the die lands on 3, then $\mathcal{U}(\text{box 2}) = 4$. While you don't know what this utility is, you can calculate its *expectation*.

$$\begin{aligned}\mathbb{E}[\mathcal{U}(\text{box 2})] &= \mathbb{C}(\mathcal{U}(\text{box 2}) = 0) \cdot 0 + \mathbb{C}(\mathcal{U}(\text{box 2}) = 4) \cdot 4 \\ &= 2/3 \cdot 0 + 1/3 \cdot 4 \\ &= 4/3\end{aligned}$$

Since $4/3 > 1$, expected utility theory says that you should prefer taking box 2 to taking box 1:

$$\text{box 1} \prec \text{box 2}$$

Why should we think that there are utilities like this? And why should we accept this norm?

One answer to this question appeals to a theorem known as a 'representation theorem'. This theorem says that, so long as your preferences between acts satisfy certain rationality constraints, those preferences will be 'representable' with a utility function $\mathcal{U}$ and a probabilistic credence function $\mathbb{C}$ such that one act is preferred to another if and only if the first act has a higher expected utility than the second (relative to that pair of credence and utility functions).

Representation Theorem (Schema) If your preferences satisfy constraints $\mathscr{C}$, then there is a probabilistic credence function $\mathbb{C}$ and a utility function $\mathcal{U}$ such that, for any two acts a and b,

$$a \succeq b \quad \text{iff} \quad \mathbb{E}[\mathcal{U}(a)] \geqslant \mathbb{E}[\mathcal{U}(b)]$$

(where $\mathbb{E}$ is the expectation associated with the probability function $\mathbb{C}$)

For instance, Leonard Savage proved a representation theorem from

the rationality constraints given in the margin. But there are other representation theorems out there—including ones proved by Ramsey, von Neumann and Morgenstern, Jeffrey and Bolker (for evidential decision theory), and Joyce (for causal decision theory).

These theorems afford us an argument (schema) for probabilism:

P1) If you are rational, then your preferences over acts will satisfy the constraints $\mathscr{C}$.

P2) If you satisfy the constraints $\mathscr{C}$, then you can be represented as maximizing expected utility relative to a probabilistic credence function.

$\therefore$ C) If you are rational, you will have a probabilistic credence function.

The second premise is just a theorem, so there's no objections to be raised there. Some objections to the argument focus on the first premise—for instance, is totality ($a \succeq b$ or $b \succeq a$) really a rational requirement? But Alan Hájek raises a deeper concern: it looks like the argument is invalid. He give the following parody of the argument: if your preferences satisfy certain constraints, then you can be represented as though your decisions were the product of warring voodoo spirits. You should satisfy these constraints. Therefore, your decisions should be the product of warring voodoo spirits.

The proponents of this argument were thinking that there's no difference between *having* a certain credence and utility function pair and *being representable* as having those credences and utilities. They were driven by a behavioristic understanding of mental state ascriptions, according to which there is no deep fact-of-the-matter about what you believe beyond the difference those beliefs make to action. If we reject this kind of behaviorism, then the argument will be invalid.

Titelbaum points out that, even if we reject the behavioristic assumptions underlying the old argument, we can reformulate the argument. He appeals to the following kind of theorem:

Revised Representation Theorem (Schema) If your preferences satisfy constraints $\mathscr{C}$, then there is a credence function $\mathbb{C}$ (unique up to positive scalar transformation) and a utility function $\mathcal{U}$ (unique up to positive linear transformation) such that, for any acts a and b,

$$a \succeq b \quad \text{iff} \quad \mathbb{E}[\mathcal{U}(a)] \geqslant \mathbb{E}[\mathcal{U}(b)]$$

(where $\mathbb{E}$ is the expectation associated with the probability function $\mathbb{C}$) Moreover, one of the scalar multiples of $\mathbb{C}$ is a probability.

He then offers the following argument for probabilism, which assumes the norm of expected utility maximization:

P1) If you are rational, then your preferences over acts will satisfy the constraints $\mathscr{C}$.

P2) If you are rational, then you will prefer acts with greater expected utility

Savage represents acts with functions from $\mathcal{W}$ to outcomes. And he assumes that, for any two acts a and b, and any proposition E, there is a third act, $a_E b$, which has the same outcome as a whenever E is true and has the same outcome as b otherwise. An act is *constant* iff it leads to the same outcome in every world.

Then, he lays down the following rationality constraints on preference: for any acts a, b, c, and d, and any propositions E and F:

1. either $a \succeq b$ or $b \succeq a$
2. if $a \succeq b$ and $b \succeq c$, then $a \succeq c$
3. $c_E a \succeq c_E b$ iff $d_E a \succeq d_E b$
4. if E is non-null, then $a \succeq b$ iff, for all f, $a_E f \succeq b_E f$
5. if $a \succ b$ and $c \succ d$, then $a_E b \succ a_F b$ iff $c_E d \succ c_F d$
6. for some constant acts a, b, $a \succ b$.
7. if $a \succ b$, then there's a finite partition $\{E_1, E_2, \ldots, E_N\}$ such that, for every i, $a \succ c_{E_i} b$ and $c_{E_i} a \succ b$.

P3) If you satisfy the constraints $\mathscr{C}$ and prefer acts with greater expected utility, then your credences are a scalar multiple of a probability

∴ C) If you are rational, your credences will be a scalar multiple of a probability

Assuming that we take the choice of 1 for your probability in $\mathcal{W}$ to be an arbitrary choice, this argument gets us probabilism. But unlike the original argument, it requires us to assume expected utility maximization. Also unlike the original argument, there is no version of the revised representation theorem which applies to evidential or causal decision theory.

5.2 *The Dutch Book Argument*

Let's suppose that your utilities are linear in dollars; and let's suppose that your fair price for a \$1 bet on any proposition A is equal to your credence in A. That is, consider the following ticket:

\$1	if A
\$0	else

If you have this ticket in your possession, then it will entitle you to \$1 if A turns out to be true. Otherwise, it is worthless. We will assume that you are willing to pay up to $\$\mathbb{C}(A)$ for this ticket. This could be justified on the grounds of expected utility maximization, for the expected utility of possessing this ticket is given by

$$\mathbb{C}(A) \cdot 1 + \mathbb{C}(\neg A) \cdot 0 = \mathbb{C}(A)$$

So if the price of the ticket is greater than $\mathbb{C}(A)$, you will not want to buy it; and if the price of the ticket is less than $\mathbb{C}(A)$, you will want to buy it. Correlatively: if the price of the ticket is greater than $\mathbb{C}(A)$, you will want to sell it; and if the price of the ticket is less than $\mathbb{C}(A)$, you will not want to sell it.

Now: suppose that there is a race between three horses and your credences about the winner are given by the Dempster Shafer belief function from figure 1, so that $bel(\{1,2\}) = 0.5$ and $bel(\{3\}) = 0$. And suppose I offer to buy bet 1 from you.

Bet 1

\$1	if horse 1 or 2 wins
\$0	else

price: 51¢

Since the price of this bet is greater than your credence that either horse 1 or 2 wins, you will be willing to sell it. Next, suppose I offer to buy bet 2 from you. Since the price of the bet is greater than your credence that horse 3 wins (which is zero), you will be willing to sell it.

Bet 2

\$1	if horse 3 wins
\$0	else

price: 1¢

But now you're in trouble. Let's consider the possible outcomes for you:

	Horse 1 or 2 wins	Horse 3 wins
Net profit from selling bet 1	−49¢	51¢
Net profit from selling bet 2	1 ¢	−99¢
Overall net profit	−48¢	−48¢

No matter what happens, you're going to lose 48¢ . A combination of bets like this (which is guaranteed to lose you money no matter what) is called a 'Dutch book'. (The origin of the name is unclear). So what we've seen is that, if your credences are given by the Dempster Shafer function from figure 1, and you use your credences as your fair betting prices, then you will purchase a Dutch book.

In fact, something similar will happen *whenever* you have non-probabilistic credences. This is a general theorem:

Dutch Book Theorem If your credences are not probabilistic, and you use those credences as your fair betting prices, then a Dutch book can be constructed against you.

This affords us the following argument:

P1) If you are rational, then you will not be susceptible to a Dutch book.

P2) You are susceptible to a Dutch book if your credences are not probabilistic.

∴ C) If you are rational, then your credences will be probabilistic.

You might worry about the first premise. We've shown that non-probabilistic credences are Dutch bookable, but we haven't shown that probabilistic credences *aren't* Dutch-bookable. Perhaps there's just no way to avoid being Dutch-bookable—or perhaps the only way to avoid being Dutch-bookable is to not use your credences as your fair betting odds. We can allay these kinds of concerns with the following converse theorem:

Converse Dutch Book Theorem If your credences are probabilistic, and you use those credences as your fair betting prices, then a (static and finite) Dutch book cannot be constructed against you.

When I say that the Dutch book is 'static', I mean that the bets comprising the Dutch book are all bought or sold while your credences remain fixed. When I say that it is 'finite', I mean that there are only finitely many bets included in the book. (We'll come back to infinite numbers of bets shortly.)

Against the second premise: suppose somebody is a recovering gambling addict, and they turn down any bets they're offered (even if they think that the bets are more than fair). They just don't want to relapse. Moreover, this person's credences are not probabilistic. This person seems to be a counterexample to the second premise. They have non-probabilistic credences, but they are not susceptible to a Dutch book.

Relatedly: the Dutch book argument seems to be dealing in the wrong kind of reasons. We want to know why it is *epistemically* required to have probabilistic credences, but the argument is pointing us to a *pragmatic* defect. If we're pragmatists who think that the epistemic norms on belief are derivative from pragmatic norms, then perhaps this will move us. But many of us are not pragmatists. We think, for instance, that there can be pragmatic costs to being epistemically

rational. (Perhaps believing certain truths will leave you depressed and have no other benefits; but we still think that it is epistemically required to believe those truths)

In response to this, many have tried to 'depragmatize' the Dutch book arguments. Roughly, their thought is that the Dutch book argument reveals an *evaluative* inconsistency in non-probabilistic credences. They take their to be a normative link between your credences and how you evaluate certain bets:

Evaluation-Credence Link If your credence in A is x, and your utilities are linear in dollars, then you should evaluate the ticket

\$1	if A
\$0	else

price: y¢

as valuable if $x > y$, disvaluable if $y < x$, and neutral if $x = y$.

Then, the 'depragmatized' version of the Dutch book argument goes like this:

P1) If your credences are not probabilistic, then you'll evaluate each of a collection of bets are individually valuable, and you'll evaluate their collection as disvaluable

P2) If your credences are rational, they will not lead you to evaluate each of a collection of bets as individually valuable and yet evaluate their collection as disvaluable.

∴C) If your credences are rational, then they will be probabilistic.

(P1) comes from the Dutch book theorem and the Evaluation-Credence Link. One objection to (P1) targets this link. We've seen that the link follows if we evaluate betting tickets in terms of their expected utilities, but as we learnt on the first problem set, expectations are closely linked to probabilities. If you're willing to give up on probabilism, you may also want to sever the link between expected utility and rational preference.

Another objections questions (P2). We could argue for this premise with the following two principles:

Equivalence Principle If two betting arrangements have the same payoffs in every possible world, then you should value them in exactly the same way.

Package Principle If you regard betting arrangement 1 as valuable and you regard betting arrangement 2 as valuable, then you should regard the *package* of both betting arrangements together as valuable.

Applied to our sample Dutch book from above: the package principle says that, if you regard selling bet 1 as valuable, and you regard selling bet 2 as valuable, then you should regard selling *both* bet 1 and bet 2 as valuable. And since selling both bets 1 and 2 together is equivalent to a guaranteed 48¢ loss, the equivalence principle says that you

should regard a 48¢ loss as valuable. Insofar as you don't, your values are inconsistent. And this inconsistency is due to the failure of your credences to be probabilistic.

Non-probabilists have objected to the package principle. They contend that it smuggles in precisely the kind of additivity that we were trying to establish. (To get your evaluation of the package of bets, just 'add up' your evaluation of each bet individually.)

5.2.1 *The Dutch Book Argument for Countable Additivity*

There are Dutch book arguments for countable additivity as well. Suppose that we have a lottery with a countable infinity of tickets in it, and your credence that any given ticket wins is zero. Then, you will be willing to sell each of the infinitely many bets of the following form:

\$1	if ticket n wins the lottery
\$0	else

price: $1/2^n$¢

With each sale, you will take in a positive monetary amount and you are 100% confident that you won't have to pay out. However, when you package all of these bets together, you face a sure loss. Together, the tickets will only bring in 1¢ since

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1$$

But exactly one of the bets will pay out, so you will lose \$1, and your net profit will be a guaranteed −99¢.

But notice that we can give a similar argument for *full* additivity. Suppose a random number will be selected from between 0 and 1, and you sell all of the uncountably many bets

\$1	if number n is selected
\$0	else

price: \$0

Individually, you judge each of these bets as fair, even though, collectively, you recognize that they are a guaranteed loss.

Defenders of countable additivity have pointed out that, in the first Dutch book, the individual sales are all *favorable*; whereas, in the second Dutch book, the individual tickets are merely *fair*, but not favorable.

5.2.2 *The Dutch Book Argument for Conglomerability*

Suppose that you violate Comglomerability. Then, there's a partition Π and a proposition A such that, for every $P \in \Pi$, $\mathbb{C}(A) > \mathbb{C}(A \mid P)$. Then, you will be happy to sell the following bet on A:

\$1	if A is true
\$0	else

price: $\mathbb{C}(A)$¢

And you will be happy to buy each of the following conditional bets:

\$1	if A is true
\$0	else

$$\text{price: } \begin{cases} \mathbb{C}(A \mid P)\text{¢} & \text{if } P \\ 0 & \text{if } \neg P \end{cases}$$

But since $\mathbb{C}(A) > \mathbb{C}(A \mid P)$, you will certainly lose money on this combination of bets. (Moreover, each of these bets is favorable.)

We'll think more about these kinds of bets when we consider the Dutch book argument for conditionalization.

5.3 *Accuracy Arguments for Probabilism*

Al Hájek poses the following SAT analogy problem:

belief : truth :: credence : _____

The idea is this: when we think about binary (on/off) belief, truth is the standard of *correctness* or *vindication*—even if the belief isn't known, or isn't justified, or is unreasonable, if the belief turns out to be true, then you got it right (in spite of your irrationality). What is the parallel notion for *degree* of belief? When are your credences *correct*? When are they *vindicated*?

Hájek himself gives the answer "chance". (Question: chance *at what time*?) Hájek is looking for something degreed to serve as the potential vindication for credence. Since truth doesn't come in degrees (let's suppose), he thinks we need something that *does* come in degrees to serve as the standard of correctness.

Another possible answer is "*calibration*".

Calibration your credences are (perfectly) *calibrated* iff, for every x, amongst the propositions you give a credence of x to, exactly $100x\%$ of them are true.

For instance, if 10% of the propositions you give a credence of 0.1 to are true, 20% of the propositions you give a credence of 0.2 to are true, and so on and so forth, then your credences are perfectly calibrated.

Titelbaum argues that calibration isn't the right answer to Hájek's SAT analogy question, either. He gives the following counterexample: suppose that there are two races for Senate: candidates A and B face off against each other, and candidates C and D face off against each other. Nate Silver and Alan Lichtman both make forecasts about who will win.

Candidate:	A^*	B	C^*	D
Silver's forecast:	90%	10%	90%	10%
Lichtman's forecast:	50%	50%	50%	50%

Suppose that the starred candidates (A and C) won. It seems like Silver should be at least *more* vindicated than Lichtman; but Lichtman's forecasts are perfectly calibrated, whereas Silver's forecast is not. So if we think of vindication in terms of calibration, we will say that Lichtman was vindicated. But this seems like the wrong result.

James M. Joyce gives a different answer to Hájek's SAT analogy: *(gradational) accuracy*. Whereas Hájek's tried to match the degreed nature of credence in the thing that was doing the vindicating, Joyce makes vindication itself a degreed notion. Truth is the thing that vindicates both belief and credence. The only that that is degreed is *how close to truth* your credences are to truth—that is, the only thing that's degreed is *how* vindicated your credences are.

Joyce thinks that we have to introduce some way of measuring *how accurate* credences are. Here's a sample way of doing that measuring:

Quadratic Measure of Inaccuracy (Local) The inaccuracy of a credence x in a proposition A at a possible world w, $\mathcal{I}(x, A, w)$, is the square of the difference between x and A's truth-value at w.

$$\mathcal{I}(x, A, w) = (\mathbf{1}_A(w) - x)^2$$

Against Hájek, consider the following case: a fair coin is about to be flipped. You believe it will land heads and you have a credence of 1 that it will land heads. I don't believe it will land heads and have a credence of 1/2 that it will land heads. The coin is flipped and in fact landed heads. While your belief may not have been *justified* or *reasonable*, it was nonetheless *true*. Carrying the analogy forward, shouldn't your (perhaps unjustified, perhaps unreasonable) high credence similarly count as vindicated? But if credence is to chance as belief is to truth, your high credence in heads won't count as vindicated; instead, my middling credence in heads will be vindicated.

The Quadratic measure is also called the 'Brier' measure—after Glenn Brier, who proposed it as a way of gauging the accuracy of weather forecasts.

Quadratic Measure of Inaccuracy (Global) The inaccuracy of a credence function, $\mathbb{C}$, at a world w, is the sum of the quadratic inaccuracy, at w, of $\mathbb{C}$'s credence in every proposition.

$$\mathcal{I}(\mathbb{C}, w) = \sum_{A \in \mathcal{A}} \mathcal{I}(\mathbb{C}(A), A, w) = \sum_{A \in \mathcal{A}} [\mathbf{1}_A(w) - \mathbb{C}(A)]^2$$

Joyce uses this gradational measure of accuracy to mount a *non-pragmatic* argument for probabilism. To give you a flavor for how that argument goes, consider an example in which there are just two possible worlds: w_H, at which the coin lands heads, and w_T, at which the coin lands tails. Let's assume for now that your credence in $\mathcal{W} = \{w_H, w_T\}$ is 100%, and that your credence in $\varnothing$ is 0% (we will come back to those credences in a bit).

Then, we can think about every possible assignment of credences to w_H and w_T as some point in the 1×1 box in figure 5.2. And the essence of Joyce's defense of probabilism can be seen in figure 5.3. In that diagram, the blue dot in the upper left hand corner is the function which assigns 1 to all truths and 0 to all falsehoods, if the coin lands tails. It is the truth-function for the world w_T. Likewise, the orange dot in the lower right hand corner is the truth-function for the world w_H. Suppose that you have the non-probabilistic credence distribution shown in figure 5.3—your credences in w_H and w_T do not sum up to 100%. Then, the orange curve shows all of the credence functions which are as inaccurate as you are, if the coin in fact lands heads. And the blue curve shows all of the credence functions which as as inaccurate as you are, if the coin in fact lands tails. Then, notice that, no matter whether the coin lands heads or tails, the probabilistic credence function shown in figure 5.3 will be *less* inaccurate than you. You are, in other words: *accuracy dominated*.

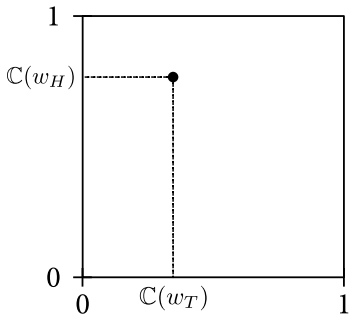


Figure 5.2: Each point in the 1×1 box represents a possible assignment of credence to the propositions $\{w_H\}$ and $\{w_T\}$.

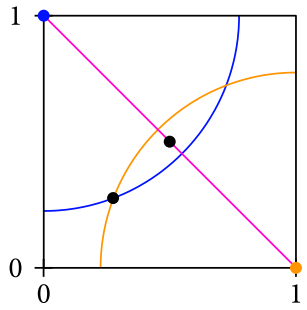


Figure 5.3: The probability functions are all of the credences on the pink line. The blue dot is the perfectly vindicated function if the coin lands heads; the orange dot is the perfectly vindicated function if the coin lands tails. If we use the Brier measure, then the given non-probability will have a greater inaccuracy than the given probability *no matter what*.

Accuracy Domination The credence function $\mathbb{C}$ is *accuracy dominated* by the credence function $\mathbb{C}^*$ iff, for every possible world $w \in \mathcal{W}$,

$$\mathcal{I}(\mathbb{C}, w) \geqslant \mathcal{I}(\mathbb{C}^*, w)$$

and, for some possible world $w \in \mathcal{W}$,

$$\mathcal{I}(\mathbb{C}, w) > \mathcal{I}(\mathbb{C}^*, w)$$

The same thing will happen to you if your credence in the necessary truth $\{w_H, w_T\}$ is anything less than 100%, and if your credence in the necessary falsehood $\varnothing$ is anything greater than 0%. The credence function which is exactly like you but gives those propositions credence 1 and 0, respectively, will accuracy dominate yours.

