# 4

# CONDITIONAL PROBABILITIES

*Kenny Easwaran*

Conditional probability is one of the central concepts in probability theory. Some notion of conditional probability is part of every interpretation of probability. The basic mathematical fact about conditional probability is that $p(A \mid B) = p(A \wedge B)/p(B)$ where this is defined. However, while it has been typical to take this as a definition or analysis of conditional probability, some (perhaps most prominently Hájek, 2003) have argued that conditional probability should instead be taken as the primitive notion, so that this formula is at best coextensive, and at worst sometimes gets it wrong.

Section 1.1 considers the concept of conditional probability in each of the major families of interpretation of probability. Section 1.2 considers a conceptual argument for the claim that conditional probability is prior to unconditional probability, while Section 1.3 considers a family of mathematical arguments for this claim, leading to consideration specifically of the question of how to understand probability conditional on events of probability 0. Section 1.4 discusses several mathematical principles that have been alleged to be important for understanding how probability 0 behaves, and raises a dilemma for probability conditional on events of probability 0. Section 2 and Section 3 take the two horns of this dilemma and describe the two main competing families of mathematical accounts of conditional probability for events of probability 0. Section 4 summarizes the results, and their significance for the two arguments that conditional probability is prior to unconditional probability.

## 1 BACKGROUND

### 1.1 *What is Conditional Probability?*

Before considering the arguments suggesting that conditional probability is a primitive notion (either equal to unconditional probability in fundamentality, or perhaps even more basic), we should consider just what conditional probability is.

Some have argued, following some cryptic remarks of Frank Ramsey, that conditional probability can be understood as the probability of a conditional. However, without a clear interpretation of what a conditional means, this provides little help for clarifying the concept of conditional probability. There are deep difficulties with this identification, since to-

gether with certain plausible logical principles for conditionals, it entails various triviality results about unconditional probability. (Edgington, 1995, summarizes much of this literature and argues that there is some interpretation of the conditional that allows for this identification, and Bacon, 2015, shows how much logic for conditionals can be preserved.) At any rate, the defenders of this principle hope to use conditional probability to clarify the meaning of conditionals, rather than vice versa. Since the meaning of a conditional has so much obscurity, this identification is of no help in trying to analyze the meaning of conditional probability.

Perhaps a more useful (and also Ramsey-inspired) way to think of conditional probability is to look at some of the roles it plays in order to see what features it needs to have. But since there are many different phenomena that have all been said to be interpretations of probability, and conditional probability plays different roles in each, I will break this consideration down into several parts. In this discussion, I will not consider each separate interpretation of probability, but I will instead consider them in three broad families. (For more on specific interpretations, see Hájek, 2007.)

The first family (which I will use as my primary reference point in much later discussion) is the set of broadly "Bayesian" interpretations that treat probability as some sort of informational state. The second family is the set of broadly "physical" interpretations that treat probability as a feature of some part of the world itself, rather than an information state. The third family is the set of "mathematical" applications of probability, some of which I don't think rise to the level of an interpretation, but are worth mentioning separately.

### 1.1.1 *Bayesian Interpretations*

Among the interpretations I am calling "Bayesian" are both various objective and subjective notions. I mean this class to include "logical probabilities" (Keynes, 1921; Carnap, 1950; Maher, 2006) and "evidential probabilities" (Williamson, 2002), as well as the more familiar objective and subjective Bayesian interpretations of probability as some sort of rational degree of belief (Easwaran, 2011a, 2011b). These interpretations of probability are used in a broad variety of applications in psychology, economics, decision theory, philosophy of science, and epistemology.

However, in all of these applications, it seems that there are three main roles that conditional probability is said to play. First, conditional probability is said to play some sort of fairly direct role in constraining the way that probabilities change over time. Second, conditional probability is used in the analysis of various measures of confirmation (which often claim to describe the potential value of various pieces of information,

whether or not anyone ever gains that information). And third, conditional probability is important in certain accounts of decision theory. If there are roles for conditional probability other than these, then some of my later evaluation of the different mathematical accounts of conditional probability may need to be modified.

The role of conditional probability in updating is perhaps the most familiar one. The traditional notion of Bayesian updating is said to occur when there is some new evidence $E$ that the agent gains with certainty. In this case, the probability function after the update $p_{\text{new}}$ and the probability function before the update $p_{\text{old}}$ are said to satisfy, for every $A$, $p_{\text{new}}(A) = p_{\text{old}}(A \mid E)$. Following Jeffrey (1965), many have thought that this sort of update scenario is implausible, because there is never any particular evidence that is gained with certainty. Instead, there is said to be an evidential partition $\mathbf{E}$, which is a set of propositions $\{E_i : i \in I\}$, such that it is antecedently certain that there is exactly one $i$ such that $E_i$ is true. No member of this partition becomes certain, but their probabilities change in a way that drives the change of all other propositions. This notion of "driving the change" is summarized by a constraint known as *rigidity*: for any $A$, $p_{\text{new}}(A \mid E_i) = p_{\text{old}}(A \mid E_i)$. The specification of these conditional probabilities is said to be enough, in conjunction with the new probabilities for each $E_i$, to specify the new probability function uniquely, by means of the Law of Total Probability. When the partition is finite, this takes the form $p(A) = \sum p(E_i)p(A \mid E_i)$, though in the infinite case we need to be a bit more careful. As I will discuss in Section 1.4.3, the natural way to generalize this will be notated as $p(A) = \int p(A \mid \mathbf{E_I})\, \mathrm{d}p$, though further complexities arise.

At least since the work of Hosiasson-Lindenbaum (1940), conditional probability has also been very important in analyzing the notion of confirmation. Much of this literature has focused on finding numerical measures of the degree to which particular evidence would support particular hypotheses. Where $H$ is some hypothesis, and $E$ is some potential evidence, some well-known measures are said to take the value $p(H \mid E) - p(H)$, or $p(H \mid E)/p(H)$, or $p(E \mid H)/p(E \mid \neg H)$. (These and other measures are discussed by Fitelson, 1999.) The probabilities that show up in these formulations are of four types. There are two unconditional probabilities, $p(E)$ and $p(H)$, which are called "priors" for the evidence and the hypothesis respectively. (Priors for their negations sometimes appear as well, but since $p(\neg E) = 1 - p(E)$ and $p(\neg H) = 1 - p(H)$ these are not relevantly different.) There are also two types of conditional probability that arise. $p(H \mid E)$ is called the "posterior" of the hypothesis, because (according to the update rule mentioned above), it gives the probability the hypothesis would have after hypothetically learning the evidence. And $p(E \mid H)$ and $p(E \mid \neg H)$ are called "likelihoods" of the hypothesis and its negation. Some

philosophers have focused on measures involving only likelihoods, because they are said to be more objective than priors and posteriors (Royall, 1997). But at any rate, these are the conditional probabilities whose values are relevant to confirmation.

In decision theory, the most traditional analysis of the value of an action doesn't depend on conditional probability at all (Savage, 1954). There are said to be a set **A** of actions available to the agent and a set **S** of possible states of the world independent of the agent, and together these are said to determine outcomes of the act. The agent has a value $V(A \wedge S)$ for each outcome. When everything is finite, the value of an act $A \in \mathbf{A}$ is given by $V(A) = \sum_{S \in \mathbf{S}} p(S)V(A \wedge S)$. (Again, when **S** is infinite, things are more complicated, as will be discussed in Section 1.4.2.) However, Jeffrey (1965) and others have worried about cases in which one can't identify states of the world independent of the agent. In this case, Jeffrey suggests that we should have $V(A) = \sum_{S \in \mathbf{S}} p(S \mid A)V(A \wedge S)$, replacing the unconditional probability of a state with its probability conditional on each action. Joyce (1999) and other "causal decision theorists" have argued that this "evidential decision theory" is wrong for certain cases, and replace the conditional probability $p(S \mid A)$ with something like $p(A \mathrel{\square\!\!\rightarrow} S)$, the probability of the subjunctive conditional. Regardless of how this is to be interpreted, the relevant conditional probabilities for decision theory are what I will call "action probabilities," and they must be defined for states of the world conditional on the possible acts of an agent.

Thus, on the Bayesian interpretations of probability, the conditional probabilities that arise in any relevant application appear to be of three forms—posteriors, likelihoods, and action probabilities. Posteriors must be defined for every *hypothesis* conditional on every piece of possible *evidence* (for confirmation theory), or for *every* proposition conditional on every piece of possible evidence (for updating). Likelihoods must be defined for every piece of possible evidence conditional on every hypothesis. And action probabilities must be defined for every state of the world conditional on every possible action. (On Jeffrey's interpretation, action probabilities may just be a special case of posteriors, since the role of an act for him is in some sense as a special piece of evidence, but for Joyce and others the role is somewhat different, though it may not even be a conditional probability in the traditional sense.) In each case, the set of things that may be conditioned on form a "partition"—they are a set of propositions such that it is certain in advance that exactly one of them is true. This fact will be significant for later discussion.

### 1.1.2 *Physical Interpretations*

Another family of interpretations of probability take probability to be something separate from any sort of information state. One historically influential such interpretation is Popper's account of chance as a sort of "propensity" of the world to evolve in a certain way (Popper, 1959b). Many statisticians have wanted some sort of objective physical notion of probability like this, but without the metaphysical baggage. This has given rise to frequentist statistical practice, described for instance by Mayo and Cox (2006), on which the relevant probabilities are the proportion of cases in which particular outcomes "would arise in a hypothetical long-run of repeated sampling" (p. 79).

These interpretations are possibly more heterogeneous than the Bayesian ones I discussed above, but we can still identify particular families of uses to which conditional probabilities are put. First, conditional probabilities are sometimes said to govern the way in which chances change over time. Second, conditional probabilities are sometimes used to analyze notions of causation or independence. Third, there are various uses conditional probabilities are put to in frequentist statistical practice. And fourth, there may be a relevant notion of expected value computed from physical probabilities.

For changing chances, David Lewis claims that "a later chance distribution comes from an earlier one by conditionalizing on the complete history of the interval in between" (1980, p. 280). That is, if $p_{\text{old}}$ is the probability function giving the chances at some earlier time and $p_{\text{new}}$ gives the chances at a later time, and $H$ is the history of all events that occur between these two times, then for any $A$, $p_{\text{new}}(A) = p_{\text{old}}(A \mid H)$. This requires a notion of probability conditional on any $H \in \mathbf{H}$, where $\mathbf{H}$ is the set of all histories that could transpire between one time and another.

Some analyses of causation have said that $A$ is a cause of $B$ iff $p(B \mid A) > p(B)$, where $p$ is some physical notion of probability. There are many obvious problems with this account, turning on cases where there are common causes (the probability of a parent having blond hair given that a child has blond hair is higher than the unconditional probability of a parent having blond hair, even though the child's hair color is not a cause of the parent's), other events intervening (the probability of getting in a car crash given that you've had a drink may be lower than the unconditional probability of getting in a car crash, if drinking makes you less likely to drive, even though drinking does tend to cause car crashes), and similar sorts of problems. Sophisticated versions of this theory have now turned to the sort of "causal modeling" developed by Pearl (2000) and Spirtes, Glymour, and Scheines (2000). On this picture, events $A$ and $B$ are taken to be particular values of variables $\mathbf{A}$ and $\mathbf{B}$, which may have two values

($A$ occurs or does not occur) or more (if $A$ is seen as one of a class of ways for something to happen). These variables are represented by nodes in a graph with arrows connecting some nodes to others. Physical probabilities are given by a probability distribution for the values of one variable conditional on each specification of the values of the variables with arrows pointing directly to it. There are then *two* notions of conditional probability, depending on whether we "intervene" on one variable or merely "condition" on it (Meek & Glymour, 1994). This difference can be seen by considering the probability of someone having a sun tan given that their vitamin D levels are high—conditioning involves looking at people with *currently* high levels of vitamin D and measuring their tan, while intervening involves artificially *giving* people high levels of vitamin D and measuring their tan. Variable **A** is then said to be a cause of **B** iff intervening on **A** in different ways gives different conditional distributions for **B**, and is said to be independent if the conditional probabilities are the same. (Vitamin D likely turns out not to be a cause of sun tan, but to have correlation due to common cause.) Again, the relevant probabilities always involve conditioning on the elements of a partition. For far more on this, see Hitchcock (2010).