In fact, this isn't just true for this particular example. It is true *in general*.

Theorem If we measure inaccuracy with the Brier measure, then: (1) every non-probabilistic credence function is accuracy dominated; and (2) no probabilistic credence function is accuracy dominated.

This is the basic idea behind Joyce's *non-pragmatic* argument for probabilism: you should have a probability function because, otherwise, your credences will be accuracy-dominated. Having a non-probability function guarantees that you are further from the truth that you could otherwise be. And this is irrational. So having non-probabilistic credences is irrational.

Now, there are arguments for using the Brier measure. But Joyce doesn't want to rest his hat on this particular way of measuring inaccuracy. Instead, he points to a collection of features that he thinks any measure of accuracy should have, and he proves a theorem showing that, for *any* measure of accuracy which has these features, non-probabilities will be accuracy-dominated, and probabilities will not be accuracy-dominated. The features are given in the margin.

One of the key assumptions that we will dwell on is *convexity*: which says that the 'equal inaccuracy' curves in figure 5.3 must be convex—the mixture of any two points on those curves must be strictly less inaccurate (more accurate) than the points on the curve themselves are. Joyce's summary of this is as follows:

> [Convexity] is motivated by the intuition that extremeism in the pursuit of accuracy is no virtue. It says that if a certain change in a person's degrees of belief does not improve accuracy, then a more radical change in the same direction and of the same magnitude should not improve accuracy either...If it did not hold, one could have absurdities like this: "I raised by confidence levels in $[A]$ and $[B]$ and my beliefs became less accurate overall, so I raised my confidence levels in $[A]$ and $[B]$ again, by exactly the same amounts, and the initial accuracy was restored.

But Paul Horwich and Patrick Maher have both defended a measure of accuracy which does not satisfy convexity. This is the *linear* measure of inaccuracy.

Linear Measure of Inaccuracy (Local) The inaccuracy of a credence x in a proposition A at a possible world w, $\mathcal{I}(x, A, w)$, is the absolute value of the difference between x and A's truth-value at w.

$$\mathcal{I}(x, A, w) = |\mathbf{1}_A(w) - x|$$

Linear Measure of Inaccuracy (Global) The inaccuracy of a credence function, $\mathbb{C}$, at a world w, is the sum of the quadratic inaccuracy, at w, of $\mathbb{C}$'s credence in every proposition.

$$\mathcal{I}(\mathbb{C}, w) = \sum_{A \in \mathcal{A}} \mathcal{I}(\mathbb{C}(A), A, w) = \sum_{A \in \mathcal{A}} |\mathbf{1}_A(w) - \mathbb{C}(A)|$$

The linear measure will not work for the accuracy-dominance argument. Consider the three credence distributions over (w_H, w_T): $(1/3, 1/3)$, $(1/2, 1/2)$, and $(2/3, 2/3)$ (shown in figure 5.4).

If we use the linear measure, then all three of these credence distributions will have exactly the same inaccuracy at every possible world. So the non-probability won't be accuracy-dominated by the probability. (And, in general, there won't be anything accuracy-dominating

Here are the features Joyce assumed a reasonable measure of accuracy would have:

Structure For every $w \in \mathcal{W}$, $\mathcal{I}(\mathbb{C}, w)$ is a non-negative continuous function of $\mathbb{C}$ which goes to infinity in the limit as $\mathbb{C}(A)$ goes to infinity for any $A \in \mathcal{A}$.

Extensionality At each possible world $w \in \mathcal{W}$, $\mathcal{I}(\mathbb{C}, w)$ is a function of nothing other than the truth-values of the propositions $A \in \mathcal{A}$ and the credence which $\mathbb{C}$ gives to those propositions.

Truth-Directedness If $\mathbb{C}(A) = \mathbb{C}^*(A)$ for every $A \in \mathcal{A}$ other than B, and $\mathbb{C}(B)$ is closer to the truth-value of B at w than $\mathbb{C}^*(B)$ [i.e., $|\mathbf{1}_B(w) - \mathbb{C}(B)| < |\mathbf{1}_B(w) - \mathbb{C}^*(B)|$, then $\mathcal{I}(\mathbb{C}, w) < \mathcal{I}(\mathbb{C}^*, w)$.

Normality If $|\mathbf{1}_A(w) - \mathbb{C}(A)| = \mathbf{1}_A(w^*) - \mathbb{C}(A)|$ for every $A \in \mathcal{A}$, then $\mathcal{I}(\mathbb{C}, w) = \mathcal{I}(\mathbb{C}, w^*)$.

Convexity If $\mathcal{I}(\mathbb{C}, w) = \mathcal{I}(\mathbb{C}^*, w)$, then $\mathcal{I}([\mathbb{C} + \mathbb{C}^*]/2, w) \leqslant \mathcal{I}(\mathbb{C}, w)$, with equality only if $\mathbb{C} = \mathbb{C}^*$.

Symmetry If $\mathcal{I}(\mathbb{C}, w) = \mathcal{I}(\mathbb{C}^*, w)$, then for any $\lambda \in [0, 1]$, $\mathcal{I}(\lambda\mathbb{C} + (1 - \lambda)\mathbb{C}^*, w) = \mathcal{I}((1 - \lambda)\mathbb{C} + \lambda C^*, w)$.

He then was able to prove a theorem showing that, if $\mathcal{I}$ satisfies these assumptions, then (1) every non-probabilistic credence function is accuracy-dominated; and (2) no probabilistic credence function is accuracy dominated.

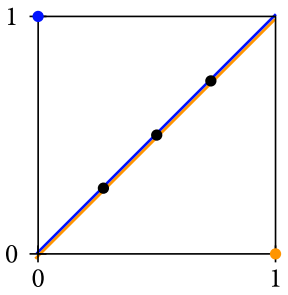


Figure 5.4: The blue line is the set of all credences as inaccurate as $(1/2, 1/2)$ if the coin landed tails; the orange line is the set of all credences as inaccurate as $(1/2, 1/2)$ if the coin landed heads.

the non-probability $(1/3, 1/3)$ if we measure inaccuracy with the linear measure.

Using more recent results, we can offer a different argument for probabilism. This argument again singles out a class of accuracy measures and then shows a comparable theorem: any non-probability is going to be accuracy dominated on this measure, and no probability will be.

To appreciate this later argument, let's start by noticing something odd about the linear measure of accuracy. Suppose that you know for sure that the coin is biased 2/3rds towards heads, and (therefore) your credence that the coin lands tails is 1/3rd and your credence that the coin lands heads is 2/3rds: that is, your credences are given by the probabilistic $(2/3, 1/3)$. Then, ask yourself: which credence function do you *expect* to be most accurate? That is: ask yourself: which credence function has the greatest *expected* accuracy (or the lowest expected inaccuracy)?

The answer will depend upon which accuracy measure we use. Suppose first that we use the quadratic, or Brier, measure. Then, it will turn out that the (unique) credence function which maximizes expected accuracy will be your current credences: $(2/3, 1/3)$ (see the quick calculations in the margin if you know calculus). But the same is not true for the linear measure. If you use the linear measure, then the credences with the lowest *expected* inaccuracy are the extremal credences $(1, 0)$, which is certain the coin will land heads.

So Joyce notes that there's something *self-defeating* about holding the credences $(2/3, 1/3)$ and measuring inaccuracy with the linear measure—by your own lights, it would be better to switch to the extremal credences $(1, 0)$. Joyce assumes that, if your credences are self-defeating in this way, then they are irrational.

If we use the quadratic measure, then the expected inaccuracy of the credences (c, d) is given by

$$2/3 \cdot [(1-c)^2 + d^2] + 1/3[c^2 + (1-d)^2]$$

We can find the minimum by taking the first-order condition for the choice of c (while treating d as a constant):

$$\begin{aligned} 2/3 \cdot [-2(1-c)] + 1/3[2c] &= 0 \\ -4 + 4c + 2c &= 0 \\ c &= 2/3 \end{aligned}$$

And likewise, taking the first-order condition for the choice of d (while treating c as a constant):

$$\begin{aligned} 2/3 \cdot [2d] - 1/3[2(1-d)] &= 0 \\ 4d - 2 + 2d &= 0 \\ d &= 1/3 \end{aligned}$$

(You can verify the second-order conditions and the boundary conditions yourself, but if you trust me, I can let you know that this is the unique global minimum)

Self-Recommendation is Required If some credence function other than your own has an expected inaccuracy at least as low as yours, then it is irrational to hold your current credences.

Equivalently: if it is rational to hold your credences, then no other credence function has a lower expected inaccuracy than your own. So rational credences should expect themselves to be best. Lewis offers the following analogy:

> It is as if *Consumer Bulletin* were to advise you that *Consumer Reports* was a best buy whereas *Consumer Bulletin* itself was not acceptable; you could not possibly trust *Consumer Bulletin* completely thereafter.

Lewis's point is that, if your credences are not self-recommending, then, if you rely on them, you won't rely on them. So you can't rely on a non-self-recommending credence function.

If every *probability* function uniquely recommends itself (on a certain measure of accuracy), then that measure of accuracy is said to be *strictly proper*.

Strict Propriety $\mathcal{I}$ is a *strictly proper* measure of inaccuracy iff, for any

probability function $\mathbb{P}$ and any credence function $\mathbb{C}$,

$$\mathbb{E}_{\mathbb{P}}[\mathcal{I}(\mathbb{P}, w)] \leqslant \mathbb{E}_{\mathbb{P}}[\mathcal{I}(\mathbb{C}, w)]$$

with equality only when $\mathbb{P} = \mathbb{C}$.

I will inform you of the following theorem (but won't go through the proof):

Theorem If $\mathcal{I}$ is any strictly proper and continuous measure of inaccuracy, then (1) every non-probability is accuracy-dominated; and (2) no probability is accuracy-dominated.[1]

[1] This theorem was proven by Predd *et al*, 2009. "Probabilistic Coherence and Proper Scoring Rules". In *IEEE Transactions on Information Theory*, vol. 55 (10): 4786–4792.

Putting together all of the pieces, we have the following non-pragmatic argument for probabilism:

P1) For any probability function, there is *some* evidence you could hold that would make it rationally permissible to have that probability function.

P2) *Self-Recommendation is Required*: If some credence function other than your own has an expected inaccuracy at least as low as yours, then your credences are irrational.

P3) Small changes in credence shouldn't lead to big changes in accuracy, so accuracy should be measured in a continuous way.

∴ C1) Accuracy should be measured in a strictly proper and continuous way.

P4) *Theorem*: If accuracy is measured in a strictly proper and continuous way, then any non-probabilistic credence function is accuracy-dominated.

P5) *Accuracy Domination is Irrational*: if your credences are accuracy-dominated, then they are irrational.

∴C2) If your credences are not probabilistic, then they are irrational.

You might think that (P1) is question-begging; but it's important to recognize that it's much weaker than probabilism. (P1) just says that, in *some* circumstances, having a probability is *permissible*. But probabilism says that, in *all* circumstances, having a probability is *required*. Note also that even a proponent of the Dempster-Shafer theory will be happy to accept (P1). Joyce further argues for (P1) by pointing out that, for any probability function, you could have the evidence that the objective chances are given by exactly that probability function.

Review Questions

1. What is probabilism?
2. What is a representation theorem, and how could it be used to argue for probabilism? What is Hájek's objection to this argument?
3. What is a 'Dutch book'? What is the 'Dutch book argument' for probabilism? Explain how this argument can be 'depragmatized'.
4. Imagine that we are going to flip a coin, and you have credences in the two propositions 'the coin lands heads' and 'the coin lands tails'. Assume the quadratic measure of accuracy to argue that your credences should be probabilistic. (Draw a picture.)
5. What is *strict propriety*, and why does assuming that a measure of accuracy should be strictly proper help us to argue for probabilism? How does Joyce argue that a measure of accuracy should be strictly proper?

6
Arguments for Conditionalization

6.1 What does Conditionalization say?

Recall, the Bayesian interprets (at least some) probabilities as a rational person's degrees of belief or *credences*. They are committed to two fundamental norms governing these degrees of belief. Firstly, they should be probabilities. Secondly, when you acquire new evidence, they should be revise in line with the norm of *conditionalization*:

> *Conditionalization* You should learn from your evidence by conditioning on it. That is, if your *prior* credence in A, at t_0, is $\mathbb{C}_(A)$, then your *posterior* credence in A, at t_1, after learning E, should be $\mathbb{C}(A \mid E)$.

Let's start by getting clearer about what this norm says. First, let's clear up a potential misunderstanding with the notation '$\mathbb{C}(A \mid E)$'. As I'm using the notation here, it is your *prior conditional* credence in A, given E. This is a synchronic feature of your doxastic state at t_0; $\mathbb{C}(A \mid E)$ says how many times more likely than E you think the conjunction AE is. So long as $\mathbb{C}(E) > 0$, $\mathbb{C}(A \mid E)$ is equal to the ratio $\mathbb{C}(AE) \div \mathbb{C}(E)$.

Let's carefully distinguish $\mathbb{C}(A \mid E)$ from how confident you *plan* to be in A, if you end up learning E. I'll write this second quantity '$\mathbb{C}_E(A)$'.

And we should further distinguish your credence revision *plans* from the credences you actually end up adopting, after learning. Throughout, let's suppose that you will learn something in between the times t_0 and t_1. Then, '$\mathbb{C}$' will be your prior credence function at t_0, and '$\mathbb{C}^+$' will be your posterior credence function at t_1. Then, we can separate out the thesis of conditionalization into two subsidiary these:

> *Plan Conditionalization* Before learning, you should plan to have credence $C(A \mid E)$ in A after learning E. That is: you should have
>
> $$\mathbb{C}_E(A) = \mathbb{C}(A \mid E) = \mathbb{C}(AE) \div \mathbb{C}(E)$$

Honor Your Plans After learning, your credence in A should be equal to the credence you planned to have in A, if you learned E. That is: you should have

$$\mathbb{C}^+(A) = \mathbb{C}_E(A)$$

These two norms together imply that, at t_1, your posterior credence in A will be equal to your prior conditional credence in A, given E.

6.1.1 Examples

We've already seen the principle of conditionalization at work in several examples throughout the course. But let's start with an illustrative example to make vivid what kind of normative commitment Conditionalization is:

Example 11. *Daniel the Democrat is intensely partisan. Whenever he hears about a Democratic politician involved in a scandal, he is disposed to be very confident that the Democratic politician has done no wrong. On the other hand, whenever he hears about a Republican politician involved in a scandal, he is disposed to be very confident that the Republican politician has done something wrong. That's not because Daniel thinks that the actions of politicians gives evidence about which acts are wrong. Beforehand, he thinks that the wrongness of being lax with classified information, using your office to enrich political donors, and so on, is independent of whether a Democrat or Republican does these things. Nonetheless, whenever Daniel learns that a Democrat has done one of these things, he is disposed to think that their acts were not wrong; and whenever he learns that a Republican has done one of these things, he is disposed to think that their acts were wrong.*

On the other hand, Melissa the Moderate is equally inclined to think that ϕ-ing is wrong, whether the person who is found to have ϕ-ed is a Democrat or a Republican. She is sometimes inclined to think that a Democrat has done wrong, and sometimes inclined to think that a Republican has done wrong.

The Bayesian says that Daniel is irrational—why? Because he is not updating his credences by conditionalization. We said that he starts out thinking that whether ϕ-ing is wrong is independent of whether a Democrat or a Republican ϕs. So if $\mathbb{C}$ is Daniel's credence function, then we have

$$\mathbb{C}(\phi\text{-ing is wrong} \mid \text{ a Democrat } \phi\text{s}) = \mathbb{C}(\phi\text{-ing is wrong}) = \mathbb{C}(\phi\text{-ing is wrong} \mid \text{ a Republican } \phi\text{s})$$

Yet

$$\mathbb{C}_{\text{a Democrat } \phi\text{s}}(\phi\text{-ing is wrong}) < \mathbb{C}(\phi\text{-ing is wrong}) < \mathbb{C}_{\text{a Republican } \phi\text{s}}(\phi\text{-ing is wrong})$$

So Daniel does not plan to condition on the evidence that a Democrat or a Republican has mishandled classified information, used their office to enrich political donors, and so on. On the other hand, Melissa could very well be conditionalizing on her evidence. So the Bayesian has no objection to Melissa's dispositions to learn from evidence.

Here's another example:

Example 12. *We're going to roll a six-sided die. I will see how it landed, but you will not. I won't tell you exactly how it landed; but I will tell you whether it landed on an odd or even number. I roll the die and tell you that it landed odd. How confident should you be that it landed on a high number* (4, 5, *or* 6)*, if you learn that it landed odd? How confident should be you that it landed on a high number if you learn that it landed even?*

w	$\mathbb{C}(\{w\})$	$\mathbb{C}(\{w\} \mid O)$	$\mathbb{C}(\{w\} \mid E)$
1	1/6	1/3	0
2	1/6	0	1/3
3	1/6	1/3	0
4	1/6	0	1/3
5	1/6	1/3	0
6	1/6	0	1/3

This example is simple, but I want to approach it meticulously, since dotting every i and crossing every t will help us later on. You have credences defined over the possibilities $\{1, 2, 3, 4, 5, 6\}$, given by the probability table in the margin.

In this example, there are two things you might learn: you might learn that the die landed on an odd number, $O = \{1, 3, 5\}$, and you might learn that it landed on an even number, $E = \{2, 4, 6\}$. Your prior conditional credences in high ($H = \{4, 5, 6\}$), given odd, is one third, since

$$\mathbb{C}(H \mid O) = \frac{\mathbb{C}(HO)}{\mathbb{C}(O)} = \frac{\mathbb{C}(\{5\})}{\mathbb{C}(\{1, 3, 5\})} = \frac{1/6}{3/6} = \frac{1}{3}$$

So *Plan Conditionalization* says that you should plan to be one third confident of high, if you learn odd, $\mathbb{C}_O(H) = 1/3$. If you then learn that it landed odd, *Diachronic Conditionalization* says that you should follow through on this plan and adopt the posterior credence $\mathbb{C}^+(H) = 1/3$.

Your prior conditional credence in high, given even, is two thirds, since

$$\mathbb{C}(H \mid E) = \frac{\mathbb{C}(HE)}{\mathbb{C}(E)} = \frac{\mathbb{C}(\{4, 6\})}{\mathbb{C}(\{2, 4, 6\})} = \frac{2/6}{3/6} = \frac{2}{3}$$

So *Plan Conditionalization* says that you should plan to be two thirds confident of high, if you learn even, $\mathbb{C}_E(H) = 2/3$. If you then learn that it landed even, *Diachronic Conditionalization* says that you should follow through on this plan and adopt the posterior credence $\mathbb{C}^+(H) = 2/3$.

Example 13. *There are three prisoners, A, B, and C, scheduled to be executed tomorrow. However, in a show of leniency, the king has decreed that one prisoner be selected* (*at random*) *to be spared execution. The guards know who will be spared, but the prisoners do not. B approaches one of the guards, and says "I know that you're not allowed to tell me whether I will be executed or not; but I already know that at least one of A and C will be executed. So if you tell me that one of them is slated for execution, you'll give me no information about whether* I *will be executed." The guard agrees and tells B that A will be executed tomorrow. How confident should B now be that they will be executed?*

w	$\mathbb{C}(\{w\})$	$\mathbb{C}(\{w\} \mid \neg A)$	$\mathbb{C}(\{w\} \mid \neg C)$
A	1/3	0	1/2
B	1/3	1/2	1/2
C	1/3	1/2	0

Before learning, there are three possibilities: either A was spared from execution, B was spared from execution, or C was spared from execution. Since each was equally likely to be spared, B has prior credences defined over the possibilities $\{A, B, C\}$, given by the probability distrribution in the margin.

There are two things B might learn: they might learn that A was not spared, $\neg A = \{B, C\}$, and they might learn that C was not spared, $\neg C = \{A, B\}$. Their prior conditional credence in B, given $\neg A$, is one half, and their prior conditional credence in B, given $\neg C$, is one half,

$$\mathbb{C}(B \mid \neg A) = \frac{1}{2} \qquad \mathbb{C}(B \mid \neg C) = \frac{1}{2}$$

So, following the norm of conditionalization, B will get more confident that they were spared *no matter what* they learn. But wait—how could this be? How could getting the guard to slip information in this way allow B to manufacture for themselves the guaranteed evidence that they are less likely to die?

This is known as the 'three prisoner's paradox'. It is closely related to the Monty Hall puzzle,

Example 14 (Monty Hall Puzzle)**.** *You're on the Monty Hall show. Before you are three doors. Behind one of the doors is a new car, and behind two of the doors is a goat. You choose door* 2*. At this point, Monty opens door* 1 *to reveal a goat and asks whether you want to change your mind and take the prize behind door* 3 *instead.* (*He always reveals a goat to the guests after their initial choice and gives them the option to switch in this way.*) *Should you switch*?

w	$\mathbb{C}(\{w\})$	$\mathbb{C}(\{w\} \mid \neg 1)$	$\mathbb{C}(\{w\} \mid \neg 3)$
1	1/3	0	1/2
2	1/3	1/2	1/2
3	1/3	1/2	0

In Monty Hall, after you make your initial choice, there is a one third chance that you've selected the prize. And you know that Monty will either reveal a goat behind door 1 or a goat behind door 3. If you condition on 'there's no car behind door 1' or 'there's no car behind door 3', then your posterior probability for the car being behind door 2 will be one half, no matter what Monty reveals.

The usual Bayesian line on both of these examples is that the analysis above has made a mistake. The mistake is that we've conditioned on *the wrong thing*. You conditioned on something that you learned, but you didn't condition on *everything* that you learned. The Bayesian is committed not just to conditioning on *something* that you learn; they are committed to conditioning on *everything* that you learn.

In the three prisoner's puzzle, B didn't *just* learn that A isn't spared; they *also* learned that the guard *told them this*. Let's use '$G\neg A$' for this proposition. If the guard had told B instead that C wasn't spared, then B would have additionally learned this stronger proposition—call it '$G\neg C$'. Before learning, B had the prior credence distribution in the margin.

	A	B	C
$G\neg A$	0	1/6	1/3
$G\neg C$	1/3	1/6	0

If they condition on $G\neg A$, then their probability that C was spared will rise to two thirds. If they condition on $G\neg C$, then their probability that A was spared will rise to two thirds. But either way, their probability that *they* are spared will remain fixed at one third.

The same thing happens in Monty Hall. You don't *just* learn that there's a goat behind door #1. You *additionally* learn that *Monty revealed* a goat behind door #1, $M\neg 1$. If we include this additional evidence, then your probability that there's a car behind door #2 will stay fixed at one third, and your probability that there's a car behind door #3 will

	1	2	3
$M\neg 1$	0	1/6	1/3
$M\neg 3$	1/3	1/6	0

rise to two thirds. So the answer to the puzzle is (counterintuitively?): yes, you should switch.

Notice a strange thing about our initial reasoning in examples 2 and 3: *no matter what* was learned, your credence in a given proposition was going to go up. In the Three Prisoner's Puzzle, this seemed like the wrong result—it seemed like prisoner B shouldn't be able to guarantee themselves confirmation that they'd been spared in this way. They shouldn't be able to reason to a foregone conclusion. Let's lay this down as a general principle:

No Guaranteed Confirmation If you might learn something that raises your credence in A, then there must be something else you might learn that would lower your credence in A.

Let's use $\mathscr{E}$ for the set of possible evidence you might receive. For instance, in example 1, $\mathscr{E} = \{O, E\} = \{\{1,3,5\},\{2,4,6\}\}$. And in example 2, we started out making the assumption that $\mathscr{E} = \{\neg A, \neg C\}$; but then our solution to the puzzle was that in fact $\mathscr{E} = \{G\neg A, G\neg C\}$. Then, what *No Guaranteed Confirmation* says is this: if there's some $E \in \mathscr{E}$ such that $\mathbb{C}_E(A) > \mathbb{C}(A)$, then there must be some $E \in \mathscr{E}$ such that $\mathbb{C}_E(A) < \mathbb{C}(A)$.

There's a specific reason that we ended up violating *No Guaranteed Confirmation* in the Three Prisoner's Puzzle. It was because we assumed that $\mathscr{E}$ was not a partition. If $\mathscr{E}$ is a finite partition (each cell of which has positive credence) and you update by conditionalization, then you will always satisfy *No Guaranteed Confirmation*.

The reasons we have for liking *No Guaranteed Confirmation* should carry over to prohibit even *expected* confirmation. So it looks like there's reason to favor the following principle:

No Expected Confirmation Your expectation of the degree to which any proposition is confirmed should be zero.

$$\mathbb{E}\left[\mathbb{C}_{\mathscr{E}}(A) - \mathbb{C}(A)\right] = 0$$

Notation: $\mathbb{C}_{\mathscr{E}}(A)$ is a definite description for 'the credence you plan to have in A after you learn whichever proposition in $\mathscr{E}$ you happen to learn'. In a world where you learn $E \in \mathscr{E}$, '$\mathbb{C}_{\mathscr{E}}(A)$' will refer to $\mathbb{C}_E(A)$.

If we assume that you know for sure what your current credence in A is, this will imply the following norm, which is known as the principle of *Reflection* (or, sometimes, the general principle of Reflection):

Reflection Your expectation of your planned posterior credence in any proposition should be equal to your prior credence in that proposition. That is: your plans should be such that, for any proposition A,

$$\mathbb{E}[\mathbb{C}_{\mathscr{E}}(A)] = \mathbb{C}(A)$$

Assuming that $\mathscr{E}$ is a finite partition (each cell of which has positive credence), Reflection follows from Plan Conditionalization.

6.2 *The Dutch Strategy Argument for Conditionalization*

Return to example 12. Suppose that you don't plan to update your credences by conditionalization. For instance, suppose that you plan

to have a credence of $1/4$ in H, if you learn O, and you plan to have a credence of $3/4$ in H, if you learn E.

$$\mathbb{C}_O(H) = 1/4 \qquad \mathbb{C}_E(H) = 3/4$$

So $\mathbb{C}_O(H) < \mathbb{C}(H \mid O)$ and $\mathbb{C}_E(H) > \mathbb{C}(H \mid E)$, in violation of the norm of conditionalization.

Then, here's a strategy we could enact that will surely leave you poorer, no matter what. Before you learn anything at all, you will be happy to buy the following conditional bet:

Bet #1

\$1	if HO is true
\$0	else

price: $\begin{cases} \$1/3 & \text{if } O \\ 0 & \text{if } E \end{cases}$

Here, we are assuming (as always with these 'Dutch book' arguments) that you value dollars linearly, and that your 'fair price' for a conditional bet like this is given by your conditional credence.

Similarly, before you learn anything at all, you will be happy to *sell* the following conditional bet:

Bet #2

\$1	if HE is true
\$0	else

price: $\begin{cases} \$2/3 & \text{if } E \\ 0 & \text{if } O \end{cases}$

Now, if you learn O, bet 2 will be called off, and only bet #1 will still be relevant. At that stage, your new credence in H will be $1/4$, so you will be happy to *sell* bet #3:

Bet #3

\$1	if H is true
\$0	else

price: \$1/4

But now, the combination of bets #1 and #3 add up to a guaranteed loss:

	H	$\neg H$
Net profit from buying bet 1	\$2/3	−\$1/3
Net profit from selling bet 3	−\$3/4	\$1/4
Overall net profit	−\$1/12	−\$1/12

On the other hand, if you learn E, bet 1 will be called off, and only bet #2 will still be relevant. At that stage, your new credence in H will be $3/4$, so you will be happy to *buy* bet #4:

Bet #4

\$1	if H is true
\$0	else

price: \$3/4

But now, the combination of bets #2 and #4 add up to a guaranteed loss:

	H	$\neg H$
Net profit from selling bet 2	−\$1/3	\$2/3
Net profit from buying bet 4	\$1/4	−\$3/4
Overall net profit	−\$1/12	−\$1/12

No matter what happens, you're going to lose a twelfth of a dollar. So there's a so-called *Dutch Strategy* against you.

Dutch Strategy a *Dutch Strategy* is a contingency plan for buying and selling bets which is guaranteed to leave you poorer no matter what.

Moreover, something similar will happen any time you have determinate credence revision plans which you'll certainly follow through on which do not abide by conditionalization.

Dutch Strategy Theorem If you stand to learn one of a partition of propositions and you follow determinate plan to revise your credences which is not the conditionalization plan, then there will be a strategy for trading bets with you which is guaranteed to lose you money no matter what. And, if you plan to revise your credences by conditioning on the member of the partition that you learn, then there is no such strategy.

The Dutch Strategy Theorem was first shown (to my knowledge) by David Lewis, in a handout that was reported by Paul Teller, and published much later by Lewis himself. The converse result (that conditionalization is *not* susceptible to a Dutch strategy) was proven by Brian Skyrms.

A comprehension test: suppose that, in example 1, you learn that the die landed even and condition on this. Before learning, I buy off of you a \$1 bet on H for fifty cents.

\$1	if H is true	price: \$1/2
\$0	else	

After you learn that the die landed even, I sell you a \$1 bet on H for the price of \$2/3.

\$1	if H is true	price: \$2/3
\$0	else	

But the combination of these two bets is guaranteed to lose you \$1/6 no matter what.

Is this a Dutch strategy against the conditionalizer? (If so, is it a counterexample to the theorem?) Why or why not?

This affords us the following argument for Plan Conditionalization:

P1) If your credence revision plans are rational, then they will not be susceptible to a Dutch strategy.

P2) *Dutch Strategy Theorem*

∴C) If your credence revision plans are rational, then you will plan to conditionalize on what you learn.

Why shouldn't your credence revision plans be susceptible to a Dutch strategy? We could give a purely pragmatic argument for this—but, as in the case of the Dutch book argument for probabilism, we might instead want to give a 'depragmatized' version of the argument. We might say that, by adopting these plans for revising your credences, you have *committed* yourself to certain contingency plans for the buying and selling of bets. You have endorsed these contingency plans. But, by your own lights, these contingency plans are sure losers. So, under two different modes of description that you can recognize are equivalent, you evaluate these contingency plans as both good and bad. So, on the 'depragmatized' understanding, susceptibility to a Dutch strategy reveals an underlying evaluate inconsistency.

Let's grant the first premise; and let's trust the Dutch Strategy theorem (which is true). Even granting this, the argument's conclusion is weaker than the full norm of conditionalization. What we've shown in this argument is only that your credence revision *plans* should abide by the norm of conditionalization. So we've only justified what we earlier called *Plan Conditionalization*. We haven't additionally justified what I earlier called *Conditionalization*. To justify that additional assumption, we might attempt to run an argument like this:

P1) If you are rational, then you will not end up actually buying/selling a collection of bets which are guaranteed to lose you money no matter what

P2) *Dutch Strategy Theorem*

∴C) If you are rational, then you will actually condition on whatever you learn

But this argument is both invalid (because the conclusion doesn't follow from the premises) and it has a premise that the defender of conditionalization themselves should reject. To appreciate both of these points, go back to the comprehension test in the margin above: Being susceptible to a Dutch strategy is a *very importantly* different notion from having two of your time slices *actually buy* a Dutch book. A conditionalizer can end up buying into a Dutch book. *That's* not the thing that's meant to be bad about being susceptible to a Dutch strategy. What's meant to be bad about being susceptible to a Dutch strategy is that you could be sold a Dutch book *no matter what you learn*. In the comprehension test, even though the conditionalizer could be sold a Dutch book if they learnt that the die landed even, they could *not* have been similarly ensnared in a Dutch book if they had instead learnt that the die landed odd.

This is a problem for any attempt Conditionalization by separating it out into the two theses we considered above (*Plan Conditionalization* and *Honor Your Plans*), and then separately justifying each with some kind of Dutch strategy argument. The problem is that there's no Dutch strategy argument for honoring your plans. (We'll see a similar problem recur when we consider the accuracy-based arguments for conditionalization.)

There are further problems, even if we restrict ourselves to the justification of *Plan Conditionalization*. For the argument we gave is at best enthymematic. For the argument to be valid, we must make the following additional assumptions:

P3) If you are rational, you will have definite plans for how to revise your credences if you learn E, for each $E \in \mathscr{E}$.

P4) The set of propositions you might learn, $\mathscr{E}$, will always form a partition.

To see why the first assumption is important, suppose that, in example 3, whether you learn odd or even, you'll flip a coin. If the coin lands heads, then your posterior credence in H will be one third; but if the coin lands tails, then your posterior credence in H will be two thirds. This will scotch the Dutch strategy we considered above (comprehension check: why?). And, in general, you won't be susceptible to a Dutch strategy if you adopt an indeterministic credence revision plan like this. But this indeterministic credence revision plan is not the conditionaliztion plan. So we need some reason to rule out this kind of plan.

To see why the second assumption is needed, notice that the *Dutch Strategy Theorem* depends upon the assumption that $\mathscr{E}$ forms a partition. Go back to the Three Prisoner's Puzzle (example 2), and suppose that in fact prisoner B might either learn $\neg A$ or $\neg C$ (and no more). Then, $\mathscr{E}$ is the non-partition $\{\neg A, \neg C\}$. If prisoner B adopts the conditionalization plan, then before learning, they will happily sell a \$1 bet on them being spared for the price of \$1/3,

Bet # 5

\$1	if B is true
\$0	else

price: \$1/3

And after they learn from the guard (no matter what they learn from the guard), they will buy the very same bet back at the price of \$1/2,

Bet #6

\$1	if B is true
\$0	else

price: \$1/2

But the combination of these two bets is guaranteed to lose them a 1/6th of a dollar *no matter what*.

	B	$\neg B$
Net profit from selling bet 5	−\$2/3	\$1/3
Net profit from buying bet 6	\$1/2	−\$1/2
Overall net profit	−\$1/6	−\$1/6

So, in situations where $\mathscr{E}$ does not form a partition, the plan to conditionalize on what you learn is susceptible to a Dutch strategy.

6.3 *The Accuracy Arguments for Conditionalization*

Return to example 11: why is Melissa more rational than Daniel? The Dutch Strategy argument points out that Daniel opens himself up to a Dutch strategy—but this justification looks overly pragmatic. Is there any purely *epistemic* or *accuracy-based* reason why Daniel's credence-revision dispositions should be irrational? One very natural suggestion is to say something about Daniel's relationship to the *truth*. In this section, we'll consider three different accuracy arguments that Daniel's credence-revision dispositions are less rational than Melissa's.

As with probabilism, these arguments are all centered around the idea that *accuracy* is the sole epistemic good. As in the case of probabilism, these arguments take for granted a conception of the epistemic good—accuracy—together with a kind of epistemic consequentialism. Let's make these assumptions explicit:

Veritism The only epistemic value is accuracy.

Epistemic Consequentialism To be epistemically rational is to pursue epistemic value in an instrumentally rational way.

According to the epistemic consequentialist, we can think of ourselves as facing a hypothetical choice between different doxastic states. Some of these doxastic states will be better than others, though we won't always know which are better than which others. Nonetheless, we should have a doxastic state that it would be instrumentally rational to choose in this hypothetical decision, if our only final goal was epistemic goodness. In our accuracy-based justification of probabilism, we appealed to a simple norm of instrumental rationality, known as *dominance*:

Dominance If one available choice will be better than another, no matter what the world ends up being like, then the second choice is irrational.

6.3.1 *Accuracy Dominance Argument for Conditionalization*

Note that, in example 11, there's no *guarantee* that Daniel's credences will be less accurate than Melissa's. For it could easily turn out that the Democrats *really are* more moral than the Republicans, and that all and only the Republican scandals *really were* morally wrong. In that case, Daniel's methods of forming beliefs would have led him to all and only true beliefs about American political scandals.

So, at first glance, it doesn't look like we can give an argument that Daniel is *accuracy-dominated*. Nonetheless, R.A. Briggs & Richard Pettigrew show that there's *some* sense in which Daniel is accuracy-dominated by Melissa. They suppose that we have a measure of inaccuracy satisfying the properties given in the margin, and they moreover assume that we can measure the accuracy of an updating *plan* by adding together the accuracy of the *prior* function $\mathbb{C}$ together with the accuracy of the *posterior* function.

More carefully, we suppose that there is some evidence partition, $\mathscr{E} = \{E_1, E_2, \ldots, E_N\}$, such that one of the E_is in $\mathscr{E}$ will be the strongest thing you learn. And you have some (definite) plan for responding to each piece of evidence you might receive, $\mathbb{C}_{\mathscr{E}}$. Then, Briggs & Pettigrew are interested *not* in the accuracy of your posterior credences on their own, but instead they are interested in the total accuracy of *both* your prior credences *and* your posterior credences, $\langle \mathbb{C}, \mathbb{C}_{\mathscr{E}} \rangle$. And they say that you can measure the accuracy of this pair by just adding together the accuracy of each credence function in the pair:

$$\mathcal{I}\left(\langle \mathbb{C}, \mathbb{C}_{\mathscr{E}} \rangle, w\right) = \mathcal{I}\left(\mathbb{C}, w\right) + \mathcal{I}\left(\mathbb{C}_{\mathscr{E}}, w\right)$$

(Note that $\mathbb{C}_{\mathscr{E}}$ will be a different probability function depending upon what is learnt. When we ask about the inaccuracy of $\mathbb{C}_{\mathscr{E}}$ at w, we are really asking about the inaccuracy of $\mathbb{C}_E$ at w, where E is the cell of the partition $\mathscr{E}$ which contains w.)

If $\mathbb{C}$ is a probability function and $\mathbb{C}_{\mathscr{E}}$ is the conditionalizing plan, then say that the pair $\langle \mathbb{C}, \mathbb{C}_{\mathscr{E}} \rangle$ is a *probabilistic conditionalizing plan*. Briggs & Pettigrew prove the following theorem:

Non-Conditionalization is Accuracy-Dominated (Briggs & Pettigrew) If $\langle \mathbb{C}, \mathbb{C}_{\mathscr{E}} \rangle$ is not a probabilistic conditionalizing plan, then there is some other credal plan which accuracy dominates it. And, if $\langle \mathbb{C}, \mathbb{C}_{\mathscr{E}} \rangle$ is a probabilistic conditionalizing plan, then there is no other credal plan which accuracy dominates it.

This theorem affords us the following argument for conditionalization:

P1) Having an accuracy-dominated credal plan is epistemically irrational

P2) Inaccuracy is measured with a separable, continuous, extensional, and strictly proper function $\mathcal{I}$.

See Briggs, R. A. & Pettigrew, Richard (2020). "An Accuracy-Dominance Argument for Conditionalization". *Noûs* 54 (1): 162-181.

Briggs & Pettigrew assume that the measure of inaccuracy, $\mathcal{I}$, satisfies the following constraints:

Seperability The inaccuracy of a credence function, $\mathbb{C}$, at a world, w, is the sum of the inaccuracy of $\mathbb{C}$'s credence in A at w, for each A to which $\mathbb{C}$ assigns a credence. That is, there's some *local* measure of the inaccuracy of a credence x in the proposition A at the world w, $\mathcal{I}(x, A, w)$, such that

$$\mathcal{I}\left(\mathbb{C}, w\right) = \sum_{A \in \mathcal{A}} \mathcal{I}\left(\mathbb{C}(A), A, w\right)$$

Continuity Small changes in your credence in A don't lead to big changes in the inaccuracy of your credence in A, holding everything else fixed. That is: $\mathcal{I}(x, A, w)$ is a continuous function of x.

Extensionality The only feature of the world that's relevant to the inaccuracy of a credence x in A at w is the truth-value of A at w. That is: if $\mathbf{1}_A(w) = \mathbf{1}_B(w')$, then $\mathcal{I}(x, A, w) = \mathcal{I}(x, B, w)$

Strict Propriety Every probability function expects itself to be strictly more accurate than any other credence function. That is, for every probability function $\mathbb{P}$, and any credence function $\mathbb{C} \neq \mathbb{P}$,

$$\sum_{w \in \mathcal{W}} \mathbb{P}(w) \cdot \mathcal{I}\left(\mathbb{P}, w\right) < \sum_{w \in \mathcal{W}} \mathbb{P}(w) \cdot \mathcal{I}\left(\mathbb{C}, w\right)$$

P3) Non-probabilistic and non-conditioning credal plans are accuracy-dominated on any separable, continuous, extensional, and strictly proper measure.

∴ C) Having a non-probabilistic or non-conditioning credal plan is epistemically irrational.

This is a beautiful argument; but what it shows is something less than what we might have wanted. As Briggs & Pettigrew themselves acknowledge, it only establishes what they call the *wide scope* norm of conditionalization:

> *Conditionalization (Wide-Scope)* Suppose that, between t_0 and t_1, you will learn the true cell of the partition $\mathscr{E}$ (and no more). Then, it is rationally required that: your have a probabilistic prior $\mathbb{C}$ and an update plan $\mathbb{C}_{\mathscr{E}}$ such that $\mathbb{C}_{\mathscr{E}}$ is the plan to conditionalize on the true member of $\mathscr{E}$.