In frequentist statistical practice, there are a variety of conditional probabilities that arise. One of the most well-known such conditional probabilities is the $p$-value of a piece of evidence. This is the frequency with which evidence *at least as extreme* as the observed value would occur in hypothetical repetitions of the same experimental protocol, assuming that the "null hypothesis" is correct. We might notate this as $p(E^+ \mid H_0)$, where $E^+$ is the event of evidence at least as extreme being observed, and $H_0$ is the null hypothesis (though see Section 1.2 for discussion of whether this should really be thought of as a conditional probability). The $p$-value is often used as a criterion for statistical rejection, and it is common to reject the null hypothesis (in favor of some alternative) if the $p$-value falls below some pre-arranged threshold. The "power" of a statistical test is said to be the frequency with which the same experimental protocol would result in rejection of the null hypothesis, assuming that the alternative is in fact true. We might think of this as $p(R \mid H')$, where $H'$ is the alternative to the null hypothesis, and $R$ is the event of an experimental result that our protocol recommends rejection on. In statistical tests for which we want to estimate the value of some unknown parameter, our experimental protocol often ends not with rejection, but with specification of a "confidence interval." For instance, a 95% confidence interval is the set of parameter values for which the $p$-value would be at least .05 if that value were treated as the null—we can think of the confidence interval as the set of values that wouldn't be rejected at a given $p$-level. These probabilities are not the same as the likelihoods discussed above for Bayesian probabilities (because these

are not probabilities of the actually observed evidence, but rather of the event "an observation at least as extreme would occur"), but they are still probabilities conditional on each hypothesis.

Finally, although many contemporary decision theorists follow Savage (1954) in using some sort of Bayesian probability as the basis of computation of expected value, von Neumann and Morgenstern (1947) use a physical probability as their basis for a theory of rational decisions. Similar issues involving objective correlations between "states of the world" and an agent's actions might motivate some use of conditional probability in calculations of expected value, and these will be like the "action probabilities" I mentioned above.

Again, in all cases, the set of things that can be conditioned on forms a partition.

### 1.1.3 *Mathematical Interpretations*

There are some other interpretations of probability that don't quite fit in with those mentioned above. The most interesting such interpretation is that of probability as actual relative frequency. For instance, the World Health Organization reports that 68% of deaths worldwide in 2012 were due to non-communicable diseases, such as cancer, diabetes, and cardiovascular diseases. We can interpret this as a sort of probability, and say that the probability that a person who died in 2012 died of a non-communicable disease is .68. On this interpretation, for any descriptions $A, B$, we can say that $p(B \mid A)$ is the fraction of things fitting description $A$ that also fit description $B$. Any description whatsoever can be used in either position, provided that there is a meaningful way to count instances of each.

This bears much similarity to the "classical interpretation" of probability attributed by Hájek (2007) to early probability theorists. The idea again is that in many traditional games of chance, physical probabilities or Bayesian probabilities may be usefully approximated by counting all the different possible outcomes of the game and seeing how many of them are of the sort of interest.

Tools like this have also been applied in pure mathematics, in what is called the "probabilistic method." This method was introduced by Erdős (1947) to derive bounds for Ramsey numbers. (These numbers were first investigated by Ramsey, 1930, in an attempt to work on the decision problem for logic, but have since been generalized to the size of any sort of structure that is needed to guarantee the existence of subsets with given complexity.) Erdős considers the complete graph on $n$ vertices where edges are arbitrarily colored in two colors. He then defines a probability function on subsets of this graph, and shows that if $n$ is large enough, then the probability of selecting $k$ vertices at random such that all edges between

them are the same color is non-zero. In particular, this means that for any coloring of the graph on $n$ vertices, there must be $k$ vertices whose edges are all the same color. The importance of Erdős' result is that the bound he arrived at for $n$ is substantially smaller than that arrived at by Ramsey, and is in most cases still the best known. This method has since been deployed in many other problems in combinatorics.

The classic applications of this method don't make any use of conditional probability. More advanced applications might, but in general, the interpretation of the probability function is not really of any interest. Instead, the probabilities (and perhaps any conditional probabilities) are just tools for mathematical computation. Any mathematical account of "conditional probability" could be useful, whether or not it has any application to other interpretations of probability. Thus, this interpretation of probability gives no particular constraint to our theorizing about conditional probability, and if anything, encourages us to explore as many different mathematical accounts as possible, in case one is of use in some mathematical problem or other.

### 1.2 *Backgrounds vs. Conditions*

There are two main families of argument that all probabilities must really be conditional. One family of argument (considered in this section) is conceptual, and claims that for many different interpretations, some sort of background is essential to even determine probabilities. The second family of argument (considered in Section 1.3) is mathematical, and uses problems involving division by zero to argue that conditional probability must be prior to unconditional probability. Although the mathematical arguments are sometimes clearer and seem more convincing, I will consider the conceptual arguments first, since the mathematical arguments lead more naturally to the issues that arise in the rest of this article. This section is independent of the rest of the article, and can be skipped by readers more interested in the mathematical issues.

This section considers the claim that a background is essential to the possibility of probability. I will consider versions of this argument for each interpretation of probability, and argue that for most interpretations of probability, this "background" is different enough in kind from the sort of thing that one can conditionalize on, that it should be treated separately from conditional probability. I claim that only for probabilities thought of as actual frequencies is it correct to say that every probability requires a background, and that this background makes every probability essentially a conditional probability. For some of the other interpretations, we will at least find that many numbers traditionally thought of as unconditional probabilities may be better thought of as conditional probabilities, but for

all of these other interpretations there is conceptual room to argue that some probabilities really are unconditional.

For this argument, again it will be useful to consider different interpretations of probability in some detail. However, I will skip a few of the most purely mathematical interpretations for which there are no important conceptual requirements, and will consider the other interpretations in somewhat different connections than I did before.

### 1.2.1 *Degree of Belief*

For subjective degree of belief, some have argued that all probabilities are really conditional on a background. I will argue that the role of the background is different from the role of the conditioning event in conditional probability. De Finetti (1974) says "every evaluation of probability is conditional; not only on the mentality or psychology of the individual involved, at the time in question, but also, and especially, on the state of information in which he finds himself at that moment" (p. 134). That is, rather than representing a subject $S$'s degree of belief at $t$ in a proposition $A$ as $p(A)$, many authors suggest that it should be represented as $p(A \mid K_{S,t})$, where $K_{S,t}$ is the conjunction of all the propositions that $S$ knows at $t$.

However, if it is possible (or reasonable, or rational) for different subjects with the same knowledge to have different degrees of belief, then including the knowledge as a proposition in an expression of conditional probability doesn't address the fundamental issue. There would not be one single probability function such that conditionalizing it on the knowledge that each subject has at each time yields the degrees of belief that agent does or should have. While the "information" may be a proposition of the same sort as the bearers of probability, the "mentality or psychology of the individual" is not.

Thus, unless we assume that the knowledge an agent has uniquely determines the probabilities that are rationally permitted for her (a thesis known as *Uniqueness*, contrasted with its negation, *Permissivism*; see Kopec and Titelbaum, 2016), it seems more accurate to represent a subject $S$'s degrees of belief at a time $t$ as $p_{S,t}(A)$. There is a separate Bayesian probability function for each subject at each time. This probability function will reflect an agent's knowledge, which may mean that it gives probability 1 to any proposition that is known. If this is the right way to treat knowledge, then $p_{S,t}(A) = p_{S,t}(A \mid K_{S,t})$. But the conditional probability is no more fundamental here.

However, some philosophers, such as Horowitz and Dogramaci (2016), argue that the knowledge or evidence that one has does uniquely determine the rational degrees of belief to have. On this picture, the degrees of belief that are rational for a subject at a time really do turn out to be a

matter of conditional probability, $p_{\text{rational}}(A \mid K_{S,t})$. What the Subjectivist Bayesians think of as a subject-and-time-relative unconditional probability is actually aimed at following an objective conditional probability function. However, even on this interpretation, there is an important theoretical consideration of what the rational degrees of belief would be for an agent with no knowledge whatsoever. The defender of the claim that conditional probabilities are fundamental would represent this as $p_{\text{rational}}(A \mid T)$, where $T$ is some tautology, but it seems just as reasonable to represent this as $p_{\text{rational}}(A)$, so that there are some unconditional probabilities after all. The question then becomes: do the unconditional rational probabilities suffice to determine all the conditional rational probabilities? But this is largely a mathematical question, and not a conceptual one, and this is the fundamental question behind Section 1.3 and Section 1.4, with full theories described in Section 2 and Section 3.

I should also note that there is a view like this one available for a more permissive or subjectivist viewpoint. This viewpoint is associated with the work of Isaac Levi (1980). There is no one objectively rational evidential probability function. Instead, there are just many different "confirmational commitments" that one might have. When this confirmational commitment is conditionalized on the knowledge a subject has, we can find the degrees of belief that the subject is committed to. Thus, what I referred to above as $p_{S,t}(A)$ would instead be referred to as $p_C(A \mid K_{S,t})$, where $C$ is the particular confirmational commitment the agent has. A major advantage this view has, if correct, is that it allows us to extend Bayesian updating to cases in which one revises one's beliefs by giving up something that was taken as evidence, by removing this proposition from one's knowledge. However, this view also requires such hypothetical revisions to yield well-defined commitments for giving up *any* of one's beliefs. And again, there may still be unconditional probabilities on this view (namely, the commitments one has prior to any evidence), though there is still a mathematical question of whether they suffice to determine the conditional probabilities that we usually focus on.

### 1.2.2 *Chance and Frequentism*

Some have argued that for the chance or frequency interpretation of probability, the role of experimental setup or preconditions for repeatability mean that all chance is conditional. I will again argue that the role of the background here is distinct from the role of the conditioning event in conditional probability, so that these interpretations also have no conceptual reason for making conditional probability prior to unconditional.

On one picture, chances are relative to a world and a time (Lewis, 1980). Thus, the chance of $A$ at a time $t$ in world $w$ is fundamentally given by

$p_{w,t}(A)$. Chances may update by conditionalization, so that if $t'$ is later than $t$, then $p_{w,t'}(A) = p_{w,t}(A \mid H_{t,t'})$, where $H_{t,t'}$ is the description of the complete history of the world from $t$ to $t'$. If there is some earliest time 0, then one may even be able to say that $p_{w,t}(A) = p_{w,0}(A \mid H_{0,t})$, so that the chances at all later times are fundamentally given by the conditional chances at the beginning of time. But this still leaves unconditional chances at the earliest time. And if there is no earliest time, then it seems that we must allow unconditional chances at *every* time to count as equally fundamental, because there is no privileged earlier reference point from which they are all conditionalized. And on any of these pictures, the world must enter in as a separate background parameter distinct from the things conditionalized on. The history up to $t$ alone does not suffice to determine the chances at $t$. (Just consider the following two worlds where nothing happens other than a series of coin flips. In one world the flips are independent and have chance .6 of coming up tails, while in the other they are independent and have chance .5 of coming up tails. It is possible for the first six flips to come up the same way in the two worlds while still maintaining different chances for the seventh flip. This can happen on any view on which chances are determined either by the Humean pattern including the future, or by non-Humean laws.)

On another picture of chance, the chances are determined not by the laws and the world, but by an experimental setup. The chance of a coin coming up heads may be 0.5 when the setup of the coin flipping situation is properly specified. But without a specification that the coin is flipped, that the flip is fair, that the coin is balanced, etc., it just may not be the case that it makes sense to say what the chance is that the coin will come up heads. On some ways of taking this, experimental *outcomes* are the result of chance processes, but experimental *setups* are the result of free choice of the experimenter. Conditional probability is a relationship between two events that are both in the domain of the probability function, while the experimental setup is a *precondition* for the existence of these probabilities at all. As Humphreys points out (Humphreys, 1985, 2004), Bayes' Theorem and other mathematical results allow us to invert conditional probabilities by means of some mathematical calculations. If there were such a thing as $p(\mathit{outcome} \mid \mathit{setup})$, then there would have to be something that is $p(\mathit{setup} \mid \mathit{outcome})$. But the setup is not the sort of thing that has a chance, as it is the result of a free choice, and the outcome is not the sort of thing that characterizes a chance process, so this conditional probability is either senseless or irrelevant. If we want to notate the role of the setup in determining the chance of the outcome, we should write it as $p_{\mathit{setup}}(\mathit{outcome})$, not $p(\mathit{outcome} \mid \mathit{setup})$.

This viewpoint on chance is similar to the one that frequentist statisticians have of probability. The only probabilities that make sense on this

view are the results of repeatable experiments. Scientific hypotheses help specify these probabilities, but do not themselves have probabilities, since they are not the results of repeatable experiments. This sort of thing is often notated by philosophers as $p_{setup}(E \mid H)$, where $E$ is some evidence consisting of experimental outcomes, and $H$ is a scientific hypothesis. The function represents something like the fraction of times that this outcome would occur if one were, hypothetically, to repeat this experimental setup many times, assuming the hypothesis is true. If this is the right way to represent the situation, then every statement of probability must have some scientific hypothesis or other that determines it, so every probability must be conditional.