It does *not* give an argument for what they call the *narrow scope* norm of conditionalization:

> *Conditionalization (Narrow-Scope)* Suppose that, between t_0 and t_1, you will learn the true cell of the partition $\mathscr{E}$ (and no more), and suppose that your prior credence function is the probabilistic $\mathbb{C}$. Then, it is rationally required that you have an update plan $\mathbb{C}_{\mathscr{E}}$ which is the plan to conditionalize on the true member of $\mathscr{E}$.

Why doesn't this follow? Because the sunk-cost fallacy is a fallacy. Suppose that Danielle drunkenly and foolishly buys a \$1 bet on the slowest horse in the race for 50¢. The next day, she comes to her senses and tries to sell it back, but the market is only accepting it for 40¢. It can be perfectly rational for Danielle to sell the bet at this price, given that she's rationally very confident that the slow horse is going to lose. But notice that the *plan* to buy the bet for 50¢ and then sell it back for 40¢ is dominated. True enough, but that shouldn't stop Danielle from selling the bet today. Tomorrow's choices are in the past and beyond her control. All she can control now is what she does today; and today, selling is the best choice.

Daniel's position might be like Danielle's. Daniel finds himself with his prior credences—those are already set, and now beyond his control. His only choice now is which update plans to adopt. Now, maybe he's made some bad choices with his priors, but just like Danielle's choices yesterday shouldn't figure into her decision about whether to sell the bet today, Daniel's past choices shouldn't figure into his deliberation about which update plans to adopt today. Since $\mathbb{C}_{\mathscr{E}}$ is the only thing under his control, he should only be concerned with the accuracy of that plan. But since his motivated reasoning plan could easily end up making him more accurate *tomorrow* than Melissa's conditionalizing plan makes her tomorrow, it's not clear that the Briggs & Pettigrew result gives Daniel a reason to condition.

6.3.2 *An Expected Accuracy Maximization Argument for Conditionalization*

A different accuracy-based argument for conditionalization can be found in the work of Hannes Leitgeb and Richard Pettigrew. Their result can be used to argue for a genuinely *narrow-scope* version of conditionalization, but it relies on some additional decision-theoretic and evidential assumptions.

See Leitgeb, Hannes & Pettigrew, Richard (2010). "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy". *Philosophy of Science* 77 (2): 236-272. As I discuss in my *Learning and Value Change*, Leitgeb and Pettigrew actually give two independent strategies for justifying conditionalization. Here, I'm only discussing the second.

They assume that you start out with some prior (probabilistic) credence function, $\mathbb{C}$, and then you acquire the evidence, E. They assume that acquiring this evidence changes which possibilities are epistemically possible for you. Whereas before, all of the worlds in $\mathcal{W}$ were epistemically possible, now, only the worlds in E are epistemically possible for you. This change in your credal state might prompt you to exchange your credence function for another. How should you choose? They suggest—drawing on a long tradition from decision theory—that you should select whichever one has the greatest *expected* epistemic value. That is: they are drawing on the following decision-theoretic norm:

Maximize Expected Value When choosing from amongst a collection of options, you should choose one with the greatest expected value.

Because the prior $\mathbb{C}$ was probabilistic, and because accuracy is measured in a strictly proper way, the prior $\mathbb{C}$ maximized expected value (when the expectation is taken relative to itself). That's just what it means for the measure of accuracy to be strictly proper. But things have changed—not all the worlds in $\mathcal{W}$ are still live possibilities. So Leitgeb and Pettigrew now suggest that you should choose a credence function, $\mathbb{C}^+$, which minimizes *this* weighted sum:

$$\sum_{w \in E} \mathbb{C}(w) \cdot \mathcal{I}(\mathbb{C}^+, w)$$

And, so long as $\mathcal{I}$ is a strictly proper measure of (in)accuracy, the unique choice of $\mathbb{C}^+$ which minimizes this weighted sum will be $\mathbb{C}(- \mid E)$.[1]

So Leitgeb & Pettigrew afford us the following defense of conditionalization:

P1) When you learn E, you should adopt a posterior credence function which minimizes expected inaccuracy within those worlds compatible with the evidence.

P2) Inaccuracy is measured with a strictly proper measure $\mathcal{I}$.

P3) If $\mathcal{I}$ is strictly proper, then $\mathbb{C}(- \mid E)$ uniquely minimizes expected inaccuracy within the worlds compatible with E.

$\therefore$ C) When you learn E, you should adopt your prior credence function conditioned on your evidence.

According to this justification of conditionalization, Daniel is irrational because (roughly), once the evidence of the scandal is in, he should expect his motivated reasoning to take him further from the truth than Melissa's non-motivated reasoning.

[1] *Proof*: $\mathbb{C}(- \mid E)$ is a probability function. Since $\mathcal{I}$ is strictly proper, the unique choice of $\mathbb{C}^+$ which minimizes

$$\sum_{w \in \mathcal{W}} \mathbb{C}(w \mid E) \cdot \mathcal{I}(\mathbb{C}^+, w) = \sum_{w \in E} \mathbb{C}(w \mid E) \cdot \mathcal{I}(\mathbb{C}^+, w)$$

will be $\mathbb{C}^+(- \mid E)$. (The expectation on the top is equal to the weighted sum on the bottom since $\mathbb{C}(w \mid E) = 0$ whenever $w \notin E$.) If a choice of $\mathbb{C}^+$ minimizes this function, it will also minimize the function after we multiply it by a positive constant. So it will also minimize

$$\mathbb{C}(E) \cdot \sum_{w \in E} \mathbb{C}(w \mid E) \cdot \mathcal{I}(\mathbb{C}^+, w) = \sum_{w \in E} \mathbb{C}(w \mid E) \cdot \mathbb{C}(E) \cdot \mathcal{I}(\mathbb{C}^+, w) = \sum_{w \in E} \mathbb{C}(w) \cdot \mathcal{I}(\mathbb{C}^+, w)$$

(P2) is just a theorem, so any objection will have to be to the normative claim (P1) or the contention that accuracy is strictly proper, (P3). We discussed reasons to think that inaccuracy is strictly proper in last class. At first glance, you might suspect that (P1) is just a relatively uncontroversial application of the norm *Maximize Expected Value*. But this isn't quite right. We need to carefully distinguish these two functions:

$$\sum_{w \in W} \mathbb{C}(w) \cdot \mathcal{I}(\mathbb{C}^+, w) \qquad \sum_{w \in E} \mathbb{C}(w) \cdot \mathcal{I}(\mathbb{C}^+, w)$$

The thing on the *left* is the expected inaccuracy of $\mathbb{C}^+$. And we already know what (uniquely) minimizes this—it is $\mathbb{C}$ itself. That's just what it means for $\mathcal{I}$ to be a strictly proper measure of inaccuracy. The norm of *Maximize Expected Value*, together with the assumption of Veritism, tells us that we should try to minimize this function; so it is what tells us that it's irrational to move from $\mathbb{C}$ to any other credence function. But the thing on the *right* is not the expected inaccuracy of $\mathbb{C}^+$. That's so because the function $f[V] = \sum_{w \in E} \mathbb{C}(w) \cdot V(w)$ does not satisfy the properties of an expectation—recall, an expectation should satisfy $\mathbb{E}[c] = c$, if c is a constant. But $f[c] = \mathbb{C}(E) \cdot c$, which is going to be less than c. So this weighted sum on the right isn't an expectation of epistemic value, because it's not an *expectation*.

The justification offered by Leitgeb and Pettigrew points us towards a puzzle about any attempt to justify a norm like conditionalization using considerations of *expected* accuracy: we already know, from the fact that $\mathcal{I}$ is strictly proper, that your prior maximizes expected accuracy. So we know that the conditionalization posterior $\mathbb{C}(- \mid E)$ has a lower expected accuracy than $\mathbb{C}$ itself. So how could maximizing expected accuracy every give us any reason to exchange the prior for the conditionalization posterior? Won't this mean just exchanging a doxastic state with higher expected accuracy for one with lower expected accuracy?

6.3.3 *Another Expected Accuracy Maximization Argument*

No—but the reason is somewhat subtle. Strict propriety of $\mathcal{I}$ means that $\mathbb{C}$ will have a greater expected accuracy than any *fixed* probability function. But that doesn't mean that it will have a greater expected accuracy than any *definite description* for a probability function. Recall: a definite description for a probability function, $\mathcal{P}$, is a function from worlds to probability functions. The interpretation is that the value this function takes on at a world, w, is the probability function which '$\mathcal{P}$' refers to at the world w.

Consider, then, the *omniscient* credence function, $\mathcal{O}$. This is a function from a world w to the probability function which is certain of $\{w\}$. The expected inaccuracy of $\mathcal{O}$ is *zero*,[2] which is certainly less than the expected inaccuracy of $\mathbb{C}$ unless $\mathbb{C}$ happens to be certain about what world is actual.

Now think about $\mathbb{C}_{\mathscr{E}}$—your update plans for how to respond when

[2] Why? Because the probabilities which $\mathcal{O}_w$ gives to every proposition A is just A's truth-value at w, so the distance between $\mathcal{O}_w(A)$ and $\mathbf{1}_A(w)$ will be zero, for every A. And the weighted sum of any number of zeros is zero.

you learn the true member of the partition $\mathscr{E}$. This is not any particular probability function. Instead, it is a *definite description* for a probability function. So strict propriety isn't going to say that the expected inaccuracy of $\mathbb{C}_{\mathscr{E}}$ must be greater than the expected inaccuracy of $\mathbb{C}$.

An argument from Hilary Greaves and David Wallace justifies conditionalization by showing that the conditioning plan is the one that minimizes expected inaccuracy. But any such argument must be carefully formulated—after all, we know already that $\mathcal{O}$—the plan to become certain of the truth—will minimize expected accuracy. In particular, it will have lower expected accuracy than the conditioning plan. So Greaves & Wallace want to limit the scope of the kinds of update plans they are going to consider. They say that they are only interested in *available* update plans.

See Greaves, Hilary & Wallace, David (2006). "Justifying conditionalization: Conditionalization maximizes expected epistemic utility". *Mind* 115 (459): 607-632.

Just to formalizesome of the things we've been talking about informally, let's say that an update plan, $\mathcal{U}$, is a function from worlds $w \in \mathcal{W}$ to probability functions over the relevant algebra of propositions, $\mathcal{A}$.

Update Plans An *update plan*, $\mathcal{U}$, is a function from worlds in $\mathcal{W}$ to probability functions over $\mathcal{A}$. $\mathcal{U}_w$ is the posterior probability function the update plan prescribes adopting in the world w

On this definition, $\mathcal{U}$ is an update plan. It is just the plan to become omniscient after learning. But Greaves & Wallace contend that this update plan is not *available*.

Available Update Plans If you are going to learn the true member of the (finite) partition $\mathscr{E} = \{E_1, E_2, \ldots, E_N\}$, then an update plan $\mathcal{U}$ is *available* iff, for any $E \in \mathscr{E}$, and any two worlds $w, w^* \in E$, $\mathcal{U}_w = \mathcal{U}(w^*)$.

In other words: an update plan is *available* iff it only distinguishes between worlds in which you have different evidence. Worlds in which you have the same evidence are not worlds in which you can plan to update differently.

Then, Greaves & Wallace prove the following theorem:

Conditionalization Maximizes Expected Accuracy (Greaves & Wallace) If inaccuracy is measured with a strictly proper measure, and you stand to learn the true member of the (finite) partition $\mathscr{E} = \{E_1, E_2, \ldots, E_N\}$ then the available update plan which minimizes expected accuracy is the conditionalization plan.

(The proof is in the margin.)

This affords us the following justification of conditionalization:

P1) When you're learning which member of a partition is true, you should choose an available update plan that minimizes expected inaccuracy.

P2) Inaccuracy is measured with a strictly proper function.

P3) *Conditionalization Maximizes Expected Accuracy*

∴C) When learning which member of a partition is true, you should plan to condition on whatever you learn.

A proof of Greaves and Wallace's theorem: The expected inaccuracy of an update plan $\mathcal{U}$ is

$$\sum_{w \in \mathcal{W}} \mathbb{C}(w) \cdot \mathcal{I}(\mathcal{U}_w, w)$$
$$= \sum_{E \in \mathscr{E}} \sum_{w \in E} \mathbb{C}(w) \cdot \mathcal{I}(\mathcal{U}_w, w)$$

Since $\mathcal{U}$ is available, $\mathcal{U}_w = \mathcal{U}_{w^*}$ for each $w, w^* \in E$ and each $E \in \mathscr{E}$, so we can write '$\mathcal{U}_E$' for the posterior recommended for any world in E, and the above becomes

$$= \sum_{E \in \mathscr{E}} \sum_{w \in E} \mathbb{C}(w \mid E) \cdot \mathbb{C}(E) \cdot \mathcal{I}(\mathcal{U}_E, w)$$
$$= \sum_{E \in \mathscr{E}} \mathbb{C}(E) \sum_{w \in E} \mathbb{C}(w \mid E) \cdot \mathcal{I}(\mathcal{U}_E, w)$$

Our choice of $\mathcal{U}_E$ is independent of our choice of $\mathcal{U}_{E'}$, for any $E \neq E'$. So when we minimize the weighted sum above, we must for each $E \in \mathscr{E}$ make whatever choice minimizes the inner sum $\sum_{w \in E} \mathbb{C}(w \mid E) \cdot \mathcal{I}(\mathcal{U}_E, w)$. But since $\mathbb{C}(- \mid E)$ is a probability function and $\mathcal{I}$ is strictly proper, we know that the choice $\mathcal{U}_E = \mathbb{C}(- \mid E)$ uniquely minimizes this inner sum. And the same goes for every $E \in \mathscr{E}$. So the conditionalization plan will uniquely minimize expected inaccuracy.

Later, we're going to consider reasons to worry about what happens if and when you are learning something *other* than which cell of a partition is true. But for now, let's consider some reasons to worry about the underlying epistemic consequentialism which all of these arguments are presupposing.

6.4 Objections to Epistemic Consequentialism

6.4.1 A Digression on Instrumental Rationality

Let's start off by acknowledging a fact that we've been ignoring up to this point: norms like *Dominance* and *Maximize Expected Value* cannot be correct in general. Consider the application of those norms to the following decision:

Shakedown Crazy Joe Gallo tells you that you have a nice store, mentions that it would be terrible if anything happens to it, and offers you the mafia's "protection". The protection costs $100 a month.

We can suppose that all you care about is how much money you have at the end of the month. If your store is vandalized, you'll lose $1000. If you pay the protection money, you'll lose $100. Consider the decision table in the margin.

	Store Vandalized	Store not Vandalized
Pay	−$1100	−$100
Don't	−$1000	$0

There are two possible states of the world: either the store will be vandalized or it won't. And there are two available acts: either you pay the protection fee or you don't. Notice that, *no matter whether* the store is vandalized or not, paying the protection fee will leave you $100 poorer than not paying. So a naïve application of dominance would tell you that you shouldn't pay. Likewise, suppose that your probability that your store will be vandalized is x. Then, the expected value of paying will be less than the expected value of not paying,

$$\begin{aligned} x \cdot (-1100) + (1-x) \cdot (-100) &\overset{?}{<} x \cdot (-1000) + (1-x) \cdot 0 \\ -1100x + 100x - 100 &\overset{?}{<} -1000x \\ -100 &\overset{\checkmark}{<} 0 \end{aligned}$$

But obviously, in this decision, you should pay—if you don't pay, Gallo is going to vandalize your store, and if you pay, he won't! The trouble here is that there is *act-state dependence*. Which state (store vandalized or not) obtains depends upon which act you choose. The norms *Dominance* and *Maximize Expected Value* work well when there's no act-state dependence, but when there is act-state dependence, they need to be modified.

It turns out that there's controversy about how they should be modified. Some people think that the kind of dependence that matters is *probabilistic* dependence; whereas others think that the kind of dependence that's relevant is *causal* dependence. The classic way of bringing out the difference is with this decision from Nozick:

Newcomb's Problem Before you are two boxes: one transparent and one opaque (the 'mystery box'). You are leaving with the mystery

box and its contents no matter what—it's a gift. Your only choice is whether or not to take or leave behind the transparent box. Inside the transparent box is $1000. Yesterday, I made a quite reliable prediction about what you would choose. If I predicted that you would take just the mystery box ('one box'), then I put $1,000,000 inside the mystery box. And if I predicted that you would take both the mystery box and the transparent box ('two box'), then I left the mystery box empty.

There is considerable controversy about how to react to this decision. Some think that, because one-boxing raises the probability that there's $1,000,000 in the mystery box, you should one box. Others think that, because two-boxing will certainly cause you to get $1000 more than one-boxing would, you should two-box.

Two dominant decision theories have been built up around these two different intuitions. According to the first, *evidential decision theory*, we should generalize *Dominance* and *Maximize Expected Value* in this way:

Evidential Dominance If $\mathcal{S}$ is a partition of states which are probabilistically independent of how you choose, and one option, A, is better than another, B, in every state in $\mathcal{S}$—that is, if the value of AS is greater than the value of BS, $\mathcal{V}(AS) > \mathcal{V}(BS)$, for every $S \in \mathcal{S}$—then B is irrational.

Maximize Evidential Expected Value A is a more rational choice than B whenever

$$\sum_{w \in \mathcal{W}} \mathbb{C}(w \mid A) \cdot \mathcal{V}(w) \quad > \quad \sum_{w \in \mathcal{W}} \mathbb{C}(w \mid B) \cdot \mathcal{V}(w)$$

According to the evidentialists, you should one-box because one-boxing maximizes expected *evidential* value.

According to the second dominant decision theory, *causal decision theory*, we should generalize *Dominance* and *Maximize Expected Value* in this way:

Causal Dominance If $\mathcal{K}$ is a partition of states which are causally independent of how you choose, and one option, A, is better than another, B, in every state in $\mathcal{K}$—that is, if the value of AK is greater than the value of BS, $\mathcal{V}(AK) > \mathcal{V}(BK)$, for every $K \in \mathcal{K}$—then B is irrational.

Maximize Causal Expected Value A is a more rational choice than B whenever

$$\sum_{w \in \mathcal{W}} \mathbb{C}(A \boxdot\!\!\rightarrow w) \cdot \mathcal{V}(w) \quad > \quad \sum_{w \in \mathcal{W}} \mathbb{C}(B \boxdot\!\!\rightarrow w) \cdot \mathcal{V}(w)$$

The definition of *Maximize Causal Expected Value* in the body presupposes that, for any antecedent A, there is some world w such that $A \boxdot\!\!\rightarrow w$ is true. If you deny this, then the definition will have to be generalized. I won't bother with the details here.

One fact that will be helpful in the next subsection: Skyrms' thesis says that $\mathbb{C}(A \boxdot\!\!\rightarrow w)$ should be equal to $\mathbb{E}[Ch(w \mid A)]$. If we accept Skyrms' thesis, then we can say that causal expected value is your expectation of *chance*'s expectation of $\mathcal{V}$, conditional on A. And we can say that the evidential expected value is your conditional expectation of *chance*'s expectation of $\mathcal{V}$.

6.4.2 *Epistemic Consequentialism and Act-State Dependence*

Hilary Greaves raises some interesting and troubling cases for the the-

See Greaves, Hilary (2013). "Epistemic Decision Theory". *Mind* 122 (488): 915-952.

sis of epistemic consequentialism. All of these cases are ones that involve act-state dependence, in the sense that how accurate your credences are will depend upon which plans for updating your credences you adopt.

Promotion You are up for promotion, but your boss is insecure and will only promote you if you lack confidence. If your credence in "I will get the promotion" is x, then the objective chance that you'll get the promotion will be $1 - x$.

In *Promotion*, both evidential and causal forms of epistemic consequentialism recommend $x = 1/2$.

Leap You stand on one side of a chasm, about to leap. You are forming a credence in the proposition "I will make it to the other side". Confidence makes you more likely to succeed, so that, if your credence in this proposition is x, the objective chance that you make it to the other side is x.

In *Leap*, both evidential and causal forms of epistemic consequentialism recommend $x = 1$ or 0.

Embezzlement You have conclusive evidence that your colleague Charlie has embezzled funds, and you have access to n files attesting to the propositions $A_1, A_2, \ldots, A_n$. Yesterday, Charlie made a very reliable prediction about what credence you would have in 'Charlie embezzled funds'. If he predicted a credence of x, then with probability x he altered the contents of the files (randomizing them so that their conclusions are as likely to be true as false). If he predicted a credence of x, then the objective chance of a proposition in the file being true is $1 - x/2$.

In *Embezzlement*, for $n > 4$, evidential epistemic consequentialism recommends that you be certain Charlie is not guilty and certain in the propositions $A_1, \ldots, A_n$. And causal epistemic consequentialism recommends that you are certain that Charlie is guilty and have a credence of one half in each of the propositions $A_1, \ldots, A_n$.

Imps You are walking through the garden of epistemic imps. Before you is a child. In a nearby summerhouse are n invisible imps who are able to read your mind. If your credence that there's a child before you is x, then they will come out to play with a chance of $1 - x/2$.

In *Imps*, for $n > 4$, both the evidential and the causal versions of epistemic consequentialism recommend that you have a credence of $x = 0$ that there is a child in front of you.

According to Greaves, the correct answers for these cases are as follows: In *Promotion*, you should have a credence of $1/2$ that you'll get the promotion. In *Leap*, you should have a credence of either 0 or 1 that you'll make it to the other side. In *Embezzlement*, you should have a credence of 1 that Charlie is guilty and a credence of $1/2$ in each of the propositions $A_1 \ldots A_n$. And, in *Imps*, you should have a credence of 1 that there's a child before you and a credence of $1/2$ that each of the imps has come out to play.

The trouble is that there's no version of epistemic consequentialism which delivers these verdicts. The evidential version of the theory gets both *Embezzlement* and *Imps* wrong. And the causal version of the theory gets *Imps* wrong. It says that you should be willing to pay an *epistemic bribe*, sacrificing accuracy with a known proposition to raise your accuracy with other, unrelated propositions. Insofar as you think it's irrational to pay this kind of epistemic bribe, you might want to reject the underlying epistemic consequentialism and (therefore) reject the justifications of conditionalization we offered above.

Jason Konek and Ben Levinstein have offered a response to these worries. According to them, we should be causal decision theorists

See Konek, Jason & Levinstein, Ben (2019). "The Foundations of Epistemic Decision Theory". *Mind* 128 (509): 69-107.

when it comes to *action*, but not when it comes to *belief*.[3] The reason has to do with the differences between epistemic and practical value. According to them, there wasn't really anything wrong with the basic imperative to maximize expected value—instead, the problem lay with the *kind* of value whose expectation was being maximized. Epistemic value has a mind-to-world direction of fit, whereas practical value has a world-to-mind direction of fit. In the case of practical decision theory, the value of an act lies in what that act *would do* to the world But in the case of epistemic decision theory, the value of a credence is *not* what the credence would do to the world, but rather how well the credence would reflect what's actually the case.

[3] At least, Levinstein is willing to grant this for the purposes of this argument that we should be causal decision theorists. In other work he defends a different decision theory.

So they endorse the following general theory of instrumental rationality:

Maximize Expected Value A is a more rational option than B iff A's expected value is greater than B's expected value

Practical Value The practical value of A is given by the final value of the world that would result, were you to perform A

Epistemic Value The epistemic value of a credence function is given by how accurately that credence function describes the actual world.

Without going through all of the mathematical details, the first two theses give you causal decision theory for acts. But the final thesis gives you something different for credences. In particular, you get all of the recommendations Greaves endorses for *Promotion*, *Leap*, *Embezzlement*, and *Imps*.

Review Questions

1. What does *Conditionalization* say? What does *Reflection* say? Use the Monty Hall puzzle to illustrate why, if the set of propositions which you might learn, $\mathscr{E}$, does not form a partition, *Conditionalization* and *Reflection* can give contradictory advice.

2. What is a *Dutch strategy*? Suppose that, over a period of time, I successfully sell you a combination of bets which are guaranteed to lose you money no matter what. Is this enough to show that you were susceptible to a Dutch strategy? Why or why not?

3. What is the *Dutch Strategy Theorem*, and how could it be used to argue for *Conditionalization*?

4. What is the *Accuracy Dominance Avoidance* argument for *Conditionalization*?

5. What is Greaves & Wallace's *Expected Accuracy Maximization* argument for *Conditionalization*?

6. What is the thesis of epistemic consequentialism, and what role does it play in Greaves & Wallace's argument for conditionalization? Describe Greaves' *Imps* case and explain why it poses a problem for epistemic consequentialism.

7
Alternatives to Conditionalization

7.1 Conditionalization and Certainty

Notice that conditionalization tells us that you must be certain of your evidence, since

$$\mathbb{C}(E \mid E) = \mathbb{C}(E \wedge E) \div \mathbb{C}(E) = \mathbb{C}(E) \div \mathbb{C}(E) = 1$$

Moreover, conditionalization makes certainty *permanent*. Conditionalization can lower a proposition's credence from $1 - \epsilon$ to ϵ. But once you are certain that E, conditionalization will never lower E's credence from 1. That's because, if $\mathbb{C}(E) = 1$, then $\mathbb{C}(E \wedge F) = \mathbb{C}(F)$. So, if $\mathbb{C}(E) = 1$, then

If $\mathbb{C}(E) = 1$, then $\mathbb{C}(\neg E) = 0$ by additivity. Then, $\mathbb{C}(\neg E \wedge F) = 0$ by monotonicity. By additivity again, $\mathbb{C}(F) = C(E \wedge F) + \mathbb{C}(\neg E \wedge F) = \mathbb{C}(E \wedge F) + 0 = \mathbb{C}(E \wedge F)$.

$$\mathbb{C}(E \mid F) = \mathbb{C}(E \wedge F) \div \mathbb{C}(F) = \mathbb{C}(F) \div \mathbb{C}(F) = 1$$

So conditionalization says both that you must be certain of whatever your evidence tells you, and that you must *remain* certain of whatever your evidence tells you, no matter what. But consider cases like the following:

Example 15 (Observation by Candlelight (Jeffrey, 1965))**.** *The agent inspects a piece of cloth by candlelight, and gets the impression that it is green, although he concedes that it might be blue or even* (*but very improbably*) *violet. If G, B, and V are the propositions that the cloth is green, blue, and violet, respectively, then the outcome of the observation might be that, whereas originally his degrees of belief in G, B, and V were* .30, .30, *and* .40, *his degrees of belief in those same propositions after the observation are* .70, .25, *and* .05.

Example 16 (Undercutting Defeat)**.** *You look at a red wall in normal lighting conditions, and thereby come to learn that the wall is red. However, afterwards, you are informed by a reliable* (*but in this case, incorrect*) *informant that the wall is actually white, and bathed in red lighting.*

In the first example, from Richard Jeffrey, your credences in various propositions change, but it does not seem that there is any proposition which is learnt *with certainty*. Even if there is some proposition which exactly describes the precise visual experience you have undergone, this proposition needn't be in the algebra over which your credences are defined. So there's no proposition (in your algebra, at least) which

Even if there is such a proposition, we might think that you haven't *learned* that proposition. Consider, for instance, Ayer's "speckled hen": you look at a speckled hen. Perhaps your experience contains some precise number of speckles, but it doesn't seem that you should be certain about how many speckles the hen has, given that you can't reliably discriminate the number of speckles on a hen without careful counting.

is learnt with certainty. But it looks like you should still revise your credences in some way. So Jeffrey concludes that conditionalization does not always apply. There are some cases in which you should revise your credences in some way other than by conditioning on what you've learnt.

In the second example, introduced as a problem for conditionalization by Jonathan Weisberg,[1] we can identify a proposition which is learnt—this is the proposition "the wall is red". But we want to say that your credence in this proposition should start out high and then, after hearing from your informant, it should become low. But as we've seen, conditionalization can never lower a proposition's credence from 100%. So if you update by conditioning on the proposition "the wall is red", then your credence in this proposition will rise to 100% and it will not fall when you hear from the informant.

[1] Weisberg, Jonathan (2009). "Commutativity or Holism? A Dilemma for Conditionalizers". *British Journal for the Philosophy of Science 60* (4): 793-812.

You might think that there is a different proposition which is learnt in example 2: perhaps what you learn is just that the wall *seems* red. Weisberg contends that this just moves the bump in the carpet. Suppose that the wall in fact seems red, but that later a reliable (but in this case, incorrect) informant tells you that you've recently ingested a drug that makes you make mistakes about how things seem to you. This drug would make you falsely think that things seem red when in fact they seem green. In that case, Weisberg thinks that your credence in "the wall seems red" should fall, but it will not fall if you've conditioned on this proposition.

7.2 *Jeffrey Conditionalization*

In response to the first kind of case, Jeffrey proposed a generalization of conditionalization. To understand what this generalization says, think about how an arbitrary proposition, A, can be broken down, into the parts of it that overlap the cells of the partition $\{G, B, V\}$:[2]

[2] We can assume that you knew for sure that the cloth would be either green, blue, or violet, so that $\{G, B, V\}$ partitions your credal state.

$$\mathbb{C}(A) = \mathbb{C}(A \wedge G) + \mathbb{C}(A \wedge B) + \mathbb{C}(A \wedge V)$$

Or, equivalently, using the definition of conditional probability,

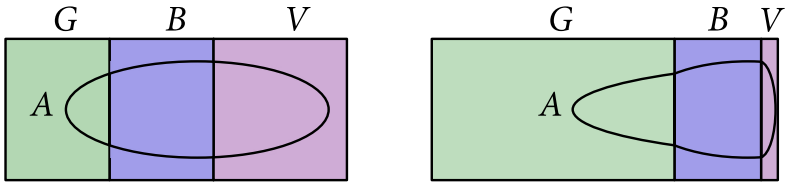
$$\mathbb{C}(A) = \mathbb{C}(A \mid G) \cdot \mathbb{C}(G) + \mathbb{C}(A \mid B) \cdot \mathbb{C}(B) + \mathbb{C}(A \mid V) \cdot \mathbb{C}(V)$$

Jeffrey's idea is that, if your credences are *directly* affected along the partition $\{G, B, V\}$, then you should adjust your probabilities for G, B, and V, but you should *keep the conditional probabilities* $\mathbb{C}(A \mid G)$, $\mathbb{C}(A \mid B)$, and $\mathbb{C}(A \mid V)$ fixed. So your posterior, after learning, should be

$$\mathbb{C}^+(A) = \mathbb{C}(A \mid G) \cdot 0.7 + \mathbb{C}(A \mid B) \cdot 0.25 + \mathbb{C}(A \mid V) \cdot 0.05$$

You can visualize this with figure 7.1, where credence corresponds to area, and the cell G has been 'stretched out', having its overall credence raised, and B and V have both been 'smushed in', having their overall credence lowered. The proposition A goes along for the ride, getting stretched or smushed in each cell in proportion to its area in that cell.

In general, Jeffrey assumes that you will have your credences *di-*



(a) Your Prior, before observation

(b) Your Posterior, after Jeffrey Conditioning

Figure 7.1: Jeffrey Conditioning on the weighted partition, or '*Jeffrey shift*', $\{(G, .7), (B, .25), (V, .05)\}$

rectly affected along some partition $\{E_1, E_2, \ldots, E_N\}$, making it rational for you to adopt the new posterior credences $q_1, q_2, \ldots, q_N$, respectively, in each cell of this partition. We can represent this with a 'weighted partition',

$$\{(E_1, q_1), (E_2, q_2), \ldots, (E_N, q_N)\}$$

which is to be interpreted as follows: it is now rational for you to have a credence of q_i in the proposition E_i (for each i). Notice that, assuming probabilism, this means that the 'weights' in a weighted partition will have to sum to 1. I'll call the experience whose evidential import gets represented with a weighted partition like this a *Jeffrey shift*. And then, Jeffrey says that, if you undergo a Jeffrey shift like this, you should adopt the new posterior

$$\mathbb{C}^+(A) = \mathbb{C}(A \mid E_1) \cdot q_1 + \mathbb{C}(A \mid E_2) \cdot q_n + \cdots + \mathbb{C}(A \mid E_N) \cdot q_N$$

Jeffrey Conditionalization If your prior credence function is $\mathbb{C}$, and you undergo a Jeffrey shift making it rational to have a posterior credence of q_i in cell E_i of the partition $\{E_1, E_2, \ldots, E_N\}$, then your posterior credence in any proposition, A, should be

$$\mathbb{C}^+(A) = \sum_{i=1}^{N} \mathbb{C}(A \mid E_i) \cdot q_i$$

A small bookkeeping point: if $\mathbb{C}(A \wedge E_i) = 0$, then we can say by convention that $\mathbb{C}(A \mid E_i) = 0$, too, to make sure that each term will be well-defined.

Notice that Jeffrey conditionalization has regular conditionalization as a special case. If we let the weighted partition be $\{(E, 1), (\neg E, 0)\}$, then Jeffrey conditionalization says that your posterior should be

$$\mathbb{C}^+(A) = \mathbb{C}(A \mid E) \cdot 1 + \mathbb{C}(A \mid \neg E) \cdot 0 = \mathbb{C}(A \mid E)$$

7.2.1 *Arguments for Jeffrey Conditionalization?*

Brian Skyrms And Brad Armendt have given two (importantly different) Dutch strategy arguments for Jeffrey conditionalization. There is an attempted expected accuracy maximization argument for Jeffrey conditionalization given by Ben Levinstein, but this argument only works given a measure of inaccuracy which isn't strictly proper—so, to my knowledge, there is no expected accuracy maximization argument for Jeffrey conditionalization.

See Levinstein, Benjamin Anders (2012). "Leitgeb and Pettigrew on Accuracy and Updating." *Philosophy of Science* 79 (3):413-424. For the criticism, see Gallow, J. Dmitri (2019). "Learning and Value Change". *Philosophers' Imprint* 19:1–22.

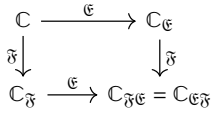
7.2.2 *Objection to Jeffrey Conditionalization: Non-commutativity*

Suppose you take one look at the cloth, and you get the Jeffrey shift $\{(G, .7), (B, .25), (V, .05)\}$. Then, you take another look and get the second Jeffrey shift $\{(G, .5), (B, .25), (V, .25)\}$. You will first stretch out the G cell to a credence of 70%, and next shrink it back down to a credence of 50%. You'll likewise shrink V to 5% and then stretch it back out to 25%. Then, consider your friend, you looks at the cloth and *first* gets the Jeffrey shift $\{(G, .5), (B, .25), (V, .25)\}$, and *then* gets the Jeffrey shift $\{(G, .7), (B, .25), (V, .05)\}$. This friend will first stretch G out to 50%, and then stretch it out even further to 70%. They'll first shrink V to 25%, and then shrink it even further, down to 5%. So even if you and your friend started out with the same prior, you'll end up at a different posterior. You will have a credence of $0.5, 0.25, 0.25$ in G, B, and V, respectively, whereas your friend will have a credence of $0.7, 0.25, 0.05$ in G, B, and V.

But *both you and your friend had exactly the same experiences*—just in a different order! Surely the order in which evidence is acquired shouldn't make a difference to what it's rational to believe after all the evidence is in.

Commutativity If you learn two things, $\mathfrak{E}$ and $\mathfrak{F}$, then the rational thing to believe, after learning both of these things, shouldn't depend upon the order you learnt them in. So, if we use '$\mathbb{C}_{\mathfrak{E}}$' for your credences updated on $\mathfrak{E}$, then we should have

$$\mathbb{C}_{\mathfrak{E}\mathfrak{F}} = \mathbb{C}_{\mathfrak{F}\mathfrak{E}}$$



It looks like Jeffrey conditionalization is non-commutative. But since commutativity looks like a rational constraint on any plausible learning procedure, this seems like a problem for Jeffrey conditionalization.[3]

One response to this problem, articulated by Marc Lange,[4] is that we need to be more careful about how we understand the *things* in the definition of commutativity. In the context of Jeffrey conditionalization, we can think of the weighted partition, or the Jeffrey shift, as the evidence acquired. But Lange says that, if we think of things this way, then we shouldn't want commutativity of *evidence*, since what evidence we receive might depend upon our background beliefs. In particular, how confident we should be that the cloth is violet after our glimpse through candlelight might depend upon how confident we were *beforehand* that the cloth was violet. So the Jeffrey shift is going to include *both* information about what happened in experience *and* information about your prior degrees of belief. Instead of commutativity of *Jeffrey shifts*, what we should want is commutativity of *learning experiences*.

So Lange suggests that, if you and your friend receive those same *Jeffrey shifts*, in those two different orders, then you and your friend couldn't have undergone the same *experiences*. After all, for you, the second experience *disconfirmed* G; whereas, for your friend, the first

[3] Versions of this complaint are found in Domotor, Zoltan (1980). "Probability kinematics and representation of belief change". *Philosophy of Science* 47 (3):384-403, Skyrms, Brian (1966). *Choice and chance*. Belmont, Calif.,: Dickenson Pub. Co., and Van Fraassen, Bas C. (1989). *Laws and symmetry*. New York: Oxford University Press.

[4] Lange, Marc (2000). Is Jeffrey Conditionalization Defective By Virtue of Being Non-Commutative? Remarks on the Sameness of Sensory Experiences. Synthese 123 (3):393-403.

As Field puts the point:

> ...it is clear that the probability q which I attached to an observation sentence E after I have been subjected to s sensory stimulation will depend not only on the sensory stimulation but also on the probability I attached to E before the stimulation.

experience *confirmed G*.

7.3 Field Conditionalization

This kind of position can be buttressed by considering an alternative way of understanding and representing Jeffrey conditionalization, offered by Hartry Field,[5] Instead of taking the *input* to an update rule to be the posterior credences that are rationalized after your experience, Field takes the input to be the *degree to which the experience has confirmed* any given cell of the partition $\{E_1, E_2, \ldots, E_N\}$. Let's represent the degrees of confirmation given to cell E_i with 'α'. If $\alpha_i > 1$, then E_i has been confirmed. If $\alpha_i < 1$, then E_i has been disconfirmed. If $\alpha_i = 0$, then E_i has been neither confirmed nor disconfirmed.

[5] See Field, Hartry (1978). "A note on Jeffrey conditionalization". *Philosophy of Science* 45 (3):361-367.

I'll represent the input to Field's rule a *Field shift*. It will be represented with a differently-weighted partition,

$$\{(E_1, \alpha_1), (E_2, \alpha_2), \ldots, (E_N, \alpha_N)\}$$

In the case of a Jeffrey shift, we assumed that $\sum_i q_i = 1$. In the case of a Field shift, we should assume that any confirmation one proposition receives is balanced out with disconfirmation somewhere else, so that $\prod_i \alpha_i = 1$.

Then, Field says that your posterior should be given by taking each cell of the partition E_i, stretching or smushing it by the factor α_i, taking every part of the proposition A lying inside that cell 'along for the ride', and then (finally) re-normalizing so that everything sums back to 100%. More carefully, we should have

$$\mathbb{C}^+(A) = \frac{\mathbb{C}(A \mid E_1) \cdot \mathbb{C}(E_1) \cdot \alpha_1 + \cdots + \mathbb{C}(A \mid E_N) \cdot C(E_N) \cdot \alpha_N}{\mathbb{C}(E_1) \cdot \alpha_1 + \ldots \mathbb{C}(E_N) \cdot \alpha_N}$$

Or, equivalently, since $\mathbb{C}(A \mid E_i) \cdot \mathbb{C}(E_i) = \mathbb{C}(AE_i)$,

$$\mathbb{C}^+(A) = \frac{\mathbb{C}(AE_1) \cdot \alpha_1 + \cdots + \mathbb{C}(AE_N) \cdot \alpha_N}{\mathbb{C}(E_1) \cdot \alpha_1 + \ldots \mathbb{C}(E_N) \cdot \alpha_N}$$

Field Conditionalization If your prior credence function is $\mathbb{C}$ and you undergo a Field shift confirming cell E_i of the partition $\{E_1, E_2, \ldots, E_N\}$ to degree α_i, then your posterior credence in any proposition A should be

$$\mathbb{C}^+(A) = \frac{\sum_{i=1}^N \mathbb{C}(AE_i) \cdot \alpha_i}{\sum_{i=1}^N \mathbb{C}(E_i) \cdot \alpha_i}$$

Now, notice that, while *Jeffrey* conditionalization does not commute on Jeffrey shifts, *Field* conditionalization does commute on Field shifts.

To see that Field Conditionalization commutes on Field shifts, let $\mathfrak{E} = \{(E_1, \alpha_1), \ldots, (E_N, \alpha_N)\}$, and let $\mathfrak{F} = \{(F_1, \beta_1), \ldots, (F_M, \beta_M)\}$. Then,

$$\begin{aligned}
\mathbb{C}_{\mathfrak{E}\mathfrak{F}}(A) &= \frac{\sum_{j=1}^M \mathbb{C}_{\mathfrak{E}}(AF_j) \cdot \beta_j}{\sum_{j=1}^M \mathbb{C}_{\mathfrak{E}}(F_j) \cdot \beta_j} \\
&= \frac{\sum_{j=1}^M \frac{\sum_{i=1}^N \mathbb{C}(AE_iF_j) \cdot \alpha_i}{\sum_{i=1}^N \mathbb{C}(E_i) \cdot \alpha_i} \beta_j}{\sum_{j=1}^M \frac{\sum_{i=1}^N \mathbb{C}(F_jE_i) \cdot \alpha_i}{\sum_{i=1}^N \mathbb{C}(E_i) \cdot \alpha_i} \cdot \beta_j} \\
&= \frac{\sum_{j=1}^M \sum_{i=1}^N \mathbb{C}(AE_iF_j) \cdot \alpha_i \cdot \beta_j}{\sum_{j=1}^M \sum_{i=1}^N \mathbb{C}(F_jE_i) \cdot \alpha_i \cdot \beta_j}
\end{aligned}$$

And this is precisely the same function you'll get if you first update on $\mathfrak{F}$ and next update on $\mathfrak{E}$. (You can see this by noting the symmetry in the above equation.)