However, I claim that on the frequentist interpretation, $H$ should not be thought of as being conditioned on, but must instead be part of the background, just like a world, confirmational commitment, or experimental setup. The clearest reason for this is that on the frequentist account, $H$ is from an importantly different ontological category than $E$, while conditional probability involves pairs of entities of the same ontological category. $H$ is either true or false, and not the outcome of a repeatable experiment. A hypothesis, for the frequentist, is not the sort of thing that has a probability, so it is not the sort of thing that can be conditioned on. In statistical practice, the difference is often indicated by using a semicolon to set off the hypothesis that is the precondition for the probabilities, rather than the vertical line, which is used for conditional probabilities. Thus, we should write "$P(E; H)$" rather than "$P(E \mid H)$".

Furthermore, there *is* a notion of conditional probability that the frequentist *can* talk about, that is quite different. On the hypothesis that an urn has 3 white and 7 black balls, the conditional probability of the second draw (without replacement) being black given that the first is white is $7/9$, while the unconditional probability of the second draw being black is $7/10$. In this case we can calculate the conditional probability as the unconditional probability of a white draw followed by a black one, divided by the unconditional probability of the first draw being white, all given the background of the urn hypothesis, which has no probability of its own for the frequentist. The Bayesian can say that all of these probabilities are conditional on the hypothesis, because the Bayesian thinks that the hypothesis is the sort of thing that has a probability. But the frequentist shouldn't say this. So the frequentist has no special need for primitive conditional probabilities.

### 1.2.3 *Actual Frequencies*

Some have argued that on the actual frequency interpretation of probability, all probabilities are fundamentally conditional. For this interpretation, I

agree. When probability is interpreted as frequency of some property within an actual reference class, every probability really is conditional.

The interpretation of probability as actual finite frequency says that $p(B \mid A)$ is the fraction of entities with property $A$ that also have property $B$. There is a particular number that is the frequency of heart attacks among 40-to-50-year-old American males in a given year, which we can calculate by counting how many 40-to-50-year-old American males there were that year, and counting how many of them had heart attacks that year. There is another frequency of heart attacks among all Americans, and another among all humans, calculated similarly. But if there is such a thing as the frequency of heart attacks independent of *any* reference class (even the entire universe), it is just a number, not a probability.

In this case, it looks like the reference class is the same sort of entity as the event whose probability is being measured. We can talk about the frequency of 40-to-50-year-old males among American heart attack victims, by counting how many heart attack victims there were that year, and finding what fraction of them were 40-to-50-year-old American males. Furthermore, if we ask for the *conditional* frequency of heart attacks among 40-to-50-year-old American males *given that* they smoke, this appears to be the same as the "unconditional" frequency of heart attacks among 40-to-50-year-old American males who are smokers. Conditionalizing really just is conjunction with the reference class. Thus, the reference class really is the same sort of thing as a conditioning event. Thus, on the actual finite frequency interpretation, we really do have a good case for every probability being conditional.

### 1.2.4 *Logical and Evidential Probabilities*

For logical and evidential probabilities (as well as perhaps some objective versions of the degree of belief interpretation of probability), some have argued that all probabilities are fundamentally conditional. For these interpretations, I don't specifically reject this argument. However, there is a special case of "empty background" that might be considered to be an unconditional probability that is equally fundamental to the conditional probabilities, so the upshot of the argument here is more equivocal.

Logical probability is often said to be a relation of partial entailment between two propositions. That is, "$p(B \mid A) = 1$" is said to mean the same thing (or something very similar to) "$A \vdash B$." Saying that $p(B \mid A) = 2/3$ is saying that $A$ "2/3 entails" $B$. Since entailment is a binary relation, this logical probability is said to be an essentially conditional relation. This is the point of view described, for instance, by Keynes (1921). (A similar viewpoint, though not identical, is expressed with regards to the "evidential probabilities" of Williamson, 2002.)

Both roles here are played by arbitrary propositions, so there are no ontological distinctions between the two sides of the conditional probability. There is no category mistake in reversing a logical entailment (though of course the degree of entailment can differ). Furthermore, just like with actual finite frequencies, there doesn't appear to be any other notion of conditional probability that is interestingly distinct from this one. The probability of $A$ given $B$, with $C$ as background, doesn't obviously have any interpretation that would be clearly different from the probability of $A$ with $B \wedge C$ as background. Thus, just as with actual frequencies, one might be able to argue on conceptual grounds that all logical probabilities are inherently conditional.

However, unlike with frequencies, the opponent of this view has a response. Deductive logic can be expressed as the study of logical entailment relations, but it can also be expressed as the study of theorems. One can think of theorems either as sentences entailed by a tautology, or as sentences entailed by no premises whatsoever. Similarly, it may be possible to consider the set of logical probabilities conditional on a tautology either as the degree of partial entailment the tautology gives to each sentence, or as the degree of partial theoremhood each sentence has.

If we can interpret $p(B \mid A)$ as the degree to which $A$ partially entails $B$, we may also be able to interpret $p(A)$ as the degree of partial theoremhood of $A$. On this account, it may be further possible to recover all the partial entailments from these facts about partial theoremhood through techniques of calculating conditional probabilities, just as it is possible to recover all the deductive entailments from the facts about theoremhood through the deduction theorem. Thus, the opponent of conditional probability as the fundamental notion may have a response to this argument, though it will depend on the extent to which conditional probabilities really can be recovered from the unconditional ones, just as in the case of Objective Bayesianism, or Levi's confirmational commitments.

### 1.2.5 *Summary*

In summary, degree of belief, physical chance, experimental chance, and hypothetical frequency all have some fundamental ontological distinction between the bearers of probability and the backgrounds that are required for probabilities to even exist. Thus, the necessity of these backgrounds does not motivate the claim that conditional probability is primitive or fundamental. For actual frequencies, logical probability, and evidential probability, the backgrounds are of the same type as the bearers of probability, so this argument does seem to motivate the claim that conditional probability is fundamental. But for logical and evidential probability, there is a possibility of empty background, which can be re-interpreted as a

fundamental notion of unconditional probability. Further mathematical investigation is needed to see whether these unconditional probabilities suffice to determine the conditional probabilities. Only for actual frequencies is it clear that all probabilities really are conditional, because of the necessity of a background for probability.

- All probabilities are non-trivially conditional:
  - Actual frequency
- All are conditional, some conditions are empty:
  - Logical
  - Evidential
  - Unique Degree of Belief
- Background relevant, not all are conditional:
  - Chance
  - Hypothetical Frequency
  - Permissive Degree of Belief

### 1.3 *Problems for the Ratio*

The previous section considers conceptual arguments that all probabilities are fundamentally conditional. I have argued that this argument works for the interpretation of probability as actual frequency, and is equivocal for logical and evidential probability and related objective epistemic interpretations, but that it does not work for the other interpretations of probability. In this section, I consider arguments for the claim that all probability is fundamentally conditional based on the mathematical features of conditional probability. This set of arguments is the center of Alan Hájek's (2003). Although this argument is perhaps easier to feel the grip of, and is largely independent of the particular interpretation of probability, I put it second, because consideration of it leads naturally to the technical issues considered in the later sections of this article.

The immediate target of Hájek's argument is the common claim that conditional probability is just *defined* as $p(A \mid B) = p(A \wedge B)/p(B)$. As Hájek points out, it appears to be a consequence of this definition that there is no such thing as the conditional probability $p(A \mid B)$ unless $p(B)$ has a precise non-zero numerical value. He then gives a litany of cases in which it seems clear that $p(A \mid B)$ exists, even though $p(B)$ is either zero, imprecise, vague, or non-existent. Thus, we must reject the ratio analysis as a definition of conditional probability. Whether this requires conditional probability to be a (or the) fundamental concept of probability theory is a deep and

difficult question that depends on what alternatives to the ratio analysis exist. The rest of the article after this section is a long consideration of these alternatives. Section 1.4 defines the particular mathematical features of probability and conditional probability that come up in addressing this problem. Section 2 and Section 3 consider the two advanced mathematical characterizations of conditional probability that avoid the problems of the ratio definition, one of which makes conditional probability primary and the other of which allows it to (almost) be calculated from unconditional probability. Evaluation of the merits of these two mathematical accounts is thus essential for deciding whether or not to accept Hájek's argument that conditional probability is prior to unconditional probability.

I will give examples of Hájek's cases shortly. I think that most are not decisive, but there is one family of them that is quite convincing for every interpretation of probability mentioned above, apart from actual frequencies. Thus it is interesting that the two primary arguments for conditional probability being fundamental have this complementary distribution—the one interpretation for which Hájek's argument against the ratio analysis clearly fails is the one interpretation for which all probabilities clearly require a background of the same type as the bearers of probability, so that it can clearly be understood as conditional probability.

### 1.3.1 *Impossible or Ruled Out Conditions*

I will begin by considering a type of case Hájek considers that is easy to reject. I think it is important to consider how this type of case differs from the others, which are more plausibly relevant. Let $H$ be the proposition that a particular coin flip comes up heads, and $T$ be the proposition that this same flip comes up tails. Hájek claims that $p(T \mid T) = 1$ under any circumstance. In particular, he claims (p. 287) that this should be true even if $p$ is the function encoding physical chances at a time when the flip has already happened and the coin already came up heads, so that $p(T) = 0$. He also suggests that it should be true if $p$ is the function encoding degrees of belief of a rational agent who has already learned that the coin came up heads, so that $p(T) = 0$.

These cases can be rejected because there doesn't appear to be a clear meaning for these conditional probabilities. Although I don't think that conditional probabilities are the probabilities of conditionals, there is a useful analogy to be drawn with conditionals. Conditional probability is intended to capture something more like an indicative conditional, rather than a subjunctive conditional or a material conditional, and indicative conditionals generally aren't considered in cases where the antecedent has already been fully ruled out. It seems correct to say, "if Oswald didn't kill Kennedy then someone else did," but this is because we allow that

our knowledge of the circumstances of the assassination is fallible. If we imagine fully ruling out any possibility that Oswald didn't commit the assassination, then the conditional becomes harder to interpret. We can apply subjunctive or material conditionals even to cases of necessary falsehoods, but it's hard to interpret them as indicative conditionals. Maybe we can make sense of a sentence like, "if 7 hadn't been a prime number, then 8 would have been," but a sentence like "if 7 isn't a prime number, then 8 is" seems only interpretable as a material conditional. Just as indicative conditionals seem not to be acceptable when the antecedent has been fully ruled out, none of the purposes for which conditional probabilities have been proposed makes any use of probabilities conditional on antecedents that have already been ruled out. There is no question of updating on or confirming a hypothesis that has been completely eliminated.

There are processes of belief *revision*, on which one removes a belief that one already has before updating on new information, but this is a different process that uses conditional probability from the revised state rather than the current state.[1] Similarly, the probability of outcomes conditional on acts that weren't done is irrelevant to decision theory.[2] Similarly, there is no question of how the chances of events will evolve when something that didn't occur does occur (though there may be a question of how chances will evolve when something of similar type to that event does occur), and there is no question of the degree of causal relevance of something that didn't occur (though there may be a question of the degree of causal relevance of its *non*-occurrence, which of course is something that *did* occur).

### 1.3.2 *Vague, Imprecise, or Gappy Conditions*

A second class of cases that Hájek considers involve vague or imprecise probabilities (pp. 293–5). It is controversial whether imprecise probabilities even exist (see Titelbaum, this volume, and Mahtani, this volume, for further discussion). But if they do, then it's clear that they cause problems. Perhaps one is uncertain about the outcome of the next United States presidential election in such a way that one has imprecise credences about it. Or perhaps it depends on non-deterministic events in a way that leaves it with an imprecise chance. Nevertheless, if $D$ is the proposition

1 Levi's notion of confirmational commitments allows for probability conditional on propositions that are currently known to be false. But in this case, the probability function is not the current degree of belief function, but rather the confirmational commitment—the current degree of belief function is itself conditional on current knowledge. Thus, the probability conditional on something currently known to be false is a prior commitment of an indicative sort—not Hájek's probability conditional on a certain falsehood.

2 Brandenburger (2007) has argued that game theory sometimes needs to consider probabilities conditional on actions that are ruled out by rationality considerations, but these are not ruled out with certainty, the way that tails was in Hájek's examples.

that a Democrat will win the next US presidential election, and $H$ is the proposition that a completely unrelated coin flip will come up heads, it seems clear that $p(H \mid D) = 1/2$.

However, this challenge may not be a fatal objection to the ratio analysis either. One proposal about imprecise probabilities is that, rather than $p(D)$ being an imprecise value (or set or whatever), there are instead multiple precise probability functions $p_i$ that are all part of the representation of degree of belief, or chance, or whichever interpretation of probability we are considering. On each such function, $p_i(H \mid D)$ can be well-defined by the ratio formula, and if they all happen to take value $1/2$, then the conditional probability can be precise even though the unconditional probability is not. (This response is described in slightly greater detail on page 295 of Hájek's paper.)

Hájek puts the most weight on cases where there is *no* unconditional probability, but conditional probabilities are well-defined. He gives a long series of such cases on pp. 295–312. These include cases of free actions (which may be such that they *can't* have credences or chances), mere gaps in the credences or chances, and cases of non-measurable sets.