7.3.1 Objections to Field Conditionalization

Daniel Garber objects to Field conditionalization by considering situations in which you look at the cloth in candlelight repeatedly. Suppose you take one glance at the cloth, and get the Field shift $\mathfrak{E} =$

$\{(G, 2), (\neg G, 1/2)\}$—we can suppose that you know that the cloth is either green or not green. This experience confirms that the cloth is green.. If your prior credence in G was 50%, your posterior credence in G will be

$$\begin{aligned}\mathbb{C}_{\mathfrak{E}}(G) &= \frac{\mathbb{C}(G) \cdot 2}{\mathbb{C}(G) \cdot 2 + \mathbb{C}(\neg G) \cdot 1/2} \\ &= \frac{0.5 \cdot 2}{0.5 \cdot 2 + 0.5 \cdot 1/2} \\ &= \frac{1}{1.25} = 0.8\end{aligned}$$

Likewise, your But then suppose you undergo *exactly the same experience*. If we think that the same experience should lead to the same Field shift, then this second experience will also provide the very same Field shift, $\mathfrak{E} = \{(G, 2), (\neg G, 1/2)\}$. So your new credence in G will be

$$\begin{aligned}\mathbb{C}_{\mathfrak{E}\mathfrak{E}}(G) &= \frac{\mathbb{C}_{\mathfrak{E}}(G) \cdot 2}{\mathbb{C}_{\mathfrak{E}}(G) \cdot 2 + \mathbb{C}_{\mathfrak{E}}(\neg G) \cdot 1/2} \\ &= \frac{0.8 \cdot 2}{0.8 \cdot 2 + 0.2 \cdot 1/2} \\ &= \frac{1.6}{1.7} \approx 0.94\end{aligned}$$

So your credence that the cloth is green will jump again to 94%. Garber points out that by simply looking again and again at the cloth in candlelight—an experience which only very weakly confirms that the cloth is green—you can end up arbitrarily confident that the cloth is green. This looks like the wrong result.

7.4 The Partitionality Assumption

Recall that, when we were justifying the rule of conditionalization, we frequently had to assume that you were going to learn that one of a *partition* of propositions was true. In particular, this was assumed in the the Dutch Strategy argument for conditionalization and it was assumed in the Greaves & Wallace accuracy maximization argument for conditionalization.

It turns out that this assumption is closely related to the following theses about the nature of evidence:

Factivity If you learn that E, then E is true

Positive Introspection If you learn that E, then you will learn that you have learnt that E

Negative Introspection If you don't learn that E, then you will learn that you haven't learnt that E

To appreciate how these assumptions are related to the assumption that $\mathscr{E}$—the set of propositions which you might learn—forms a partition, consider a Kripke model for evidence, where we introduce a binary relation, R, which one world bears to another, wRw^*, iff w^* is consistent with everything that you've learnt at w. Write '$\Box E$' iff E

is true at *every* world consistent with everything you've learnt. And write '$\Diamond E$' iff E is true at *some* world consistent with everything you've learnt. Then, $\Box E$ means that you've learnt that E, and $\Diamond E$ means that you've not learnt $\neg E$—or, equivalently, that E is true for all you've learnt.[6] With this modal semantics in the background, we can render the three assumptions above as:

Factivity $\Box E \rightarrow E$

Positive Introspection $\Box E \rightarrow \Box\Box E$ (Equivalently: $\Diamond\Diamond E \rightarrow \Diamond E$)

Negative Introspection $\neg\Box E \rightarrow \Box\neg\Box E$ (Equivalently: $\Diamond\Box E \rightarrow \Box E$)

[6] In general, to use a Kripke semantics like this, you should check that $\Box$ satisfies the K-axiom, $\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$ and the rule of necessitation: if $\vdash \phi$, then $\vdash \Box\phi$. But these assumptions seem plausible if we interpret $\Box\phi$ as meaning that you've learnt that ϕ, and we're willing to assume logical omniscience.

Then, Factivity will require that this binary relation is reflexive. Why? Suppose that $\neg wRw$. Then, w is not consistent with everything you've learnt at w. So you've learnt something inconsistent with w. So you've learnt something false. And Positive Introspection will require that the binary relation R is transitive. Why? Suppose you had a failure of transitivity: w_1Rw_2 and w_2Rw_3, yet $\neg w_1Rw_3$. Then, at w_2, w_3 is true for all you've learnt, $\Diamond w_3$. So, at w_1, for all you've learnt, w_3 is true for all you've learnt, $\Diamond\Diamond w_3$. But at w_1, w_3 is not true for all you've learnt, since you've learnt $\neg w_3$. So you violate Positive Introspection at w_1. And finally, Negative Introspection will require this binary relation to be Euclidean.[7] Why? Suppose you have a failure of Euclideanness: w_1Rw_2 and w_1Rw_3, yet $\neg w_2Rw_3$. Then, at w_1, you've not learnt that $\neg w_3$, $\neg\Box\neg w_3$. Yet you haven't *learnt* that you've not learnt that. For, at w_2 (which is consistent with everything you've learnt at w_1), you have learnt that $\neg w_3$. So there's something you haven't learnt that you haven't learnt you've not learnt. So Negative Introspection fails.

[7] A binary relation is *Euclidean* iff whenever aRb and aRc, bRc.

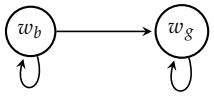
In general, any binary relation which is reflexive, transitive, and Euclidean will be what's known as an *equivalence relation*.[8] And any equivalence relation can be used to *partition* the set of things it's defined over. Given an equivalence relation R, let $[w]_R$ be the *equivalence class* of w under R: $[w]_R = \{w^* \mid wRw^*\}$. Then, you can show that the equivalence classes of an equivalence relation R will form a partition.

[8] An equivalence relation is often defined as a relation that's reflexive, symmetric, and transitive. Try to convince yourself that any relation that's reflexive, transitive, and Euclidean will also be symmetric; and try to convince yourself that any relation that's reflexive, symmetric and transitive will also be Euclidean.

So if we accept Factivity and Positive and Negative Introspection, then we can assume that the set of propositions that *might* be your total evidence, $\mathscr{E}$, will form a partition. And there is no loss of generality in making this assumption in the Dutch Strategy and Expected Accuracy Maximization arguments for Conditionalization.

However, if either Positive Introspection or Negative Introspection fail, then the set of possible evidence partitions will *not* form a partition. Let's see two simple examples of how that could happen.

Good Case/Bad Case In the good case, you look at a red wall in white lighting. In the bad case, you look at a white wall bathed in red lighting. In the good case, you learn that the wall is red. In the bad case, you don't learn anything about the wall's color.

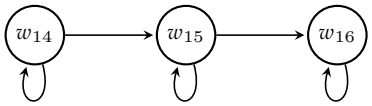


In this case, there is a failure of Negative Introspection. In the bad case, you don't learn that you're in the good case (you don't learn that

the wall is red). However, neither do you learn that you're in the bad case. So it's consistent with everything you learn that you're in fact in the good case, where you've learnt that you're in the good case. So you don't learn something (that the wall is red), but you also don't learn that you haven't learnt that thing.

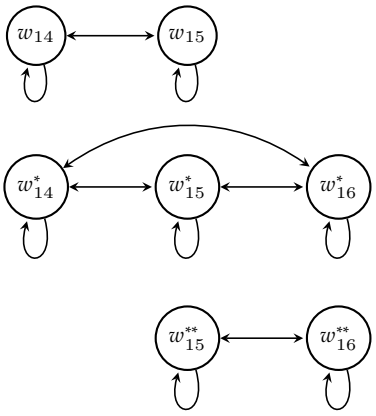
You could object that, in both the good and the bad case, you learn the same thing—namely, that it *seems* that the wall is red. This is a kind of evidence *internalism*, according to which, anytime you have the same experience, you'll have the same evidence. Some, persuaded by authors like McDowell and Williamson, have come to reject this picture and instead favor a form of evidence *externalism*, according to which two people identical from the skin in can have different evidence. (They use this position to respond to various traditional skeptical arguments.)

Imperfect Vision The tree in the distance is either 14, 15, or 16 meters tall, and you're about to take a look at it. Your vision is only good enough to discriminate between heights of 1 meter or more. So, if the tree is 14 meters tall, you'll learn that it's not 16 meters tall. And, if the tree is 16 meters tall, then you'll learn that it's not 14 meters tall. But, if it's 15 meters tall, you won't learn anything at all.



In this case, we have a failure of Positive Introspection. When the tree is 14 meters tall, for all you've learnt, it's 15 meters tall. And, if it's 15 meters tall, then for all you've learnt, it's 16 meters tall. So, for all you've learnt, for all you've learnt, it's 16 meters tall, $\Diamond\Diamond w_{16}$. Yet you've learnt that it's not 16 meters tall, so it's not the case that, for all you've learnt, it's 16 meters tall, $\neg\Diamond w_{16}$. Positive Introspection fails.

You could defend Positive Introspection from this argument by introducing some additional proposition which is learnt—for instance, suppose that when you look at the tree, you make a *guess* as to its height. And you know that your guesses are never more than 1 meter off. Then, we can model this situation with a partitional update where you're simply going to learn what your *guess* is. (In the first row in the margin, your guess was 14; in the second, your guess was 15; and in the third, your guess was 16.)



7.5 *Schoenfield Conditionalization*

Suppose that you are persuaded that there are failures of Positive or Negative Introspection. What should you do then? Miriam Schoenfield makes a suggestion and gives an expected accuracy maximization argument for it.

To first understand the rule she proposes, let's use '$\mathbf{T}E$' for the proposition that E is your *t*otal evidence—that is to say, $\mathbf{T}E$ says that you learnt E, and you didn't learn anything stronger than E—anything other than E you learnt is entailed by E. For instance, in *Good Case/Bad Case*, your total evidence in the bad case is $\{w_b, w_g\}$. So $\mathbf{T}\{w_b, w_g\}$ is true at w_b. And your total evidence in the good case is $\{w_g\}$. So $\mathbf{T}\{w_g\}$

is true at w_8. As this example demonstrates, even if *your potential evidence* $\mathscr{E}$ doesn't form a partition, $\mathbf{T}\mathscr{E} = \{\mathbf{T}E_1, \mathbf{T}E_2, \ldots, \mathbf{T}E_N\}$ *will* form a partition. (Can you say why?)

Then, Schoenfield proposes that, if your total evidence is E, you should condition *not* on E, but rather on the proposition that E is your total evidence, $\mathbf{T}E$.[9]

Schoenfield Conditionalization If $\mathbb{C}$ is your prior credence function and you learn that E (and nothing stronger), then your posterior credence in any proposition A should be

$$\mathbb{C}^+(A) = \mathbb{C}(A \mid \mathbf{T}E)$$

[9] See Schoenfield, Miriam (2017). "Conditionalization Does Not (In General) Maximize Expected Accuracy". *Mind* 126 (504):1155-1187.

Schoenfield's suggestion was made earlier (and independently) by Matthias Hild. See Hild, Matthias (1998). "Auto-epistemology and updating". *Philosophical Studies* 92 (3):321-361.

And she gives an expected accuracy maximization argument to support this update rule. Her argument builds on the framework of Greaves & Wallace we talked about earlier. Recall:

Update Plans An *update plan*, $\mathcal{U}$, is a function from worlds in $\mathcal{W}$ to probability functions over $\mathcal{A}$. $\mathcal{U}_w$ is the posterior probability function the update plan prescribes adopting in the world w.

Greaves & Wallace made the following assumption about which update plans were *available*,

Available Update Plans (Greaves & Wallace) If you are going to learn the true member of the (finite) partition $\mathscr{E} = \{E_1, E_2, \ldots, E_N\}$, then an update plan is *available* iff, for any $E \in \mathscr{E}$, and any two worlds $w, w^* \in E$, $\mathcal{U}_w = \mathcal{U}_{w^*}$.

But what about if the set of propositions that might be your total evidence, $\mathscr{E}$, does *not* form a partition? In that case, the definition of 'available' update plans looks overly restrictive. For instance, take the simple model of imperfect vision (the one on which Positive Introspection fails). In that case, suppose we say that an update plan is *available* iff, for any $E \in \mathscr{E}$ and any $w, w^* \in E$, $\mathcal{U}_w = \mathcal{U}_{w^*}$. Then, we would say that your update plan must be *constant*—it must be the same in each of w_{14}, w_{15}, and w_{16}.

So Schoenfield tightens the definition of available update plan to make more plans *available*. She says that a plan is *available* iff it gives the same advice in any two worlds where you get the same evidence:

Available Update Plans (Schoenfield) If you are going to learn the true member of $\mathscr{E} = \{E_1, E_2, \ldots, E_N\}$, then an update plan is *available* iff, for any $E \in \mathscr{E}$, and any two worlds $w, w^* \in \mathbf{T}E$, $\mathcal{U}_w = \mathcal{U}_{w^*}$.

Then, Schoenfield proves the following theorem:

Theorem (Schoenfield) If inaccuracy is measured with a strictly proper measure, and you stand to learn some true member of $\mathscr{E} = \{E_1, E_2, \ldots, E_N\}$, then the available update plan which minimizes expected inaccuracy is the plan to Schoenfield conditionalize on your evidence.

Review Questions

1. Explain why conditionalization requires certainty about your evidence, and why Jeffrey conditionalization does not. Explain why Jeffrey conditionalization is a generalization of conditionalization (that is: why Jeffrey conditionalization has conditionalization as an instance).

2. What is commutativity? Explain why Jeffrey conditionalization appears to be non-commutative. How does Lange argue that Jeffrey conditionalization is in fact commutative? What is Field conditionalization? What is the relationship between Jeffrey and Field conditionalization? And how can Field conditionalization be used to buttress Lange's defense of Jeffrey conditionalization's commutativity?

3. What is the partionality assumption, and why does Shoenfield think we should reject conditionalization if the partitionality assumption is fails? Describe two situations in which you might think that the partitionality assumption fails.

8
Self-Locating Credence and Memory Loss

8.1 De Se and De Dicto Propositions

Let's start off by distinguishing two different kinds of contents (or objects of belief/credence): *de dicto* and *de se*.[1] *De dicto* propositions say something about what the world is like. We can model *de dicto* propositions with sets of possible worlds (as we have throughout this course). *De se* propositions say something about *who* you are, *where* you are, or *when* it is. These kinds of propositions cannot be modeled with sets of possible worlds.

[1] See John Perry's *The Problem of the Essential Indexical* and David Lewis's *Attitudes De Dicto and De Se*.

A classic example: Rudolph Lingens sits in the library in Stanford, and Ludolph Ringens sits in an identical library in Princeton. Neither ever sets foot outside of the library, and both read all the same books, in the same order. The library is comprehensive enough that both Rudolph and Ludolph come to know every possible fact about what the world is like: they know precisely which world is actual, down to the least detail. But still, there is a respect in which they are ignorant—for each of Rudolph and Ludolph know of the other's existence, and neither knows which they are. They know that there is someone with their experience on the East Coast, and they know that there is someone with their experience on the West Coast, but they do not know which they are.

Can we solve this problem by introducing haecceities? This would allow us to have two worlds where before we had just one—the one where Rudolph's haecceity sits on Rudolph and Ludolph's haecceity sits on Ludoph, and the one where the two are swapped. Even so, Rudolph could come to know precisely where each haecceity sits but still fail to know which haecceity is *his*.

Think of it like this: a possible world gives you a complete map of reality. But the map does not contain a "you are here" sticker. Even once you have the full map of reality, you can remain ignorant of where you are in reality. The lesson is that *de se*, or self-locating, information goes beyond merely *de dicto*, or non-self-locating, information.

Lewis introduced a model for representing *de se* contents which has become widespread, and which we will adopt here. On this model, the contents of belief and credence are—not sets of *worlds*, but instead—sets of *centered* worlds. A centered world is a full map of reality, along with the corresponding 'you are here' sticker, which we'll call a 'center'. Now, just as we can have uncertainty about what the world is like, we can also have uncertainty about where in the world we are. And, if we are Bayesians, we will want to represent that uncertainty with a subjective probability distribution. So the subjective probability distribution will be over sets of *centered* worlds, and not just sets of

Formally, a *center* is a triple of a person, place, and time, $c = (p, x, t)$. And a centered world is a pair of a possible world and a center, (w, c)—where that person exists at that place and that time in that world.

worlds.

With this Lewisian model of *de se* content, we can distinguish propositions which are *de dicto* from those which are *de se* in a simple way. A *de dicto* proposition is one which does not distinguish between centers within a world. Picturesquely, it is a map of reality which doesn't tell you anything about where you are in the map. And a *de se* proposition is one which *does* distinguish between centers within a world. Picturesquely, it is a map of reality which tells you *something* (though not necessarily *everything*) about where you are in the map.

Formally, a *de dicto* proposition, A, is a set of centered worlds with the following property: for any two centered worlds which share a world member, (w, c) and (w, c^*), $(w, c) \in A$ iff $(w, c^*) \in A$.
Formally, a *de se* proposition, A, is a set of centered worlds with the following property: there are two centered worlds sharing a world member, (w, c) and (w, c^*), such that $(w, c) \in A$ and $(w, c^*) \notin A$.

There's another way of modeling things that's worth mentioning. You might think that the *de se* content "I am Lingans" is not importantly different from the *de dicto* content "Hesperus is Phosphorus". The set of worlds in which Hesperus is Phosphorus is just the set of all worlds (given the necessity of identity), but you can still be ignorant of this identity. And similarly (you may think) the set of worlds in which the proposition denoted by "I am Lingans" (in Lingan's mouth) is true is just the set of all worlds, but Lingans can still be ignorant of this identity. Perhaps the puzzles of *de se* content are just Frege puzzles.

In that case, you might think that what's going on when you're ignorant of whether you are Lingans is just what's going on in other Frege's puzzles. There are tons of proposed solutions to Frege's puzzle, but let's focus on one particular one here: on this view, we should distinguish between the *object* of belief/credence—the thing in which you have belief/credence—and the *state* of believing or having credence in that proposition. The *object* of belief is the proposition, whereas the *belief state* is the way in which you become related to that proposition. And there are multiple ways of being related to one and the same proposition. You might be related to the proposition that Lingans is in Stanford in one way (the one corresponding to the sentence 'Lingans is in Stanford') without being related to that proposition in another way (the one corresponding to the sentence 'I am in Stanford'). Call the ways of being belief-or-credence-like related to a proposition a *guise*. Then, the alternative approach would say that the arguments of your credence function aren't sets of centered possible worlds, but rather *guises* which (together with facts about your situation) determine sets of possible worlds.

A nice example to bring out the contrast between these two approaches: Hume believes that he is Hume, and Heimson in the insane asylum believes that he is Hume. On the Lewisian proposal, both Hume and Heimson have the same object of belief—they both self-ascribe the property of being Hume. On the guisey proposal, Hume and Heimson believe *different* propositions. Hume believes the necessarily true proposition that Hume is Hume, and Heimson believes the necessarily false proposition that Heimson is Hume. But they believe these two different propositions *in the same way*—that is to say, *under the same guise*.

If we adopt the Lewisian model, then it is easy to reformulate prob-

abilism. We simply swap out sets of worlds for sets of centered worlds, we understand contents as simply sets of centered worlds, and we use the very same axioms of probability with this change made. But if we adopt the guisey model, it's less clear what probabilism demands. For Heimson, the guise 'I am Hume' determines a necessarily false proposition. Should we take probabilism to say that 'I am Hume' must be given a credence of zero? Presumably not. So, if we are using the guisey model of self-locating probability, then we should presumably take the probability axioms to first-and-foremost govern *guises*, and not the propositions so determined.

For the Lewisian model: Let C be a set of centered possible worlds, and $\mathcal{A}$ an algebra of subsets of C. Then, (finitely additive) probabilism requires that:

1. $\mathbb{C}(C) = 1$
2. For any $A \in \mathcal{A}$, $\mathbb{C}(A) \geqslant 0$
3. For any $A, B \in \mathcal{A}$ such that $AB = \varnothing$, $\mathbb{C}(A \cup B) = \mathbb{C}(A) + \mathbb{C}(B)$.