I think that mere gaps are either best thought of as maximally imprecise probabilities and addressed supervaluationally as above, or as events that are outside of the scope of the relevant probability function. An agent who fails to have a degree of belief in some proposition is an agent who hasn't considered or grasped it, and thus fails to have any degree of belief conditional on it as well (even though there are some facts about what degree of belief she *should* have were she to have them—like $p(A \mid A) = 1$). Similarly with non-measurable sets—if they are outside the bounds of chance or credence, then there are no meaningful conditional probabilities on them either.

There may be some class of events (perhaps the actions of a free agent who is in the process of deliberation) that *can't* have probabilities, but which themselves serve as the conditions for probabilities of other events. However, some of these may in fact be better thought of as the "backgrounds" for probabilities that I considered in Section 1.2. This may be the right way to think of the "action probabilities" of decision theory, for instance, where every probability must depend on a specification of the action of the agent. However, if there were a class of events that can't have probabilities, but which also aren't essential to the specification of other probabilities, even though they can affect them, then this would be a better case.

### 1.3.3 *Probability 0 Conditions*

At any rate, I think the strongest case is one that Hájek puts less weight on (pp. 289–290). These are cases arising from consideration of infinite probability spaces, where some events have probability 0 *without* being ruled out. Consider a point on the surface of a sphere. Label the sphere with lines of latitude and longitude like those of the Earth. Let $N$ be the proposition that the point is in the northern hemisphere. Let $L_\theta$ be the proposition that the point is on the line of longitude at angle $\theta$ from the boundary between the eastern and western hemispheres. If the initial probability distribution is uniform, then it is quite plausible that $P(N \mid L_0) = 1/2$, even though $P(L_0) = 0$, so that $P(N \wedge L_0)/P(L_0)$ is undefined. Furthermore, even if the initial probability distribution isn't uniform, it seems that $P(N \mid L_\theta)$ should be defined whenever there is some possibility of $L_\theta$ being true. However, there are uncountably many distinct values of $\theta$, and at most countably many of them can have positive probability (because at most $n$ of them can have probability greater than $1/n$, for each of the countably many integers $n$, and any positive number is greater than $1/n$ for some integer $n$). Thus, there must be some way to make sense of these conditional probabilities, despite the use of probability 0. This example can be generated for probability interpreted as chances or as degrees of belief or as evidential probability, or any interpretation, as long as there are uncountably many distinct possibilities that aren't ruled out.

There are two methods that have been proposed to block this set of cases. One is to introduce additional non-zero values for the probability function to take that are nevertheless lower than $1/n$ for any positive integer $n$. I have argued elsewhere that this method is unlikely to be correct for chances or degrees of belief (Easwaran, 2014). (This proposal is discussed in more detail by Wenmackers, this volume.) Furthermore, this option bears some relationship to one of the proposals described later, in Section 3.1, so I suggest that this is in some sense not really an alternative to the methods considered here—it is effectively equivalent to letting the probability take the value 0.

The other method for blocking this sort of case is to argue that the relevant notion of probability *can't* have uncountably many disjoint possible events. In the case of Bayesian probabilities, this is motivated by some consideration of the finitude of the human mind, while in the case of chances it is motivated by some understanding of quantum mechanics as requiring the universe to be discrete in time, space, and every other meaningful parameter.

However, this sort of interpretation of quantum mechanics is implausible. Although certain parameters like charge and spin are quantized, time and space just enter into "uncertainty" relations. This means that they are

bound to other parameters in a way that interactions depending very precisely on one parameter must allow for exceedingly large variation on the other. However, this does not put any specific lower bound on the precision of any interaction, and doesn't directly motivate the idea that space and time are discrete.

Furthermore, although any particular human mind is finite, there is reason to allow consideration of every hypothesis of the form $V > p/q$, where $V$ is some physical parameter, and $p$ and $q$ are integers. Certainly, science seems to proceed as if each of these hypotheses is meaningful, even if we can never be absolutely sure which are true or false. But these countably many hypotheses together generate a family of uncountably many hypotheses of the form $x = r$ where $r$ is a real number. (The claim that all of the relevant algebras are countably generated, or generated by random variables in this way will be important in Section 2.3.2.) The example with points on a sphere is exactly like this, but so are many others that are more directly relevant in science. To reject these cases is to say that every probability function has some finite limit on the size of examples that are relevant.

This response in terms of finitism is quite effective in the interpretation of probability as actual frequency, if the classes of events one is discussing are always finite. (When the classes may be infinite, it's hard to say how to even *define* the notion of frequency involved.) But this response is no help to the statistical frequentist, who may be interested in scientific hypotheses of the relevant sort. Philosophers often make reference to examples involving a dart thrown at a board, with infinitely many points that its center might hit, or a fair coin being flipped infinitely many times, for which each sequence of heads and tails is a possible outcome. But examples involving infinity are central to much scientific practice as well.

For instance, a statistical frequentist may be interested in some hypothesis about how energetic particles are ejected from an atomic nucleus under a particular sort of process. She may further be interested in the question of how the energy distribution of these particles is correlated to the direction in which they are ejected. If we let $E_x$ be the statement that the energy of the particle is $x$, and $D_\theta$ be the statement that the particle is ejected in a direction at angle $\theta$ to the motion of the atomic nucleus, then she could be interested in all probabilities of the form $p(E_x \mid D_\theta)$. But if she hypothetically imagines the process being repeated infinitely many times, the probability of many of the $D_\theta$ is likely to be zero, given that there are uncountably many directions in which the particle could be ejected. If we limit consideration to some *actual* set of experiments, then there are likely to be only finitely many such ejections, and so the non-realized $D_\theta$ can be ignored. But the statistical frequentist is interested in *hypothetically* repeated experiments, so all of these possibilities must be considered.

To summarize, there may be a way to resist all of these cases. But it would involve some extensive use of special backgrounds for certain types of probability, a particular way of dealing with any kind of imprecision in probability functions, and a rejection of infinity. Most of the mathematical work on alternatives to the ratio analysis only address the issue of infinite probability spaces and probability 0. I think that the other problems can be avoided as in ways that I have suggested along the way. But there is certainly room for further philosophical and mathematical analysis of those suggestions, and perhaps for new alternatives, which may or may not prioritize conditional probability over unconditional probability. But the rest of this article will examine the mathematical theories that have been developed for dealing with the problems that arise around infinite probability spaces and the resulting events of probability 0.

### 1.4 *Additivity, Disintegrability, and Conglomerability*

Once we consider these infinite families of hypotheses, it seems that we must have some way of making sense of $p(A \mid B)$ even when $p(B) = 0$. There are many different mathematical theories that allow this to work out, and these will be the subject of the further sections of this article. The reason there are so many different theories is due to a fundamental dilemma around infinity, which will take some time to explain.

Every such theory begins with the idea that the "definition" $p(A \mid B) = p(A \wedge B)/p(B)$ should be replaced with an axiom $p(A \mid B)p(B) = p(A \wedge B)$. We can then consider whether further information allows us to define $p(A \mid B)$ from the unconditional values, or at least in some sense ground it in them, or whether we must take $p(A \mid B)$ as a fundamental function separate from the unconditional probability function $p(A)$. However, even allowing for this function, there are difficulties when the set of possibilities is infinite.

In this section I will discuss some of the mathematical properties involved, and show that the idea that conditional probability can be understood as a function $p(A \mid B)$ conflicts with the natural generalization of Additivity in cases of infinity. We must either give up on Additivity (and related principles generalizing the Law of Total Probability), or else accept that conditional probability is given by a function $p(A \mid B, \mathcal{E})$ for a further parameter $\mathcal{E}$. The mathematical theory of conditional probabilities for infinite sets is an interplay between the two horns of this dilemma.

In this section I will formally treat the bearers of probability as sets of possibilities, and will largely bracket concerns about the interpretation of probability until the end.

### 1.4.1 *Additivity*

When dealing with infinity, a fundamental question for probability theory is whether and how to generalize the notion of Additivity. One of the standard axioms of probability is that if $A_1$ and $A_2$ are disjoint events (that is, there is no possibility on which they both occur) then $p(A_1 \cup A_2) = p(A_1) + P(A_2)$. Kolmogorov and others have considered a generalization of this axiom to countable cases.

**Definition 1** *The $A_i$ for $i \in I$ form a* partition *of $A$ iff each $A_i$ entails $A$, and whenever $A$ is true, exactly one of the $A_i$ is true.*

(If no particular $A$ is mentioned, then I am considering a partition of the set of all possibilities.) Thinking of the $A_i$ as sets, that means that they are disjoint, and their union is $A$. I will refer to this partition with boldface $\mathbf{A_I}$, and with the index set $I$ as subscript, while italic $A_i$, with a member $i$ of $I$ as subscript, will refer to the member of $\mathbf{A_I}$ that is indexed by element $i$.

One way to state Countable Additivity is as the requirement that for any countable partition $\mathbf{A_I}$ of $A$, we have $p(A) = \sum_{i \in I} p(A_i)$. Kolmogorov actually framed his axiom in a slightly different form as a sort of continuity—whenever the $B_i$ for $i \in \mathbb{N}$ are a family of sets whose intersection is empty, we have $\lim_{n \to \infty} p(\bigcap_{i=0}^{n} B_i) = 0$.

However, I think that it is more perspicuous to phrase this generalization in a third way, in order to more clearly demonstrate the further generalizations to uncountable sets. The following is a theorem of standard finitely additive probability, whenever $\mathbf{A_I}$ is a partition of $A$.

**Theorem 1** *If $x \geq p(A)$, then for any finite $I_0 \subseteq I$, $x \geq \sum_{i \in I_0} p(A_i)$.*

We can then define additivity as the converse.

**Definition 2 ($\mathbf{A_I}$-Additivity)** *If for every finite $I_0 \subseteq I$, $x \geq \sum_{i \in I_0} p(A_i)$, then $x \geq p(A)$.*

The following definition is equivalent.

**Definition 3 ($\mathbf{A_I}$-Additivity)** *If $x < p(A)$ then there is some finite $I_0 \subseteq I$ such that $x < \sum_{i \in I_0} p(A_i)$.*

Countable Additivity is equivalent to $\mathbf{A_I}$-Additivity for all countable sets of indices $I$.[3] This is because, for a set of non-negative real numbers,

3 We can also naturally talk about $\kappa$-Additivity as $\mathbf{A_I}$-Additivity for all $I$ with cardinality *less* than $\kappa$. This is standard notation though it is slightly confusing that Countable Additivity, also known as "$\sigma$-Additivity," is $\aleph_1$-Additivity, while $\aleph_0$-Additivity is actually *Finite* Additivity. But this notation is relevant to distinguish between Additivity for all cardinals strictly below $\aleph_\omega$, and Additivity for all cardinals up to and including $\aleph_\omega$, which is called $\aleph_{\omega+1}$-Additivity.

the sum of that set is the smallest real number that is at least as great as every finite sum of those numbers.[4]

Countable Additivity is not entailed by the standard probability axioms, and in fact rules out certain intuitively appealing probability distributions. The classic proposed counterexample to Countable Additivity is often known as the "de Finetti lottery" (de Finetti, 1974; for more detailed discussion see Bartha, 2004, and Howson, 2008). Imagine that some natural number is chosen in such a way that no number is more likely than any other. This intuitively seems possible, and yet it is ruled out by Countable Additivity. Since every number is equally likely to be chosen, each number must have probability less than $1/n$, because otherwise some $n$ of them would exhaust all the probability. The only way for this to be the case is for each number to have probability 0. But this is a violation of Countable Additivity, because the sum of these 0s is strictly less than 1, which is the probability of the countable disjunction of these possibilities.

Considering Definition 3, we can derive a more general set of apparent problems. Let each $A_i$ stand for the event of the number $i$ being picked, and let $I$ be the set $\mathbb{N}$ of all natural numbers, so that $\mathbf{A_I}$ is a partition of the necessary claim that some number or other is picked. In this case, Definition 3 of $\mathbf{A_I}$-Additivity states that for every $x < 1$, there must be some finite $I_0$ such that $x < \sum_{i \in I_0} p(A_i)$. That is, for every $x < 1$, there is some finite set such that the probability that the number chosen is from that set is at least $x$. $\mathbf{A_I}$-Additivity doesn't just rule out uniform distributions on the natural numbers—it requires that *every* distribution concentrate most of the probability on some finite set or other.

If $\mathbf{A_I}$-Additivity holds for *all* partitions $\mathbf{A_I}$, then the probability function is said to be Fully Additive. In this case, for any partition $\mathbf{A_I}$ of a set

4 Readers may be familiar with the definition of the sum of a sequence of (non-negative or negative) numbers $a_i$ for $i \in \mathbb{N}$ as

$$\sum_{i \in \mathbb{N}} a_i = \lim_{n \to \infty} \sum_{i=1}^{n} a_i.$$

This definition doesn't work for index sets other than $\mathbb{N}$, and makes essential use of the order of the indices. When some terms are negative, this order can be important—the same set of numbers can have a different sum when added in a different order, if both the negative and positive terms separately sum to infinite values. But when all terms are non-negative, the least upper bound of the sums of finite subsets is the same as the sum of the terms in any order (because every finite initial sequence is a finite subset, and every finite subset is contained within some finite initial sequence, and since there are no negative terms, the sum of any larger subset is at least as great as the sum of any subset contained within it).