8.2 *Counterexamples to Conditionalization and Reflection*

Once we've acknowledged that there are *de se* contents of credence, we will have to give up on the unrestricted principles of conditionalization and reflection. Consider the following case:

Time's Passing Your current credence in the *de se* proposition "today is Monday" is 100%. Tonight, you will go to sleep and tomorrow, you will learn that you have woken up. Upon learning that you've awoken, you plan to be 100% sure in the *de se* proposition "today is Tuesday", and have a credence of 0% in "today is Monday".

Conditionalization can never raise a proposition's probability from zero, nor lower a proposition's probability from 1. So you cannot be updating your subjective probabilities by conditioning. And you are very, very confident that you will awake tomorrow (and learn that you have), so your expectation of tomorrow's credence in 'today is Monday' is very, very low. So your expectation of your updated credence in 'today is Monday' does not match your current credence in 'today is Monday', in violation of the principle of Reflection.[2]

[2] Recall, the principle of reflection says that $\mathbb{E}[\mathbb{C}_{\mathscr{E}}(A)] = \mathbb{C}(A)$.

Notice that we have assumed here that the content that you are currently certain of is the same one that you will be certain is false tomorrow. This will follow given Lewis's model. On the Lewisian model, the object of belief is the set of centered worlds whose centers are Monday. However, on the guisey model, the content that you are currently certain of is not the same one that you will be certain is false tomorrow—though that content will be believed *in the same way*, or *under the same guise*. If we adopt the second model, we will still have a counterexample to conditionalition and reflection *for guises*. Your credence in *a guise* can rise from 0% and fall from 100%, and your expectation of your future credence in *a guise* can fail to match your current credence in *that guise*. Recall that, on the guisey model, we should want the probability axioms to govern the *guises* and not the propositions determined by those guises. So it's natural to expect conditionalization and reflection to also apply to the guises, and not the propositions determined by those guises. But so understood, examples like *Time's Passing* will be counterexamples to conditionalization and reflection.

8.2.1 The De Se Irrelevance Theses

A natural first thought after encountering this problem is that perhaps conditionalization and reflection break down for *de se* contents, but they are still fine when it comes to *de dicto* contents. The idea is that we can factor out the *de dicto* from the *de se*, and that, so factored, the latter are *irrelevant* to the former. There's a couple of different ways you could try to make sense of this irrelevance thesis. On the first, which is relatively uncontroversial, so long as you're only learning *de dicto* things, you can use conditionalization and reflection as before, even if you have credences in various *de se* propositions, too.

Weak De Se Irrelevance Thesis If you only stand to learn *de dicto* information, then your *de dicto* credences will satisfy Conditionalization and Reflection, even if your *de se* credences do not.

That is: if A is a *de dicto* proposition, and $\mathscr{E}$ is a partition of *de dicto* propositions, then $\mathbb{E}[\mathbb{C}_{\mathscr{E}}(A)] = \mathbb{C}(A)$, and you should be disposed, for any $E \in \mathscr{E}$, to have $\mathbb{C}_E(A) = \mathbb{C}(A \mid E)$.

On the second way of making sense of the irrelevance idea, which is *very* controversial and rejected by most Bayesian epistemologists, the *de se* is irrelevant to the *de dicto even if* you're learning *de se* information. Your *de dicto* credences should still satisfy reflection, and you should still update your *de dicto* credences by conditioning on the strongest *de dicto* proposition learnt.

Strong De Se Irrelevance Thesis No matter what you stand to learn, your *de dicto* credences will satisfy Conditionalization and Reflection, even if your *de se* credences do not.

That is: if A is a *de dicto* proposition, and $\mathscr{E}$ is *any* partition of propositions (de dicto or de se), then $\mathbb{E}[\mathbb{C}_{\mathscr{E}}(A)] = \mathbb{C}(A)$, and you should be disposed, for any $E \in \mathscr{E}$, to have $\mathbb{C}_E(A) = \mathbb{C}(A \mid E^+)$, where E^+ is the strongest *de dicto* proposition entailed by E.

See Arntzenius, Frank (2003). Some Problems for Conditionalization and Reflection. Journal of Philosophy 100 (7):356-370.

Here is a possible counterexample to the *strong* version of the irrelevance thesis.

The Prisoner You are scheduled to be executed tomorrow, unless pardoned. The governor flips a coin to decide all pardons. So if the coin landed tails, then you will be pardoned; heads, and you will not. The prison guard will know the outcome, but is not able to communicate with you—but if the coin landed tails, then they will leave the lights on after midnight to let you know your fate. It is currently 6:00pm. As time passes, you will learn that some time has passed, but you won't know exactly how much time has passed.

Consider what will happen at 11:59pm. At 11:59, your *de se* credences will have evolved, and you will be somewhat confident that it is currently after midnight. Just for illustration, suppose that, at 11:59, you will be 50% sure that it is after midnight. But the lights will be on. So your credence that you have been pardoned, P, will be:

$$\begin{aligned}\mathbb{C}(P) &= \mathbb{C}(P \mid \text{ before midnight}) \cdot 50\% + \mathbb{C}(P \mid \text{ after midnight}) \cdot 50\% \\ &= 50\% \cdot 50\% + 1 \cdot 50\% \\ &= 25\% + 50\% \\ &= 75\%\end{aligned}$$

As you get closer and closer to midnight, you will get more and more confident that it is in fact *after* midnight; and so you'll get more and

more confident that you've been pardoned. This change in your beliefs seems to be entirely rational. But it's also entirely *predictable*. So your belief-revision plans seem to violate the principle of Reflection. So the strong *de se* irrelevance thesis is false.

Here's a related puzzle from Jessica Collins:

Collins' Prisoner Everything is as in *The Prisoner*, except that there are two clocks in the room, exactly one of which is correct. The first clock, A, currently reads 6:00pm. The second clock, B, currently reads 7:00pm.

Currently, you have the credence distribution shown in figure 8.1a. But think about what will happen when it is in fact 11:30pm. If clock A is the correct one, then you will have eliminated the possibility that B is correct and the coin landed heads. So you'll have raised your credence in tails from $1/2$ to $2/3$ (figure 8.1b). But if clock B is the correct one, then you will know for sure that it is not yet midnight, and so you'll retain the prior distribution from 8.1a. So your expectation of your 11:30pm credence in heads is

$$\begin{aligned}\mathbb{E}[\mathbb{C}_{11:30}(H)] &= \mathbb{C}(A) \cdot 2/3 + \mathbb{C}(B) \cdot 1/2 \\ &= 1/2 \cdot 2/3 + 1/2 \cdot 1/2 \\ &= 1/3 + 1/4 \\ &= 7/12\end{aligned}$$

which is greater than your current credence in heads ($6/12$). So, if you're just sitting around watching time pass, you will violate the principle of reflection.

	A is correct	*B is correct*
heads	*1/4*	*1/4*
tails	*1/4*	*1/4*

(a)

	A is correct	*B is correct*
heads	*1/3*	
tails	*1/3*	*1/3*

(b)

Figure 8.1: Your prior credences (figure 8.1a) and your posterior credences at 11:30, if clock A is correct (figure 8.1b) in *Collins' Prisoner*.

8.2.2 *Conditionalization, Reflection, and (Possible) Memory Loss*

There's another kind of case that causes trouble for principles of conditionalization and reflection: *memory loss*. Let's start simply, with a counterexample to Reflection from Talbott.[3]

Forgotten Lunch Today, you are certain that you ate spaghetti for lunch. This time next year, you will have no idea what you ate for lunch today.

In this case, you have violated the principle of Reflection; your expectation of your future credence in a proposition is not equal to your current credence in that proposition. (Be careful here: Reflection does *not* say that you should lower your current credence that you ate spaghetti for lunch. It says instead that your future self should be certain that you ate spaghetti for lunch; it says that the memory loss was irrational.)

A case from Frank Arntzenius teaches us that you can cause trouble for Reflection even *without* losing your memory—it is enough that the memory loss is *possible*.[4]

Two Roads to Shangri La You are traveling to Shangri La, but those who enter Shangri La cannot know *how* they enter. So a fair coin will be

[3] See Talbott, William (1991). "Two principles of bayesian epistemology". *Philosophical Studies* 62 (2):135-150.

[4] See Arntzenius, Frank (2003). Some Problems for Conditionalization and Reflection. Journal of Philosophy 100 (7):356-370.

tossed. If it lands heads, you will take the mountain pass. If it lands tails, you will travel by sea. If you go by the mountain pass, no more steps will be taken. However, if you go by the sea, then upon arriving in Shangri La, your memories will be erased and replaced by (non-veridical) memories of having taken the moutain pass. In fact, the coin lands heads, and you find yourself looking out at the mountains on your way to Shangri La.

Currently, you should be certain that you are taking the mountain pass, and you should be certain that the coin landed heads. However, upon arriving in Shangri La, you should only have a credence of 50% that the coin landed heads. So you should not conditionalize in this case, and you should violate the principle of Reflection—you should be certain that the coin landed heads, and you should know for sure that, upon arriving in Shangri La, you'll have a credence of 50% that the coin landed heads.

Notice, though, that in this case, you do not actually lose any memory. It is enough that the memory loss is *possible*. It needn't be *actual*.

8.3 The Sleeping Beauty Puzzle

The following case from Adam Elga[5] has been a lightning rod for debate about *de se* credence.

Sleeping Beauty Know all the following: on Sunday, you will be put to sleep. Monday morning, you will be awoken in a room without a calendar. Monday evening, you will be informed that it is Monday, and Monday night, a fair coin will be flipped. If the coin lands heads, you will be put back to sleep and kept asleep all through Tuesday. You will be awoken outside of the room. If, however, the coin lands tails, then you will be put back to sleep and all of your memories of having been awoken on Monday will be erased. You will then awaken on Tuesday, and have experiences indistinguishable from the experiences you had on Monday.

There are two questions we can ask about your credence in *Sleeping Beauty*:

Q1 On Monday morning, how confident should you be that the coin flipped on Monday night lands heads?

Q2 On Monday evening, after you learn that it is Monday, how confident should be you that the coin flipped on Monday night lands heads?

The most natural and straightforward answer to question 1 is 'one half'. This answer also seems to follow from the principal principle, since you do not appear to have any inadmissible information (after all—it is in fact Monday morning, and the coin is yet to be flipped; there's no funny business involving time travel, so how could you have any inadmissible information?) The answer 'one half' to question 1

[5] See Elga, Adam (2000). "Self-locating belief and the sleeping beauty problem". *Analysis* 60 (2):143–147.

	Monday	*Tuesday*
heads	*1/2*	
tails	*1/4*	*1/4*

(a)

	Monday	*Tuesday*
heads	*1/3*	
tails	*1/3*	*1/3*

(b)

Figure 8.2: In figure 8.2a, the credence distribution recommended by the halfer for Monday morning. In figure 8.2b, the credence distribution recommended by the thirder for Monday morning.

also follows from the strong *de se* irrelevance thesis. Those who give the answer 'one half' are known as 'halfers'.

But here are two arguments from Elga that, in fact, the correct answer to the Sleeping Beauty puzzle is *one third*. (Those who give the answer 'one third' to question 1 are known as 'thirders'.) The first argument appeals to a connection between rational credence and long-run frequency: suppose that you were to run this experiment over and over again. In the long-run, as the number of trials of the experiment got larger and larger, the proportion of *heads* wakings (wakings on a run of the experiment in which the coin landed heads) would be one third. Since all you know upon waking is that you're awakening in a run of the experiment, your credence that it's a heads waking should be one third.

Second argument: if the coin landed tails, then it is just as likely that today is Monday as it is that today is Tuesday. So, conditional on the coin landing tails, your credence in "today is Monday" should be one half. Moreover, the answer to the second question should be one half—once you know that it is Monday and that the coin is about to be flipped, your credence that the coin lands heads should be one half. So—by conditionalization—your prior credence that the coin lands heads, *given* that it is Monday, should be one half.

P1) $\mathbb{C}(\textit{Monday} \mid \textit{tails}) = 1/2$

$\therefore$ C1) $\mathbb{C}(\textit{Monday} \wedge \textit{tails}) = \mathbb{C}(\textit{Tuesday} \wedge \textit{tails})$

P2) $\mathbb{C}(\textit{heads} \mid \textit{Monday}) = 1/2$

$\therefore$ C2) $\mathbb{C}(\textit{Monday} \wedge \textit{heads}) = \mathbb{C}(\textit{Monday} \wedge \textit{tails})$

$\therefore$ C3) $\mathbb{C}(\textit{Monday} \wedge \textit{heads}) = \mathbb{C}(\textit{Monday} \wedge \textit{tails}) = \mathbb{C}(\textit{Tuesday} \wedge \textit{tails})$

$\therefore$ C4) $\mathbb{C}(\textit{heads}) = 1/3$

Elga's second argument for the answer 'one third' to question 1 appealed to a certain answer to question 2. The thirder is happy to say that, upon learning that it is Monday, you should have a credence of one half that the coin will land heads. But David Lewis, for one, disagreed. His answer to Q1 was 'one half', but he thought that, upon learning that it was Monday, you should condition on this information, and arrive at the distribution shown in figure 8.3b. So Lewis's answer to Q2 was 'two thirds'.

Those who give this answer to Q2 are known as 'Lewisian halfers'. On the other hand, some have thought that we should give the answer 'one half' to *both* Q1 and Q2. These people are known as 'double halfers'.

The Lewisian halfer seems to be violating the principal principle. Given that this was Lewis's main reason for giving the answer 'one half' to Q1, how can he decide to answer 'two thirds' to Q2? Here is what Lewis says: "when Beauty is told during her Monday awakening that it's Monday...she is getting evidence—centered evidence—about the future: namely that she is not now in it." I leave you to ponder how plausible this story is.

	Monday	*Tuesday*
heads	*1/2*	
tails	*1/2*	

(a)

	Monday	*Tuesday*
heads	*2/3*	
tails	*1/3*	

(b)

Figure 8.3: In figure 8.3a, the credence distribution recommended by the 'double halfer' (and the thirder) for Monday evening. In figure 8.2b, the credence distribution recommended by the 'Lewisian halfer' for Monday evening.

The double halfer has a challenge: if we are not to learn by conditionalization, then how *are* we to learn? An answer has been suggested by Joe Halpern, Mark Tuttle, and Christopher Meacham. Their proposal is called 'Compartmentalized Conditionalization'. To appreciate this rule, some notation: if E is a set of centered worlds, then let $E^\dagger$ be the set of worlds compatible with E—that is, it is the set of worlds w such that, for some center c, $(w, c) \in E$. Then, compartmentalized conditionalization says

Compartmentalized Conditionalization If $\mathbb{C}$ is your prior, and you've gained the evidence E (where E is a set of centered worlds), then your posterior credence in each world should be given by conditioning on $E^\dagger$,

$$\mathbb{C}_E(w) = \mathbb{C}(w \mid E^\dagger)$$

and, within each world, your credences should be distributed uniformly over every center compatible with E:

$$\mathbb{C}_E(w, c) = \begin{cases} 0 & \text{if } (w, c) \notin E \\ \frac{\mathbb{C}_E(w)}{\#\{c|(w,c)\in E\}} & \text{if } (w, c) \in E \end{cases}$$

Applied to the Sleeping Beauty case, this rule will give us the Double Halfer result.

So we have three possible positions on Sleeping Beauty: the thirder, the Lewisian halfer, and the double halfer position. Since the halfer position is the most initially attractive position, I want to spend the remainder of this lesson explaining why the majority of Bayesian epistemologists have nonetheless decided that the thirder is correct.[6]

[6] According to the 2020 PhilPapers survey, 53% of Decision Theorists accept or lean towards one third, compared to only 18% of Decision Theorists accepting or leaning towards one half. (The remaining either accept an alternative view or else are undecided.) Thirding also comes out the most popular answer amongst all respondants, and if you filter by Epistemologists.

8.3.1 Chance Deference and the De Se

One of the central arguments in favor of halfing is that it follows from a principle of chance deference like the one we called 'the current principle'. But it's worth noting that, once we have *de se* beliefs in the mix, that principle faces clear counterexamples. For instance, consider the following:

De Se Chance Evidence You don't know whether it is Monday or Tuesday, and you think it's as likely to be Monday as it is to be Tuesday. And you know for sure that *today's* chance of Mudskipper winning is 75% and that *yesterday's* chance of Mudskipper winning is 25%. (In fact, today is Monday.) How confident should you be that Mudskipper wins the race?

Given that today is Monday, it doesn't look like you have any Monday inadmissible evidence—how could you? There's no funny business about time travelers from the future, and it's actually Monday, so how could you have any evidence about times after Monday?

If that's right, then it looks like, by applying a principle of chance deference to the Monday chances, we are told the following ('W' is the

proposition that Mudskipper wins):

$$\mathbb{C}(W \mid \langle C\, h_{mon}(W) = 25\% \rangle) = 25\%$$
$$\text{and } \mathbb{C}(W \mid \langle C\, h_{mon}(W) = 75\% \rangle) = 75\%$$

But, of course, you know for sure that the Monday chance of W is 25% iff today is Tuesday, and you know for sure that the Monday chance of W is 75% iff today is Monday. So:

$$\mathbb{C}(W \mid \text{ today's Tuesday}) = 25\%$$
$$\text{and } \mathbb{C}(W \mid \text{today's Monday}) = 75\%$$

And your credence that today is Monday is equal to your credence that today is Tuesday, which is 50%. So, by the law of total probability:

$$\begin{aligned}\mathbb{C}(W) &= \mathbb{C}(W \mid \text{ today's Tuesday}) \cdot 50\% + \mathbb{C}(W \mid \text{today's Monday}) \cdot 50\% \\ &= 25\% \cdot 50\% + 75\% \cdot 50\% \\ &= 50\%\end{aligned}$$

So, once you have *de se* uncertainty, the principle of chance deference says that you should be merely 50% confident that Mudskipper wins the race, despite knowing for sure that *today's* chance of Mudskipper winning is 75%. This seems to suggest that, once we have *de se* credences in the mix, we will need to generalize the principle of chance deference we accept. I've defended a generalization of the principle of chance deference to handle cases like this, and this generalization implies the thirder's answers in Sleeping Beauty.[7]

[7] See Gallow, J. Dmitri (forthcoming). "Two-Dimensional De Se Chance Deference." *Australasian Journal of Philosophy*.

8.3.2 Embarrassments for Double Halfers

I'll put aside the Lewisian halfer's position for now, and let's just consider the position of the double halfer. While compartmentalized conditionalization and double halfing can seem initially very attractive, the position starts to lose some of its luster when you subject it to scrutiny.

Embarassment #1: Firstly, you might have hoped that compartmentalized conditionalization would help vindicate the irrelevance of the *de se* to the *de dicto*, but this isn't how things work out in general. Consider, for instance, the following case:[8]

[8] See David Manley, "On Being a Random Sample", manuscript.

Sleeping Beauty with Some Lights Everything is as in *Sleeping Beauty*, except that there are now two indistinguishable rooms. On Monday, you will be inside room #1; and on Tuesday, you will be inside room #2 (perhaps asleep, if the coin landed heads). Independent of how your coin landed, each room has an independent 50% chance of its lights turning on at noon.

We can model this situation with 8 possible worlds: the world in which the coin lands heads and both rooms are lit at noon, hll, the world in which the coin lands heads and room one is lit while room two is dark, hld, the world in which the coin lands heads and room one is

dark while room two is lit, *hdl*, and so on. Presumably, the double halfer will want to say that you have the following credence distribution before noon:

World:		*hll*	*hld*	*hdl*	*hdd*	*tll*	*tld*	*tdl*	*tdd*
Room:	1	2/16	2/16	2/16	2/16	1/16	1/16	1/16	1/16
	2	0	0	0	0	1/16	1/16	1/16	1/16

Then, one of two things could happen: it could be that you learn that *your* room is lit. If you update on this information by compartmentalized conditionalization, it will eliminate the worlds *hdl*, *hdd*, and *tdd*, and remove the unlit centers in other worlds, and you'll be left with the following posterior credence distribution:

World:		*hll*	*hld*	*hdl*	*hdd*	*tll*	*tld*	*tdl*	*tdd*
Room:	1	2/10	2/10	0	0	1/10	2/10	0	0
	2	0	0	0	0	1/0	0	2/10	0

Having updated with compartmentalized conditionalization on the centered information that *your* room is lit, your new credence that the coin landed heads will be 2/5ths.

On the other hand, you might instead learn that your room is dark after noon. That information is incompatible with the worlds *hll*, *hld*, and *tll*, so you'll end up with this posterior credence distribution:

World:		*hll*	*hld*	*hdl*	*hdd*	*tll*	*tld*	*tdl*	*tdd*
Room:	1	0	0	2/10	2/10	0	0	2/10	1/10
	2	0	0	0	0	0	2/10	0	1/10

Again, having updated on the *de se* information that *your* room is dark with compartmentalized conditionalization, you will end up with a credence of 2/5ths that the coin landed heads.