For *uncountable* infinite sets of non-negative numbers, it is hard to extend the sequential definition, because we don't have good methods for dealing with uncountably long sequences. However, the least upper bound of the set of all sums of finite subsets is still well-defined.

$A$, Definition 3 entails that for every $n$, there is a finite set of $A_i$ whose probability adds up to more than $p(A) - 1/n$. Let $I' \subset I$ be the union of the countably many finite sets of indices of these sets, which is thus countable. By Theorem 1, if we let $A' = \bigcup_{i \in I'} A_i$, then $p(A') \geq p(A) - 1/n$ for each $n$ (since it contains a finite subset adding to this probability). Since $A' \subset A$, we have $p(A') = p(A)$. Thus, the remainder of $A$ that is not in $A'$, $A \setminus A'$, must have probability 0. If $A$ was the set of all possibilities, and each $A_i$ is a singleton set containing a single possibility, then $A'$ is countable. Not only does each element outside of this countable set *individually* contribute probability 0, but even *collectively* they all contribute 0.[5] Thus, if Full Additivity holds, there is a sense in which we can ignore all but countably many possible outcomes, and these countably many outcomes have individual probabilities that add up to 1. A probability function in which the set of all possibilities is countable is said to be *discrete*. While there are many interesting applications of discrete probability, there are also plenty of applications for which no countable set of possibilities should account for all the probability, such as any scientific question for which every real number within some interval is a possible answer. Thus, most probability theorists do not accept Full Additivity.

We can think of different views of probability as along a sort of scale (Figure 1). At the most restrictive end there is the strongly finitistic view that there are only finitely many possibilities that probability is distributed over. Next we get the discrete view, that there are only countably many possibilities that probability is distributed over—this is classical probability theory with Full Additivity for all cardinalities. Next we get the traditional mathematical view on which the set of possibilities can be uncountable, but the probability function is required to satisfy Countable Additivity. Finally, at the most liberal end of the scale, we have the minority view in mathematics but a popular view in philosophy, where the probability space can be uncountable and the probability function is only required to satisfy Finite Additivity. (Some of the popularity of this view among philosophers may stem from confusion with probability over finite spaces, at the opposite end of the scale.) Finite and discrete probability have no problem with Additivity, and in fact allow conditional probability to be uniformly defined by the ratio. However, the consideration of scientific examples where we want to measure the unknown value of some parameter push us towards uncountable spaces. So it is useful to investigate the

5 Another way to see this is to consider the probabilities of each individual possibility. For each $n$, at most $n$ of the individual possibilities can have probability greater than $1/n$. Thus, at most countably many have non-zero probability. But if Full Additivity holds, then the sum of all the probabilities of the individual possibilities must be 1. So these countably many non-zero probabilities must add up to 1. Thus, the set of all possibilities other than the countably many with non-zero probability must be a set with probability 0.

ways in which probability functions with failures of Additivity can still be well-behaved. I believe that Countable Additivity is the most useful point on this scale, but it is worth considering the mathematical features of all four points.

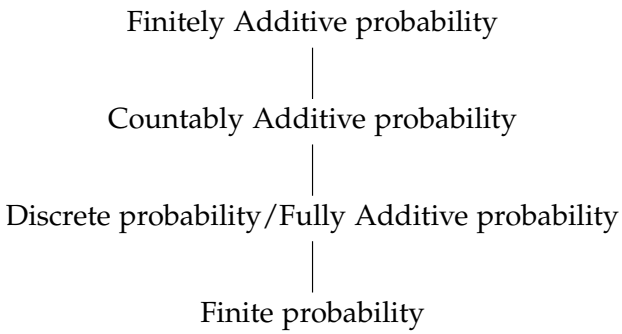


Figure 1: A scale of views

### 1.4.2 *Disintegrability and Conglomerability*

Although generalizations of Additivity are quite controversial, there are related principles that have been argued to generalize to infinite cases. These principles are defined by using integration in place of addition when infinity arises, to avoid some of the difficulties of adding up zeros. By the end of this section, I will mention some results that show that instances of these principles must fail when instances of Additivity fail. However, in Section 1.4.3, I will show that we can avoid these failures by defining conditional probability relative to a partition.

The starting point for discussion of these principles is the Law of Total Probability.

**Theorem 2 (Finite Law of Total Probability)** *If $A_1$ and $A_2$ are incompatible, and $A$ is the disjunction $A_1 \cup A_2$, then*

$$p(B \cap A) = p(B \mid A_1)p(A_1) + p(B \mid A_2)p(A_2).$$

Given two instances of the conjunction law, $p(B \cap A_i) = p(B \mid A_i)p(A_i)$, this is equivalent to an instance of Additivity: $p(B \cap A) = p(B \cap A_1) + p(B \cap A_2)$. We can state a generalization of this, where $\mathbf{A_I}$ is a partition of some set $A$.

**Definition 4** *The* $B \cap \mathbf{A_I}$-Law of Total Probability *states that*

$$p(B \cap A) = \sum_{i \in I} p(B \mid A_i)p(A_i).$$

Given that $p(B \cap A_i) = p(B \mid A_i)p(A_i)$, it is straightforward to see that the $B \cap \mathbf{A_I}$ Law of Total Probability is equivalent to $B \cap \mathbf{A_I}$-Additivity. Giving up Full Additivity means giving up certain instances of the Law of Total Probability. But there are ways of modifying the Law of Total Probability that don't directly take this additive form.

The Law of Total Probability can be related to considerations of expected value for random variables. Informally, a random variable is some quantity with a potentially unknown real number value, where for each real number $x$, there are well-defined probabilities $p(V > x)$ and $p(V = x)$. Notably, the set of events $V = x$ form a partition.

**Definition 5** *When there are only finitely many possible values for $V$, the* expected value *of $V$ is given by*

$$exp(V) = \sum_x x \cdot p(V = x),$$

*where the sum ranges over all finitely many possible values for $V$.*

This definition would yield strange results if it were applied to a variable $V$ for which Additivity fails on the partition into $V = x$.

Any violation of Additivity must involve some partition $\mathbf{A_I}$ such that $\sum_{i \in I} p(A_i) = 1 - \epsilon$. If $I$ has cardinality at most that of the set of real numbers, then we can generate a random variable whose expected value under an extension of the above definition would be paradoxical. For each $i \in I$, let $\epsilon_i$ be a distinct positive value less than $\epsilon/(1-\epsilon)$. Let $V$ be a random variable that takes on the value $1 + \epsilon_i$ iff $A_i$ is true. Then a naive extension of Definition 5 would tell us that $exp(V) = \sum_{i \in I}(1 + \epsilon_i)p(A_i)$. But by choice of $\epsilon_i$, we see that $(1 + \epsilon_i) < (1 + \epsilon/(1-\epsilon)) = 1/(1-\epsilon)$. Thus, $exp(V) < \sum_{i \in I}(1/(1-\epsilon))p(A_i) = (1/(1-\epsilon))(1-\epsilon) = 1$. That is, even though $V$ is a random variable whose value is always strictly greater than 1, this definition of expectation would yield an expected value that is strictly less than 1.

To avoid this problem, it has been standard to define expected value slightly differently in infinite cases. Instead of directly considering the probability of $V = x$ for each possible value that $V$ can take on, mathematicians just directly rule out discontinuities like the one mentioned above. If $V$ is a random variable that only has finitely many possible values, then we follow the old definition and let $exp(V) = \sum_x x \cdot p(V = x)$. If $V$ has infinitely many possible values, but has a lower bound (that is, there is some $l$ such that it is certain that $V > l$), then we can avoid this problem. If $V'$ is a random variable that always takes a value strictly less than $V$, we will say $V' < V$. We will just directly stipulate that if $V > V'$ then $exp(V) > exp(V')$. This will rule out the problem of the previous paragraph, because we could let $V'$ be the random variable that always takes

the value 1, and see that $exp(V) > exp(V') = 1$. By considering variables $V'$ that only take on finitely many distinct values, we get a set of lower bounds for what $E(V)$ could be. We say that the expectation of $V$ is the least number above all these lower bounds (the "supremum" of this set of lower bounds).

**Definition 6** *Let $V$ be a random variable with a lower bound. Then*

$$exp(V) = \sup_{V' < V} exp(V'),$$

*where $V'$ ranges over variables that only take on finitely many distinct values.*

Similarly, for random variables that have an upper bound, we can define the expectation to be the greatest number below all the upper bounds (the "infimum" of this set). We then deal with unbounded random variables by breaking them into a component with a lower bound and an upper bound. Let $V^+$ be the random variable that agrees with $V$ when $V$ is positive and is 0 otherwise, and $V^-$ be the random variable that agrees with $V$ when $V$ is negative and is 0 otherwise. Then define $exp(V)$ as follows.

**Definition 7**

$$exp(V) = \int V \, \mathrm{d}p = \sup_{V' < V^+} \sum_x x \cdot p(V' = x) + \inf_{V' > V^-} \sum_x x \cdot p(V' = x),$$

*where $V'$ ranges over random variables that only take finitely many distinct values.*

This is the definition of the Lebesgue integral of $V$ with respect to probability function $p$, and is the final generalized definition of expected value. It agrees with Definition 5 and Definition 6 in the cases where they apply.

With this new definition, we can try to save the Law of Total Probability in a slightly different form. Let $\mathbf{A_I}$ be a partition. We can consider $p(B \mid \mathbf{A_I})$ as a random variable whose value is given by $p(B \mid A_i)$ for whichever proposition $A_i$ is the unique one from $\mathbf{A_I}$ that is true. If $\mathbf{A_I}$ is finite, then the Law of Total Probability takes the form $p(B) = exp(p(B \mid \mathbf{A_I}))$. This motivates the following definition.

**Definition 8** *$B$ is* Disintegrable *over the partition $\mathbf{A_I}$ iff*

$$p(B) = \int p(B \mid \mathbf{A_I}) \, \mathrm{d}p.$$

Disintegrability is thus another generalization of the Law of Total Probability, formulated with integrals rather than (potentially infinite) sums.

Let $\mathbf{A_I}$ be any partition, $I'$ be any subset of $I$ and $A' = \cup_{i \in I'} A_i$. Define *Conglomerability* as follows.

**Definition 9** $p(B \mid \mathbf{A_I})$ *is* Conglomerable *over* $A'$ *iff*

$$\inf_{i \in I'} p(B \mid A_i) \leq p(B \mid A') \leq \sup_{i \in I'} p(B \mid A_i).$$

It is useful to compare Conglomerability to van Fraassen's principle of "reflection" (van Fraassen, 1984; Briggs, 2009).

It is not hard to see that Disintegrability of $B$ over $\mathbf{A_I}$ entails Conglomerability over each $A'$ with positive probability (because constant functions taking on the infimum or supremum of $p(B \mid A_i)$ are among the set of random variables whose expectation is considered in calculating $exp(p(B \mid \mathbf{A_I}))$). Conversely, Conglomerability of $p(B \mid \mathbf{A_I})$ over all $A'$ with positive probability entails Disintegrability of $B$ over $\mathbf{A_I}$. (Since the integral is defined by comparison to finite sums, this only requires the Finite Law of Total Probability, rather than the generalizations that fail when Additivity fails over infinite partitions.)

We might hope that these new generalizations of the Law of Total Probability in terms of integration rather than summation don't require Countable Additivity. However, this hope turns out to be misplaced. A general theorem is proven by Hill and Lane (1985), verifying that for countable probability spaces, Conglomerability and Countable Additivity are equivalent. That is, any failure of Countable Additivity entails a failure of Conglomerability, and thus Disintegrability, which is the generalization of the Law of Total Probability. (Slightly more general versions of this result were proven earlier by Schervish, Seidenfeld, and Kadane, 1984.)

Instances of this result were noted by de Finetti (1974, pp. 177–8), who also conjectured the general result but hadn't proven it. To see the basic idea, consider something like the de Finetti lottery, where each natural number has equal probability of being chosen. Let $E$ be the event that an even number is chosen. Intuitively, $p(E) = 1/2$. However, if we consider the partition into the sets $A_i = \{2i + 1, 4i, 4i + 2\}$, then intuitively $p(E \mid A_i) = 2/3$, so that the unconditional probability of $E$, which is $1/2$, is strictly outside the range spanned by its probabilities conditional on each member of the partition, which are all $2/3$. The construction by Hill and Lane notes that even without the assumptions of uniformity underlying the specific probability judgments $1/2$ and $2/3$, if $E$ and its complement are both sets of positive probability, then we can often create each $A_i$ by taking enough elements of $E$ with one element of its complement to make $p(E \mid A_i) > p(A) + \epsilon$. If we can't do this for every element of the complement, we can usually do it by taking enough elements of the complement with one element of $E$ to make $p(E \mid A_i) < p(A) - \epsilon$. The tricky part of the Hill and Lane construction is showing how to create a special partition in the case where neither of these techniques works. These results have been generalized to show that there are failures of Conglomerability

for probability distributions that satisfy Countable Additivity but fail to satisfy Additivity at some cardinality beyond the countable (Seidenfeld, Schervish, & Kadane, 2013, 2014). Thus, Disintegrability and Conglomerability don't let us get quite as much distance from Additivity as we might hope.

### 1.4.3 *The Fundamental Dilemma*

However, there is a way to separate Disintegrability and Conglomerability from Additivity.

First, we should note that Additivity only makes reference to unconditional probabilities, while Disintegrability and Conglomerability make reference to conditional probabilities. Furthermore, Disintegrability and Conglomerability make reference to conditional probabilities $p(B \mid A_i)$ only in the context of a random variable $p(B \mid \mathbf{A_I})$. In generating a contradiction to Conglomerability from a failure of Additivity, Hill and Lane needed to construct a *new* partition by joining together elements of $\mathbf{A_I}$. (This is also the case for Seidenfeld et al.) Thus, if a given set $A$ is an element of two distinct partitions $\mathbf{A_I}$ and $\mathbf{A'_{I'}}$, we can avoid the problems if we *change* the value of $p(B \mid A)$ when we move from considering $\mathbf{A_I}$ to considering $\mathbf{A'_{I'}}$. That is, we should consider conditional probability as a three-place function, $p(B \mid A_i, \mathbf{A_I})$, so that changing just the partition can change the value of the conditional probability, even if we are considering the same events $B$ and $A_i$. Some theorists find this repugnant to their sense that conditional probability $p(B \mid A_i)$ must have a single value, but it enables us to avoid the paradoxes.

This move was in fact already made by Kolmogorov (1950). Although he hadn't noticed the connections between Additivity principles and Conglomerability, he had already noticed some problems that Conglomerability apparently led to, and avoided them by turning conditional probability into a three-place function of two events and a partition.[6] (In fact, this problem was already mentioned as early as Bertrand, 1889, though due to Borel's work on this problem, and the existence of another paradox known as "Bertrand's Paradox," this has come to be known as the "Borel Paradox.")

Imagine a point uniformly chosen from the surface of a sphere, labeled with latitude and longitude like the surface of the Earth. Consider the set $P$ of "polar" points—those with latitude greater than 60 degrees north or greater than 60 degrees south. Consider the set $E$ of "equatorial" points—those with latitude between 30 degrees south and 30 degrees north. Let $L_\theta$ be the great circle of longitude $\theta$. By symmetry, it seems that $p(P \mid L_\theta)$

6 Strictly speaking, Kolmogorov worked with a "sub-$\sigma$-algebra" rather than a partition, but we will discuss the relation of these concepts in Section 2.

should be independent of $\theta$, and so should $p(E \mid L_\theta)$. Conglomerability over the partition[7] $\mathbf{L}_\theta$ requires that $p(P) = p(P \mid L_\theta)$ and $p(E) = p(E \mid L_\theta)$. But $p(P) = \frac{2-\sqrt{3}}{2} \approx 1/8$ while $p(E) = 1/2$. Note that $P$ and $E$ each cover $1/3$ of the length of $L_\theta$. Thus, conditionalizing a uniform distribution over the sphere in a way that is Conglomerable over the longitudes gives a conditional distribution that is concentrated near the equator and away from the poles.[8]

To force a problem for the two-place conditional probability function, we can fix a given line of longitude and shift which partition it is considered as a member of. Re-describe the sphere so that the poles are still on this line, but where the old equator was. This switches which points on the line are polar and which are equatorial. Conglomerability requires the very same great circle to give rise to different conditional probabilities when considered as a line of longitude for one set of coordinates, rather than as a line of longitude for a different set of coordinates. If we let $C$ be this circle, and $\mathbf{L}_\theta$ be the partition into lines of longitude for the given poles, while $\mathbf{L}_\phi$ is the partition into lines of longitude for poles where $C$ intersects the equator of the original partition, then we get $p(P \mid C, \mathbf{L}_\theta) = \frac{2-\sqrt{3}}{2}$ while $p(P \mid C, \mathbf{L}_\phi) = 1/2$. Conditioning on the same event gives different results when that event is considered as drawn from one partition rather than another.

Thus, Conglomerability already motivates the idea that conditional probability depends not just on the conditioning event, but also on the partition from which that event is drawn. Since the arguments from Conglomerabil-

7 Strictly speaking, $\mathbf{L}_\theta$ do not form a partition, because every line of longitude includes the poles. However, the example can be slightly modified without making any significant changes to anything by just removing the poles from the sphere, or arbitrarily adding the poles to one particular line of longitude and not any of the others. A slightly cleaner version of the same sort of case exists if $X$ and $Y$ are two independent normally distributed variables with mean 0 and standard deviation of 1. Exercise 33.2 of Billingsley (1995) notes that conditioning on $X - Y = 0$ relative to the partition $\mathbf{X} - \mathbf{Y}$ gives different results from conditioning on $X/Y = 1$ relative to the partition $\mathbf{X}/\mathbf{Y}$. Example 6.1 on pp. 224–5 of Kadane, Schervish, and Seidenfeld (1986) considers the case where $Y = 0$ has been ruled out and notes that conditioning on $X = 0$ relative to the partition $\mathbf{X}$ gives different results from conditioning on $X/Y = 0$ relative to the partition $\mathbf{X}/\mathbf{Y}$.

8 Some have worried that the appeal to symmetry in the argument that $p(P \mid L_\theta)$ should be independent of $\theta$ is enough like the appeal to symmetry in the intuition that the conditional probability should be uniform that *both* are suspect. However, if we take the partition into account as part of the description of the problem, then there is a relevant difference. The unconditional probability is symmetric under *any* rotation of the sphere. However, the partition into lines of longitude is only symmetric under rotations of the sphere about the poles—rotating about any other point sends some lines of longitude to great circles that are not lines of longitude. In particular, rotation *along* any particular line of longitude changes the partition, so there is no need for probability conditional on this partition to preserve uniformity under this rotation. See p. 303 of Chang and Pollard (1997) for more discussion of this symmetry breaking.

ity to Additivity rely on generation of new partitions, we might hope that allowing conditional probability to vary as the partition changes can avoid the worst consequences. And in fact it often can. As shown by Billingsley (1995, Theorem 33.3), if $p$ is a probability function satisfying Countable Additivity over the events involving two random variables, then there is a way to specify the values for $p(B \mid A, \mathbf{A})$ while satisfying Conglomerability, where $\mathbf{A}$ is the partition of possible values of one variable, and $B$ ranges over any proposition involving the two variables. In particular, this means that it is possible to give up on all forms of Additivity beyond Countable Additivity while holding on to Conglomerability.[9]

Thus, we have a choice between allowing conditional probability to be a three-place function $p(B \mid A, \mathbf{A})$ depending on a partition as well as a pair of events, and having unrestricted Conglomerability while only keeping Countable Additivity; or requiring conditional probability to be a two-place function $p(B \mid A)$ just of two events and keeping only as much Conglomerability as we do Additivity. The former option is called *Regular Conditional Probability*, while the latter is called *Coherent Conditional Probability*. ('Coherent' in this sense just means that the same pair of events has the same conditional probability regardless of what algebra it was drawn from, and is not related to the use of the word 'coherent' to mean "satisfying the probability axioms." I don't know where the term 'regular' comes from here, but it is not related to the concept requiring non-zero probabilities.) Mathematical theories of these two types will be the subjects, respectively, of Section 2 and Section 3.

Fuller consideration of the costs and benefits of these proposals will come in Section 2 and Section 3. But I will first mention several arguments for Conglomerability, which defenders of Coherent Conditional Probability must reject.

Recall that Conglomerability (Definition 9) says that for any partition $\mathbf{A}$, $\inf_{A \in \mathbf{A}} p(B \mid A) \leq p(B) \leq \sup_{A \in \mathbf{A}} p(B \mid A)$. By considering either $B$ or its negation as needed, a violation means that there is some value $x$ such that $p(B) < x$, but for every $A \in \mathbf{A}$, $p(B \mid A) > x$. If we consider the role of conditional probability in updating degrees of belief or in measuring confirmation, then this means that if one is about to perform an experiment whose possible outcomes are $\mathbf{A}$, then one can know in advance that one will get evidence confirming proposition $B$. This possibility seems intuitively costly for statistical or scientific reasoning, though there have been some attempts to mitigate it (Kadane, Schervish, & Seidenfeld, 1996).

For update via Jeffrey Conditionalization, Conglomerability is even more natural. Recall that update via Jeffrey Conditionalization proceeds by tak-

9 There are some other challenges to Conglomerability raised by Arntzenius, Elga, and Hawthorne (2004), but these also depend on changing partitions while keeping conditional probability fixed.

ing some partition **E** of possible evidence and updating one's old degrees of belief $p(E)$ to new degrees of belief $p'(E)$ for all $E \in \mathbf{E}$. This then propagates through the rest of one's beliefs by means of "rigidity," the requirement that for any proposition $A$, we have $p'(A \mid E) = p(A \mid E)$. In the finite case, the Law of Total Probability tells us that $p'(A) = \sum_{E \in \mathbf{E}} p'(A \mid E)p'(E)$, and since these values are specified, so are the probabilities for all other propositions. In the infinite case, we need some version of the Law of Total Probability for this to generalize. The natural thought is that we should have $p'(A) = \int p'(A \mid \mathbf{E})\, \mathrm{d}p'$. But this just is the formulation of Disintegrability for $p'$, which is equivalent to Conglomerability. Thus, giving up Conglomerability would require finding a new version of the Law of Total Probability that doesn't have these features, to use in defining Jeffrey Conditionalization.

Considering the role of conditional probability in decision theory, Conglomerability is also supported by a Dutch book argument. The basic idea is given by Billingsley (1995, p. 431). Basically, any sort of reasoning to a foregone conclusion (as violations of Conglomerability allow) will make for guaranteed changes in one's betting prices that can be exploited by someone who knows one's updating rule. Rescorla (2018) has given a more complete Dutch book argument, including converse theorems proving that Conglomerability suffices for immunity to this sort of Dutch book.

There is also an accuracy-based argument for Conglomerability. Some authors have suggested that the right way to think of degree of belief is as aiming at the truth. Once we have a reasonable notion of "accuracy" that measures closeness to the truth, we can then derive norms for degree of belief from principles of maximizing accuracy (Joyce, 1998; Greaves & Wallace, 2006; Pettigrew, 2016). As it turns out, an update plan for learning which member of a partition is true maximizes expected accuracy iff it satisfies Conglomerability with respect to that partition (Easwaran, 2013a).

None of these arguments is fully definitive. It is possible to reject the importance of Dutch books and accuracy conditions for degree of belief. It is conceivable that an alternative formulation of the Law of Total Probability allows for a generalization of Jeffrey Conditionalization (or that Jeffrey Conditionalization is not the right update rule). And perhaps reasoning to a foregone conclusion is not so bad for updating. And all of these problems are perhaps less bad for physical or chance interpretations of probability than for Bayesian interpretations of one sort or another. Thus, if it is very important that conditional probability really be a two-place function rather than depending on a partition as well, then there is motivation to pursue Coherent Conditional Probability.

Thus the question becomes just how bad the costs are of Regular Conditional Probabilities, with their extra parameter. Some have said that an event alone must be sufficient to determine a posterior probability

distribution, and that the fact of the partition from which the event was drawn can't be relevant. "This approach [Regular Conditional Probability] is unacceptable from the point of view of the statistician who, when given the information that $A = B$ has occurred, must determine the conditional distribution of $X_2$" (Kadane et al., 1986). This is most plausible for uses of conditional probability in update by conditionalization, where one just learns a new piece of information, and apparently doesn't learn anything about the partition from which this information was drawn.

However, I claim that by considering the situation in a bit more detail, there will always be a partition that is relevant in any application of conditional probability. Billingsley (1995, end of section 33) brings this out with a juxtaposition of three exercises. The first two exercises involve consideration of the Borel paradox with a point on the surface of a sphere, and a version involving two independent normally distributed random variables. The third exercise juxtaposes the effect in these exercises of the same information presented in two different ways (a great circle presented as one from the family of longitudes, or as the equator from a family of latitudes; the fact of two random variables being equal as a piece of information about their difference, or as a piece of information about their ratio) with a classic probability puzzle.