This is embarassing on its own for double halfers, since their lodestar is the thought that *de se* info can't lead your credence in 'a flipped coin landed heads' to deviant from one half. But it also shows that compartmentalized conditionalization doesn't secure the principle of reflection. In *Sleeping Beauty with Some Lights*, the compartmentalized conditionalizer will violate reflection, since no matter what they learn, their credence that the coin lands heads will drop from 1/2 to 2/5ths.

Embarassment #2: Secondly, consider the following version of Sleeping Beauty,[9]

This coin lands heads Everything is as in *Sleeping Beauty*, except that you are no longer told what day it is before the coin is flipped. On both Monday and Tuesday (if you're awake), you are allowed to flip a coin. The coin you flip on Monday is the one which determines whether you are awoken again on Tuesday.

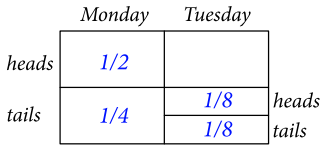
Now, consider the proposition "*this* coin lands heads", where the demonstrative 'this' is picking out the coin that you hold in your hands. This

[9] See Titelbaum, Michael G. (2012). "An Embarrassment for Double-Halfers". *Thought: A Journal of Philosophy* 1 (2):146-151.

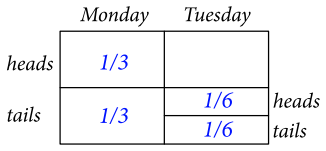
proposition is true if it is Monday and Monday's coin lands heads or if it is Tuesday and Tuesday's coin lands heads. Since it can't be both Monday and Tuesday, your credence in 'this coin lands heads' must be the sum of your credence in 'It is Monday and Monday's coin lands heads' and your credence in 'It is Tuesday and Tuesday's coin lands heads. If you're a double halfer, then your credence in 'It is Monday and Monday's coin lands heads' must be one half. But that means that, unless you're certain that Tuesday's coin won't be flipped and land heads, your credence in '*this* coin lands heads' will have to be greater than one half.

For instance, if you have the halfer's distribution from figure 8.4a, then your credence that *this* coin lands heads will be $1/2 + 1/8 = 5/8$.

On the other hand, if you have the thirder's distribution from figure 8.4b, then your credence that *this* coin lands heads will be $1/3 + 1/6 = 1/2$.



(a)



(b)

Figure 8.4: In 8.4a, the halfer's credence distribution when there's a second coin flip on Tuesday. In 8.4b, the thirder's credence distribution when there's a second coin flip on Tuesday.

8.4 Why Does it Matter?

Sleeping Beauty can seem like a philosopher's puzzle that doesn't matter—but I think that impression is mistaken. See Titelbaum's "Ten Reasons to Care about the Sleeping Beauty Puzzle" for a thorough tour of all of the puzzle's importance. For here, let me just mention one important upshot: your situation in Sleeping Beauty is not unlike the following situation:

Splitting Beauty? You have prepared an electron in a superposition of 50% x-spin up and 50% x-spin down.[10] You are about to take a measurement of the electron's x-spin. If the Everettian (or 'many worlds') interpretation of quantum mechanics is correct, then upon this observation being made, the world will split in two, and there will be one version of you (retaining all your current memories) who sees the electron with x-spin up, and another version of you (retaining all your current memories) who sees the electron with x-spin down. If, however, the GRW (or 'collapse') interpretation of quantum mechanics is correct, then there is a 50% chance that you'll see an electron with x-spin up and a 50% chance that you'll see an electron with x-spin down.

[10] More carefully, the spin state of the electron is

$$1/\sqrt{2}\,|\uparrow_x\rangle \;+\; 1/\sqrt{2}\,|\downarrow_x\rangle$$

Let's suppose that you have performed this experiment, but you've not yet observed the result. And suppose that you're 50% sure that GRW is correct, and 50% sure that Everett is correct (in which case, there will be two versions of you, one which will see $\uparrow_x$ and one which will see $\downarrow_x$. Then, you'll have the credence distribution shown below.

World:		GRW & $\uparrow_x$	GRW & $\downarrow_x$	Everett
Center:	you see $\uparrow_x$	$1/4$	0	$1/4$
	you see $\downarrow_x$	0	$1/4$	$1/4$

When you open your eyes, you are going to learn some *de se* information—you won't just learn that *there is an electron with x-spin up* (e.g.), since

this is true even at the centered world in which Everettian QM is correct and the electron in front of *you* has x-spin down. You will instead learn the stronger *de se* information that the electron in front of *you* has x-spin up.

If conditionalization is the correct rule, then learning whether the electron is x-spin up or down won't confirm or disconfirm Everettian quantum mechanics over GRW. But, if compartmentalized conditionalization is correct, then Everett will be confirmed *no matter what*. The situation here is exactly the situation we faced with the situation *Sleeping Beauty with Some Lights* above. Learning that the electron in front of us has x-spin up will eliminate the world GRW & $\downarrow_x$. Conditionalizing on the purely *de dicto* information that this world is ruled out (as compartmentalized conditionalization would have us do), we will end up $2/3$ confident in Everettian quantum mechanics, and only $1/3$ confident in GRW.

World:		GRW & $\uparrow_x$	GRW & $\downarrow_x$	Everett
Center:	you see $\uparrow_x$	$1/3$	0	$2/3$
	you see $\downarrow_x$	0	0	0

And the same thing will happen if we instead learn that the electron in front of us has x-spin down:

World:		GRW & $\uparrow_x$	GRW & $\downarrow_x$	Everett
Center:	you see $\uparrow_x$	0	0	0
	you see $\downarrow_x$	0	$1/3$	$2/3$

There are a lot of other kinds of cases to consider when thinking about how well Everettian QM fares according to the Bayesian theory of confirmation with various *de se* update rules, but the important thing to emphasize here is just that the *de se makes a difference* to how well confirmed different fundamental physical theories are (not just in the case of interpretations of quantum mechanics, but also with respect to other cosmological theories, since those theories will also make a difference to how likely our *de se* information is).

8.5 Dutch Strategy Arguments

Here's a simple Dutch strategy argument against the thirder: on Sunday, before they go to sleep, buy bet #1 off of them. After they awake on Monday, sell them bet #2. The combination of these two bets will lose them money no matter what (the table shows the thirder's net profit):

Bet #1

\$30	if the coin lands tails
\$0	else

price: \$15

Bet #2

\$30	if the coin lands tails
\$0	else

price: \$20

	the coin lands heads	the coin lands tails
Net profit from selling bet 1	\$15	−\$15
Net profit from buying bet 2	−\$20	\$10
Overall net profit	−\$5	−\$5

As Christopher Hitchcock points out,[11] this Dutch strategy argu-

[11] See Hitchcock, Christopher (2004). "Beauty and the bets". *Synthese* 139 (3):405 - 420.

ment is fallacious—for it relies upon you (the person selling the bets) having information which the thirder does not. In particular, it requires you knowing what day of the week it is. This *de se* knowledge is what allows you to only sell bet #2 *once* (on Monday), and not sell it again on Tuesday.

To appreciate this, suppose that you are forced to undergo the same experiment as the thirder (being put to sleep and having your memory erased if the coin lands tails). Then, you would end up selling the thirder bet #2 *twice* if the coin lands tails, and only *once* if the coin lands heads. So the thirder's net profit would be the following:

	the coin lands heads	the coin lands tails
Net profit from selling bet 1	\$15	−\$15
Net profit from buying bet 2	−\$20 (× 1)	\$10 (× 2)
Overall net profit	−\$5	\$5

This is not a guaranteed loss—in fact, the combination of these bets has an expected payout of \$0 (that is to say, this combination of bets is fair).

Hitchcock points out, however, that if we force you (the bookie selling the bets) to go through the experiment along with the *halfer*, then they will buy bet #1 off of you beforehand, on Sunday. And they will buy bet #3 off of you *once* if the coin lands heads, but *twice* if the coin lands tails:

Bet #3

\$20	if the coin lands heads
\$0	else

price: \$10

	the coin lands heads	the coin lands tails
Net profit from buying bet 1	−\$15	\$15
Net profit from buying bet 3	\$10 (× 1)	−\$10 (× 2)
Overall net profit	−\$5	−\$5

So, Hitchcock alleges, it is in fact the *halfer* who is susceptible to a Dutch strategy, and not the thirder.

R.A. Briggs has an interesting response to this argument: they suggest that it relies upon an implicit committment to *causal decision theory* (as opposed to evidential decision theory). Why? Because there is a *non-causal correlation* between how you act on Monday and how you act on Tuesday (if you awake on Tuesday). Since you will awake on Tuesday in exactly the state you previously awoke on Monday, you will behave in exactly the same way on Tuesday that you behaved on Monday—not because how you behave on Monday *causes* your behavior on Tuesday, but rather because your behavior on Monday and your behavior on Tuesday have a *common cause*: your mental states on Sunday.

Let's see how this makes a difference to how causalists and evidentialists will evaluate the instrumental value of taking bet #3 on Monday. A *causalist* will reason as follows: whether I take bet #3 *today* doesn't make any causal difference to whether I will take bet #3 on any other day. If the coin lands tails and there are two wakings, then, if I in fact take bet #3 on the other waking, then I would *still* take bet

#3 on that other waking if I were to refuse it today. And if I in fact refuse bet #3 on the other waking, then I would *still* refuse bet #3 on the other waking.

Let's go through that carefully, just to see how the calculation of the expectation works out. Suppose that the halfer is a causal decision theorist, and they've bought bet #1 on Sunday. Today, they find themselves offered bet #3. There are at that point 3 different possible outcomes, if they were to buy bet #3: it could be that the coin lands heads, in which case they'd net −$5. It could be that the coin lands tails and they buy bet #3 on the other waking, in which case they'd net −$5, and it could be that they do *not* buy bet #3 on the other waking, in which case they'd net $5. So the causal expected utility of buying bet #3 is given by:

$$\begin{aligned}
CEU(buy) &= \sum_{o \in O} \mathbb{C}(buy \mathbin{\square\!\!\rightarrow} o) \cdot \mathcal{V}(o) \\
&= \mathbb{C}(buy \mathbin{\square\!\!\rightarrow} -\$5) \cdot \mathcal{V}(-\$5) + \mathbb{C}(buy \mathbin{\square\!\!\rightarrow} \$5) \cdot \mathcal{V}(\$5) \\
&= \mathbb{C}(heads \vee tails \mathbin{\&} bought) \cdot \mathcal{V}(-\$5) + \mathbb{C}(tails \mathbin{\&} \neg bought) \cdot \mathcal{V}(\$5) \\
&= [(1+\beta)/2] \cdot -5 + [(1-\beta)/2] \cdot 5 \\
&= -5\beta
\end{aligned}$$

(where '*bought*' is the proposition that they have or will buy bet #3 on the other waking, and I'm using 'β' for their credence that they will buy bet #3 now.)

On the other hand, there are 2 different possible outcomes if they reject bet #3: it could be that the coin lands heads, in which case they lose $15. It could be that the coin lands tails and they buy bet #3 on the other waking, in which case they net $5. And it could be that the coin lands tails and they don't buy bet #3 on the other waking, in which case they net $15. So the causal expected utility of not buying bet #3 is:

$$\begin{aligned}
CEU(\neg buy) &= \sum_{o \in O} \mathbb{C}(\neg buy \mathbin{\square\!\!\rightarrow} o) \cdot \mathcal{V}(o) \\
&= \mathbb{C}(\neg buy \mathbin{\square\!\!\rightarrow} -\$15) \cdot \mathcal{V}(-\$15) + \mathbb{C}(\neg buy \mathbin{\square\!\!\rightarrow} \$5) \cdot \mathcal{V}(\$5) + \mathbb{C}(\neg buy \mathbin{\square\!\!\rightarrow} \$15) \cdot \mathcal{V}(\$15) \\
&= \mathbb{C}(heads) \cdot \mathcal{V}(-\$15) + \mathbb{C}(tails \mathbin{\&} bought) \cdot \mathcal{V}(\$5) + \mathbb{C}(tails \mathbin{\&} \neg bought) \cdot \mathcal{V}(\$15) \\
&= 1/2 \cdot (-15) + 1/2 \cdot \beta \cdot 5 + 1/2 \cdot (1-\beta) \cdot 15 \\
&= -15/2 + 5/2 + 15/2 - 15\beta/2 \\
&= -10\beta/2 \\
&= -5\beta
\end{aligned}$$

So the causal decision theorist will consider taking bet #3 fair—it has exactly the same causal expected utility as not buying bet #3. So the causalist halfer is subject to Hitchcock's Dutch strategy.

On the other hand, suppose that the halfer is an *evidentialist*. Then, they will not fall prey to the Dutch strategy, for they will see bet #3 as a loser. To calculate the *evidential* expected utility of buying bet #3,

recall, they will use the expectation:

$$\begin{aligned}
EEU(buy) &= \sum_{o \in O} \mathbb{C}(o \mid buy) \cdot \mathcal{V}(o) \\
&= \mathcal{V}(-\$5 \mid buy) \cdot \mathcal{V}(-\$5) + \mathcal{V}(\$5 \mid buy) \cdot \mathcal{V}(\$5) \\
&= \mathbb{C}(heads \vee tails \,\&\, bought \mid buy) \cdot \mathcal{V}(-\$5) + \mathbb{C}(tails \,\&\, \neg bought \mid buy) \cdot \mathcal{V}(\$5) \\
&= 1 \cdot (-5) + 0 \cdot 5 \\
&= -5
\end{aligned}$$

And to calculate the evidential expected utility of not buying, they will use the expectation:

$$\begin{aligned}
EEU(\neg buy) &= \sum_{o \in O} \mathbb{C}(o \mid \neg buy) \cdot \mathcal{V}(o) \\
&= \mathbb{C}(-\$15 \mid \neg buy) \cdot \mathcal{V}(-\$15) + \mathbb{C}(\$5 \mid \neg buy) \cdot \mathcal{V}(\$5) + \mathbb{C}(\$15 \mid \neg buy) \cdot \mathcal{V}(\$15) \\
&= \mathbb{C}(heads \mid \neg buy) \cdot \mathcal{V}(-\$15) + \mathbb{C}(tails \,\&\, bought \mid \neg buy) \cdot \mathcal{V}(\$5) + \mathbb{C}(tails \,\&\, \neg bought \mid \neg buy) \cdot \mathcal{V}(\$15) \\
&= 1/2 \cdot (-15) + 0 \cdot 5 + 1/2 \cdot 15 \\
&= -15/2 + 15/2 \\
&= 0
\end{aligned}$$

So the *evidentialist* halfer will not fall prey to Hithcock's Dutch strategy.

Moreover, Briggs shows that an evidentialist *thirder* will fall prey to a Dutch strategy. Again, they will sell you bet #1 on Sunday (netting them \$15 if the coin lands heads and losing them \$15 if the coin lands tails). And, on every waking event, they will buy from you bet #4.

First, let's convince ourselves that this constitutes a Dutch strategy. And later, we'll verify that the evidentialist thirder will purchase bet #4 whenever it's offered. Here are the possible payouts for the evidentialist thirder from this combination of bets:

Bet #4

\$25	if the coin lands tails
\$0	else

price: \$20

	the coin lands heads	the coin lands tails
Net profit from selling bet 1	\$15	−\$15
Net profit from buying bet 4	−\$20 (× 1)	\$5 (× 2)
Overall net profit	−\$5	−\$5

So, if the evidentialist thirder will buy bet #4 whenever it's offered, then they will be susceptible to Briggs' Dutch strategy. But why think they will buy this bet whenever it's offered? Because they will evaluate the bet by paying attention to the non-causal correlations, as follows:

$$\begin{aligned}
EEU(buy) &= \sum_{o \in O} \mathbb{C}(o \mid buy) \cdot \mathcal{V}(o) \\
&= \mathbb{C}(-\$5 \mid buy) \cdot \mathcal{V}(-\$5) + \mathbb{C}(-\$10 \mid buy) \cdot \mathcal{V}(-\$10) + \mathbb{C}(-\$5 \mid buy) \cdot \mathcal{V}(-\$5) \\
&= \mathbb{C}(heads \vee tails \,\&\, bought \mid buy) \cdot \mathcal{V}(-\$5) + \mathbb{C}(tails \,\&\, \neg bought) \cdot \mathcal{V}(-\$10) \\
&= 1 \cdot (-5) + 0 \cdot (-10) \\
&= -5
\end{aligned}$$

whereas they'll evaluate not buying the bet with:

$$\begin{aligned} EEU(\neg buy) &= \sum_{o \in O} \mathbb{C}(o \mid buy) \cdot \mathcal{V}(o) \\ &= \mathbb{C}(\$15 \mid \neg buy) \cdot \mathcal{V}(\$15) + \mathbb{C}(-\$15 \mid \neg buy) \cdot \mathcal{V}(-\$15) \\ &= 1/3 \cdot 15 - 2/3 \cdot 15 \\ &= -5 \end{aligned}$$

So they will see taking bet #4 as fair. So the *evidentialist* thirder is subject to Briggs' Dutch strategy.

Briggs goes on to show something stronger: the evidentialist halfer is *immune* to Dutch strategies. There is no Dutch strategy against the evidentialist halfer. And the causalist thirder is likewise immune to Dutch strategies.

We can summarize this as follows: if you're an evidential decision theorist, then there is a Dutch strategy argument for being a halfer. And, if you are a causal decision theorist, then there is a Dutch strategy argument for being a thirder. So Dutch strategies cannot be used *on their own* to settle whether to be a halfer or a thirder. We will also need to settle cases of act-state dependence like Newcomb's problem. But, once we're decided about the right solution to decisions like Newcomb's problem, we will have a Dutch strategy argument for one answer or another to Sleeping Beauty.

8.6 Expected Inaccuracy Arguments

What about considerations of expected inaccuracy? Can we use them to settle the Sleeping Beauty puzzle? Again, matters are complicated.[12] The issue is that there are two ways of calculating expected inaccuracy of your update plans in the case: you could care about your *total* inaccuracy—adding up the inaccuracy the update plans upon each awakening—or you could instead care about your *average* inaccuracy—adding up the inaccuracy of the update plans on each awakening, and dividing through by the total number of awakenings.

[12] See Kierland, Brian & Monton, Bradley (2005). "Minimizing Inaccuracy for Self-Locating Beliefs". *Philosophy and Phenomenological Research* 70 (2): 384-395.

Let's see how this goes with the Quadratic (or 'Brier') measure. Suppose that you care about *total* inaccuracy. Then, the expected total inaccuracy of a plan to have credence x in the proposition that the coin lands heads will be given by:

$$\mathbb{C}(heads) \cdot (1-x)^2 + \mathbb{C}(tails) \cdot [x^2 + x^2] = \frac{(1-x)^2}{2} + x^2$$

This expected inaccuracy is minimized at $x = 1/3$.[13] So, if you care about minimizing your *total* expected inaccuracy, then you will be a thirder.

[13] To appreciate this, take the first-order condition:

$$\begin{aligned} -(1-x) + 2x &= 0 \\ 3x &= 1 \\ x &= 1/3 \end{aligned}$$

You should still check the second-order and boundary conditions, but those do work out, and this is the unique global minimum.

However, suppose that you instead care about *average* inaccuracy. Then, the expected inaccuracy of aplan to have credence x in the propo-

sition that the coin lands heads will be given by

$$\mathbb{C}(\textit{heads}) \cdot (1-x)^2 + \mathbb{C}(\textit{tails}) \cdot \frac{[x^2 + x^2]}{2} = \frac{(1-x)^2}{2} + \frac{x^2}{2}$$

And this expected inaccuracy is minimized at $x = 1/2$.[14]

So it looks like we again have an argument for both conclusions—it depends upon whether we should care about minimizing our *total* inaccuracy or whether we should instead care about minimizing our *average* inaccuracy. The former option gives us an argument for thirding, whereas the latter option gives us an argument for halfing.

[14] To appreciate this, take the first-order condition:

$$\begin{aligned} -(1-x) + x &= 0 \\ 2x &= 1 \\ x &= 1/2 \end{aligned}$$

Again, you should verify that the second derivative is positive and check the boundary conditions, but this is the global minimum.

Review Questions

1. What is the Sleeping Beauty puzzle? Describe three answers to the Sleeping Beauty puzzle.
2. What is *Compartmentalized Conditionalization* say, and which answer to the Sleeping Beauty puzzle does it imply? Raise two objections to *Compartmentalized Conditionalization*.
3. Suppose you think that, if you are susceptible to a Dutch strategy, then you are irrational. Then, explain why your answer to Newcomb's Problem is going to constrain your answer to the Sleeping Beauty problem.
4. Explain why expected accuracy arguments don't settle the Sleeping Beauty problem.