Three prisoners are in a cell and two will be executed in the morning. Prisoner 3 asks the guard to tell him which of 1 or 2 will be executed (since at least one of them will) and on hearing the answer reasons that his chance of survival has gone up from 1/3 (as one of three prisoners, two of whom will be executed) to 1/2 (as one of two prisoners, one of whom will be executed). But of course, as anyone who has considered the similar "Monty Hall" problem can recognize, this reasoning ignores the fact that "Prisoner 1 is executed" and "Prisoner 2 is executed" do not form a partition, since it is possible for both to be true. The relevant learning situation is one in which the partition is "The guard says prisoner 1 will be executed" and "The guard says prisoner 2 will be executed." If these two answers are equally likely conditional on prisoner 3 surviving, then in fact the probability of survival is unchanged by this update.

This sort of example shows that even in elementary cases, we need to be careful about only updating on evidence by conditionalization in cases where it is clear that the evidence is drawn from a partition. To properly take this into account, we must be able to figure out what partition the evidence was drawn from. For Jeffrey Conditionalization, the partition is in fact part of the specification of the update situation, so this is clearer. Thus, I claim that for the first two uses of Bayesian probability (update by conditionalization or Jeffrey Conditionalization) the partition relativity of Regular Conditional Probabilities is no problem. There are some authors who argue that update situations don't always involve evidence that comes

from a partition (Schoenfield, 2016; Gallow, 2016). But I think that at least for scientific cases where evidence comes as the result of the performance of an experiment, the partition is implicit in the experimental setup. This is especially so in cases where the evidence was something that antecedently had probability 0, which are the only cases in which the issue of how to conditionalize arises.

For the uses of conditional probability in the measurement of confirmation, we have to look both at posterior probabilities and likelihoods. That is, we should be looking at probabilities of hypotheses conditional on evidence (as for updating) and for probabilities of evidence conditional on hypotheses. In this case, because of the Problem of Old Evidence (presented by Glymour, 1980, and classified and investigated at length by Eells, 1985), we must be considering conditional probabilities given before the experiment is actually performed. In order to properly compare and contrast the effect of different possible pieces of evidence, or different experiments, on different hypotheses, we must have a sense of the possible experiments, the possible pieces of evidence they could result in, and the possible hypotheses under consideration. This is particularly clear in cases where we are interested in confirmation, disconfirmation, and independence of hypotheses about random variables rather than just single propositions. A scientist who is interested in measuring the value of some physical, social, or biological parameter is going to have a whole family of propositions about its value that each may be confirmed, disconfirmed, or independent of the evidence received, and this family will define a partition for the relevant likelihoods.

For decision-theoretic cases, the relevant conditional probabilities are probabilities of outcomes conditional on actions. Here again it seems plausible that the set of actions available to an agent forms a partition. If this is right, then the relativization to a partition just brings out a feature that is already important to the situation. Thus, just like with the other Bayesian applications of conditional probability, I claim that there is no problem to the three-place formulation of conditional probability required by Regular Conditional Probabilities.

Even once we see that conditional probability depends on the partition from which the conditioning event was drawn, we might worry about how the *description* of events and partitions can affect the actual value. Rescorla (2015) argues that we should think of the same event drawn from a different partition as having something like a different "sense," so that these are just Frege puzzles of a sort. I'm not convinced that this is the right way to understand things, because the difference in conditional probability persists even when everyone involved recognizes that the same conditioning event is a member of multiple partitions. But I think that some reasoning of this sort can dissolve some worries.

Some have also worried that by redescribing the probability space, we might be able to make one partition look like another, so that we can get conflicting requirements for the same conditional probability. But Gyenis, Hofer-Szabó, and Rédei (2016) show that this is impossible—any reparameterization of a set of events and a partition gives rise to some other description on which the mathematical requirements of Conglomerability and Disintegrability give the same results.

In addition to the obvious challenge in terms of relativization, there is also a question of whether Regular Conditional Probabilities require Countable Additivity. Classic results (such as the Radon–Nikodym Theorem, or Theorem 33.3 of Billingsley, 1995) show that when the propositions involved are just about random variables, relativization of conditional probability to a partition as well as a conditioning event is sufficient to allow Conglomerability to hold even when Additivity fails at uncountable cardinalities. However, every existence theorem I know of assumes Countable Additivity. I have not investigated the proofs of Countable Additivity from Countable Conglomerability in enough detail to be sure that they hold up when conditional probabilities are allowed to vary as the partition changes. Thus, if considerations like the de Finetti lottery motivate rejection of Countable Additivity, then there may be further problems for Regular Conditional Probabilities. But as I have argued elsewhere, there are independent reasons to accept Countable Additivity that don't generalize to higher cardinalities (Easwaran, 2013b).

As the reader can probably see, I favor Regular Conditional Probabilities over Coherent Conditional Probabilities. But in the remainder of the paper, I will put forward mathematical theories of both types so that the reader can judge for herself what the appropriate uses of each might be.

## 2 REGULAR CONDITIONAL PROBABILITIES

### 2.1 *Formal Theory*

Regular Conditional Probabilities are a central motivation for the Kolmogorov (1950) axiomatization of probability. There is some set $\Omega$ of "possibilities," and the bearers of probability are subsets of this set. (Different interpretations of probability will interpret these possibilities and sets of them differently.) Not every subset of the space of possibilities is a bearer of probability, but there is some collection $\mathcal{F}$ of them that are. $\mathcal{F}$ is assumed to be a "$\sigma$-algebra" or "$\sigma$-field," which means that the empty set is an element of $\mathcal{F}$, the complement of any element of $\mathcal{F}$ is an element of $\mathcal{F}$, and if $A_i$ for $i \in \mathbb{N}$ are any countable collection of elements of $\mathcal{F}$, then $\bigcup_{i\in\mathbb{N}} A_i$ is also an element of $\mathcal{F}$. (This restriction to closure only under countable unions and complements is quite natural for the propositions

implicitly grasped by a finite mind, though one might want to restrict further to computably-definable sets or the like.)

Finally, there is a function $p$ assigning real numbers to all and only the elements of $\mathcal{F}$ subject to the following rules. For any $A \in \mathcal{F}$, $p(A) \geq 0$; $p(\Omega) = 1$; and if $A_i$ for $i \in \mathbb{N}$ are any countable collection of *disjoint* elements of $\mathcal{F}$, then $p(\bigcup_{i\in\mathbb{N}} A_i) = \sum_{i\in\mathbb{N}} p(A_i)$. That is, the probability function satisfies Countable Additivity. We refer to the triple $(\Omega, \mathcal{F}, p)$ as a *probability space*.

For any non-empty set $\Omega$, there are of course multiple different $\sigma$-algebras of subsets of that space. Trivially, the set $\{\emptyset, \Omega\}$ is always the minimal $\sigma$-algebra on $\Omega$, while the full power set consisting of *all* subsets of $\Omega$ is always the maximal $\sigma$-algebra on $\Omega$. But usually, $\mathcal{F}$ is some algebra other than these two. We say that a set $A$ is "$\mathcal{A}$-measurable" iff $A$ is an element of $\mathcal{A}$. If $\mathcal{A}$ and $\mathcal{B}$ are any two $\sigma$-algebras on $\Omega$, and every element of $\mathcal{A}$ is $\mathcal{B}$-measurable, then we say that $\mathcal{A}$ is a "sub-$\sigma$-algebra" of $\mathcal{B}$.

We often consider functions assigning a real number to every element of $\Omega$. If $V$ is such a function, then we say that $V$ is a *random variable*, or that it is *$\mathcal{F}$-measurable*, iff for all rational values $x$, the set $\{\omega \in \Omega \colon V(\omega) < x\}$ is $\mathcal{F}$-measurable. The set $\{\omega \in \Omega \colon V(\omega) \in S\}$ is often just written as $V(\omega) \in S$ or even $V \in S$, so for $V$ to be $\mathcal{F}$-measurable just is for $p(V < x)$ to exist for all rational values $x$, just as in Section 1.4.2. Furthermore, since the rational values are a countable and dense subset of the real numbers, the fact that $\mathcal{F}$ is closed under countable unions and complements means that $p(V = x)$, $p(V \geq x)$ and any other probability simply expressible in terms of values of $V$ exist as well.

As in Section 1.4.2, we can define the integral $\int_A V \, \mathrm{d}p$ for bounded random variables $V$. This definition proceeds in two parts. If $V$ only takes finitely many values on points in $A$, we say that $\int_A V \, \mathrm{d}p = \sum x \cdot p(A \cap (V = x))$, where the sum ranges over the finitely many values that $V$ takes on. Otherwise, we define $\int_A V \, \mathrm{d}p = \sup_{V' < V} \int_A V' \, \mathrm{d}p$, where the supremum ranges over all random variables $V'$ that take on only finitely many values in $A$, and such that whenever $\omega \in A$, $V'(\omega) < V(\omega)$.

With these definitions, I can finally give the official definition of a Regular Conditional Probability.

**Definition 10** *A* Regular Conditional Probability *is a three-place real-valued function $p(B \mid \mathcal{A})(\omega)$ satisfying the following three conditions:*

1. *Fixing a $\sigma$-algebra $\mathcal{A} \subseteq \mathcal{F}$ and $\omega \in \Omega$ defines a function of $B$ satisfying the probability axioms (that is, it is non-negative for all $B \in \mathcal{F}$, it takes the value 1 when $B = \Omega$, and it is Countably Additive).*

2. *Fixing a $\sigma$-algebra $\mathcal{A} \subseteq \mathcal{F}$ and a measurable set $B$ defines an $\mathcal{A}$-measurable function of $\omega$.*

3. *For any fixed $\sigma$-algebra $\mathcal{A} \subseteq \mathcal{F}$ and an $\mathcal{F}$-measurable set $B$, and for $A \in \mathcal{A}$,*
$$\int_A p(B \mid \mathcal{A})(\omega)\, \mathrm{d}p = p(B \cap A).$$

In Section 2.2 I will discuss how this notion relates to the three-place function $p(B \mid A, \mathbf{A})$ of conditional probability mentioned earlier. The basic idea of each condition is as follows. Condition 1 will ensure that conditioning on a single event relative to a single partition yields a probability function. Condition 2 will ensure that we really are conditioning on an event $A$ from the partition $\mathbf{A}$. Condition 3 will ensure that $p(B \mid A, \mathbf{A})$ satisfies Disintegrability (and thus Conglomerability). But for now I will just discuss a few formal features this mathematical function has.

As a first example, consider a probability space defined by a joint probability density for two random variables. That is, we can consider $X$ and $Y$ as two random variables, and let $\Omega = \mathbb{R}^2$, where the element $\omega = (\omega_X, \omega_Y)$ of $\Omega$ represents the possibility of $X = \omega_X$ and $Y = \omega_Y$. $\mathcal{F}$ is the $\sigma$-algebra generated by the set of sets $X < x$ and $Y < y$. (This algebra is known as the collection of "Borel sets," which is a subset of the Lebesgue-measurable sets, but sufficient for our purposes.) To say that the probability is defined by a joint probability density means that there is a measurable function $d(x, y)$ such that

$$p((x_1 < X < x_2) \cap (y_1 < Y < y_2)) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} d(x, y)\, \mathrm{d}y\, \mathrm{d}x,$$

where the integrals here are ordinary real-valued integrals. (This definition of probability over the rectangular boxes suffices to determine the probability of every measurable set.)[10]

If $\mathcal{X}$ is the $\sigma$-algebra generated by the set of sets $X < x$, then we can define a Regular Conditional Probability $p(B \mid \mathcal{X})(\omega)$ as follows. Let

$$p((x_1 < X < x_2) \cap (y_1 < Y < y_2) \mid \mathcal{X})(\omega) = \frac{\int_{y_1}^{y_2} d(\omega_X, y)\, \mathrm{d}y}{\int_{-\infty}^{\infty} d(\omega_X, y)\, \mathrm{d}y},$$

if $x_1 < \omega_X < x_2$ and 0 otherwise. (I use $\omega_X$ to represent the fixed value $X$ takes at $\omega$, while I use $y$ as the bound variable of the integral.) Again, because the rectangles $(x_1 < X < x_2) \cap (y_1 < Y < y_2)$ generate the whole $\sigma$-algebra, this suffices to define the conditional probability $p(B \mid \mathcal{X})(\omega)$ for all measurable sets $B$. Note that the values $y_1$ and $y_2$ enter on the right as limits of an integral, while the values $x_1$ and $x_2$ just determine when

10 Note that since we have assumed there is an unconditional probability function $p$, then we have assumed that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d(x, y)\, \mathrm{d}y\, \mathrm{d}x = 1$. In Section 3.2.5, when discussing Rényi's theory of conditional probability, I will allow this integral to be infinite instead, to capture the statistical theory of "improper priors."

the probability is 0. This is because the point $(\omega_X, \omega_Y)$ with respect to the $\sigma$-algebra $\mathcal{X}$ represents the set of all points with $X = \omega_X$ and any value of $Y$, and the rectangle either intersects this line at all points from $y_1$ to $y_2$ or none of them. Intuitively, the numerator of the right side says how much density is concentrated at $y_1 < Y < y_2$ and $X = \omega_X$, while the denominator normalizes this to account for how much density is at $X = \omega_X$ generally. It is tedious, but possible to check that this definition satisfies the three conditions to be a Regular Conditional Probability.[11]

The Borel paradox can be thought of as a special case of this example. If $X$ represents the longitude (from $-\pi$ to $\pi$) and $Y$ represents the latitude (from $-\pi/2$ to $\pi/2$), then the uniform unconditional probability is given by the density function $d(x, y) = \frac{\cos y}{4\pi}$ when $-\pi < x < \pi$ and $-\pi/2 < y < \pi/2$, and 0 otherwise. Using the above formula, we calculate that

$$p(y_1 < Y < y_2) \mid \mathcal{X})(\omega) = \frac{\int_{y_1}^{y_2} \frac{\cos y}{4\pi} dy}{1/2\pi} = \frac{\sin y_2 - \sin y_1}{2}.$$

By parallel reasoning, we calculate that

$$p(x_1 < X < x_2) \mid \mathcal{Y})(\omega) = \frac{\int_{x_1}^{x_2} \frac{\cos \omega_Y}{4\pi} \, \mathrm{d}x}{\cos \omega_Y / 2} = \frac{x_2 - x_1}{2\pi}.$$

That is, conditional on lines of longitude, probability is concentrated near the equator, while conditional on lines of latitude, probability is uniform.

If we want to use this sort of technique to figure out other Regular Conditional Probabilities for other sub-$\sigma$-algebras, we can often do this, if the new algebra is related to the old one by a change of coordinates. This will work if the probability space is defined by two random variables $X$ and $Y$, and there are two other random variables $f_1$ and $f_2$, such that the values of $f_1$ and $f_2$ are uniquely determined by the values of $X$ and $Y$, and vice versa. For instance, we might have $f_1 = X - Y$ and $f_2 = Y$, or $f_1 = X/Y$ and $f_2 = Y$ (if $Y = 0$ is impossible), or $f_1$ and $f_2$ as latitude and longitude in a different set of coordinates than $X$ and $Y$. In such a case, we can consider $f_1$ and $f_2$ as functions of the values of $X$ and $Y$, and represent points in $\Omega$ not as $(\omega_X, \omega_Y)$, but as $(f_1(\omega_X, \omega_Y), f_2(\omega_X, \omega_Y))$.

Assuming the functions $f_1$ and $f_2$ are measurable, we get a new density function given by

$$d(\omega_X, \omega_Y) \cdot \left[ \frac{\partial f_1(x, \omega_Y)}{\partial x} \frac{\partial f_2(\omega_X, y)}{\partial y} - \frac{\partial f_2(x, \omega_Y)}{\partial x} \frac{\partial f_1(\omega_X, y)}{\partial y} \right].$$

11 To check the third condition, it's useful to note that the $A \in \mathcal{X}$ are generated by the sets $x_1 < X < x_2$, and the probability of these sets is given by integrals like the denominator of the right-hand-side, so that this denominator cancels in the integration, leaving just the integral of the numerator over $X$, which is how we defined the unconditional probability in the first place.

This quantity on the right is the Jacobian associated with the relevant change of variables. When $f_1(\omega_X, \omega_Y) = \omega_X$ and $f_2(\omega_X, \omega_Y) = \omega_Y$, so that the "new" variables are the same as the old, the Jacobian is equal to 1, so the density is unchanged, as expected. But the fact that this Jacobian is not generally equal to 1 indicates that corresponding points in the two representations of the probability space will have different densities with respect to the two different sets of variables. Thus, even if one value of one variable occurs exactly when a corresponding value of a different variable occurs (such as $X = 0$ occurring iff $X/Y = 0$, or latitude is 0 in one set of coordinates iff longitude is 0 in another set of coordinates), the densities may have been transformed in some non-uniform way, so the Regular Conditional Probability may take different values.

A slightly different introduction to this sort of method is discussed by Chang and Pollard (1997). They argue that in most cases where Regular Conditional Probabilities are of interest, they can be calculated by a method like this one. Although their discussion is still quite technical, it may be more usable and friendly than some others.

## 2.2 *Philosophical Application*

As before, I define a "partition" to be a collection $\mathbf{A}$ of subsets of $\Omega$ such that every member of $\Omega$ is in exactly one member of $\mathbf{A}$. In Section 1.4.3, I argued that in order to maintain Conglomerability, while respecting the roles of conditional probability as posterior for conditionalization, or Jeffrey update, or as likelihood, or as action probability for decision theory, we need a notion of conditional probability that defines $p(B \mid A, \mathbf{A})$ whenever $\mathbf{A}$ is a partition. However, the formal theory given above defined a random variable $p(B, \mathcal{A})(\omega)$, where $\mathcal{A}$ is a sub-$\sigma$-algebra rather than a partition, and where $\omega$ is an element of $\Omega$ rather than a subset of it. In this section, I show that the formal definition of a Regular Conditional Probability is sufficient to give us what we need.

Partitions can be related to $\sigma$-algebras in two importantly different ways. One is that we can say that a $\sigma$-algebra $\mathcal{B}$ is *generated* by a partition if it is the *smallest* $\sigma$-algebra with respect to which every element of $\mathbf{A}$ is measurable. In this case, $\mathcal{B}$ consists of the set of all unions of countably many elements of $\mathbf{A}$, and their complements.[12] However, in many cases, the more useful $\sigma$-algebra to consider is a slightly different one. I will say that a $\sigma$-algebra $\mathcal{B}$ is *compatible* with a partition $\mathbf{A}$ iff every element of $\mathbf{A}$ is

12 We also talk about $\sigma$-algebras generated by collections of subsets other than a partition, and in those cases there can often be much more complex elements of the generated $\sigma$-algebra, such as countable unions of complements of countable unions of complements of countable unions of elements. But in the case of a partition, these more complex elements already exist just at the level of countable unions or their complements.

an element of $\mathcal{B}$, and no proper subset of an element of $\mathbf{A}$ is an element of $\mathcal{B}$, except for the empty set.[13] Then, if $\mathcal{B}$ is any $\sigma$-algebra and $\mathbf{A}$ is any partition, I will say that the *restriction of $\mathcal{B}$ to $\mathbf{A}$* is the *largest* sub-$\sigma$-algebra of $\mathcal{B}$ that is compatible with $\mathbf{A}$. This consists of all elements of $\mathcal{B}$ whose intersection with any element of $\mathbf{A}$ is either empty or the full element of $\mathbf{A}$—it is the set of all $\mathcal{B}$-measurable sets that don't crosscut any element of $\mathbf{A}$.

Given these definitions, for $A, B \in \mathcal{F}$ and $\mathbf{A} \subseteq \mathcal{F}$ a partition containing $A$, I will define $p(B \mid A, \mathbf{A})$ as $p(B \mid \mathcal{A})(\omega)$, where $\omega$ is any element of $A$ and $\mathcal{A}$ is the restriction[14] of $\mathcal{F}$ to $\mathbf{A}$. If $A$ is empty, then $p(B \mid A, \mathbf{A})$ is undefined. This corresponds to the fact that conditional probability is intended to be an indicative conditional for updating rather than revision of beliefs, as discussed in Section 1.3. Otherwise, since $p(B \mid \mathcal{A})(\omega)$, considered as a function of $\omega$, is required to be $\mathcal{A}$-measurable, it must be constant on the atoms of $\mathcal{A}$. But because $\mathcal{A}$ is the restriction of $\mathcal{F}$ to $\mathbf{A}$, the atoms are the elements of $\mathbf{A}$. Since $A$ is an element of $\mathbf{A}$, this means that it doesn't matter which $\omega \in A$ is chosen. Thus, as long as $p(B \mid \mathcal{A})(\omega)$ is a well-defined function, so is $p(B \mid A, \mathbf{A})$, whenever $A$ is non-empty. The stipulations in the definition of a Regular Conditional Probability then mean that $p(B \mid A, \mathbf{A})$ satisfies the probability axioms (including Countable Additivity) when $A$ and $\mathbf{A}$ are fixed, and that Conglomerability is satisfied over $\mathbf{A}$. Thus, if conditional probability should be defined relative to any partition, and Conglomerability must be satisfied, then conditional probability must be related to a Regular Conditional Probability in this way.

## 2.3 *Existence and Uniqueness of Regular Conditional Probabilities*

The question motivated by the arguments of Section 1.3 is whether unconditional probabilities suffice to determine a notion of conditional probability, or whether conditional probability should be taken as fundamental. The mathematical definition of a Regular Conditional Probability as $p(B \mid \mathcal{A})(\omega)$ is as a function that satisfies some axioms connecting it to the unconditional probability space $(\Omega, \mathcal{F}, p)$. In some cases, we have been able to demonstrate that Regular Conditional Probabilities exist. If they don't exist in probability spaces that are philosophically important, then

13 In more standard terminology, $\mathbf{A}$ consists of the "atoms" of $\mathcal{B}$, where an atom of a $\sigma$-algebra is any non-empty element of the $\sigma$-algebra such that no non-empty proper subsets are also members of the $\sigma$-algebra. Not every $\sigma$-algebra has atoms, but if there are any atoms, they are disjoint. The atoms form a partition iff every element of the space is a member of some atom, in which case the $\sigma$-algebra is said to be "atomic."

14 Section 2.3.2 will show what goes wrong if we try to use the sub-$\sigma$-algebra generated by $\mathbf{A}$ instead of the restriction to it.

Conglomerability must be given up. And if Regular Conditional Probabilities are not unique, then we must either accept that conditional probability is at least as fundamental as unconditional probability, or give some further conditions that suffice to uniquely determine the Regular Conditional Probability uniquely. In this section I will consider some mathematical problems of particular Regular Conditional Probabilities and argue that they don't arise in philosophical application, so they will always exist and have the desired features. Furthermore, I will show that unconditional probability is almost sufficient to define all conditional probabilities in the relevant probability spaces, and give some ideas of what else might suffice to define conditional probability uniquely from unconditional probability.

### 2.3.1 *In Bad Sub-$\sigma$-algebras There Is No Regular Conditional Probability*

It is mathematically well-known that there are probability spaces $(\Omega, \mathcal{F}, p)$ and sub-$\sigma$-algebras $\mathcal{A}$ for which there is no Regular Conditional Probability. A classic example is the case where $\Omega$ is the set $[0,1]$ of real numbers between 0 and 1, $\mathcal{A}$ is the set of all Borel subsets of this set, $\mathcal{F}$ is generated by $\mathcal{A}$ plus one set that is not Lebesgue-measurable, and $p$ is Lebesgue measure on $\mathcal{A}$ and assigns probability $1/2$ to the additional set generating $\mathcal{A}$. (This example is discussed in Billingsley, 1995, Exercise 33.11.)

However, Theorem 33.3 of Billingsley (1995) states that when $\mathcal{F}$ is the $\sigma$-algebra generated by the values of a random variable, this problem can never arise. There will always be a Regular Conditional Probability for every sub-$\sigma$-algebra. This result generalizes to cases where $\mathcal{F}$ is the $\sigma$-algebra generated by the values of finitely many random variables, as appears to be the case for most scientific applications of probability.

Furthermore, due to the finitistic limits of the human mind, I claim that this in fact includes all epistemically relevant cases. As I suggested near the end of Section 1.3, I think the right interpretation of human finitude doesn't mean that the probability space is finite. Rather, it means that the probability space is generated by the countably many sentences of some finitary language. I claim that the sentences in this language fit within the $\sigma$-algebra over this space generated by a particular artificial random variable.

To see this, define the random variable $T$ by enumerating the sentences of the language as $\phi_i$ and letting

$$T(\omega) = \sum_{\phi_i \text{ is true}} \frac{1}{2^i}.$$

Any possibility $\omega$ will make infinitely many sentences true and infinitely many sentences false, and no two such possibilities can result in the same real value, so this random variable distinguishes all possible worlds. We

need to check further that the set of values that are logically consistent is itself measurable. But by the Compactness Theorem of first-order logic, any logically inconsistent set contains one of the countably many logically inconsistent finite sets, and each of these sets is an intersection of finitely many closed sets of values. Thus, the set of consistent values is the complement of a countable union of closed sets, and is thus measurable. Thus, I claim that any epistemically reasonable probability space uses a $\sigma$-algebra generated by a random variable, conditionalized on a measurable set. Thus, Theorem 33.3 of Billingsley (1995) entails that Regular Conditional Probabilities exist.

Even without this sort of argument, the existence theorem can be generalized. These generalizations are investigated by Hoffmann-Jørgensen (1971), Faden (1985), Pachl (1978).

### 2.3.2 *In Bad Sub-algebras, the Regular Conditional Probability Behaves Badly*

Another problem that sometimes arises is highlighted by Blackwell and Dubins (1975) and Seidenfeld, Schervish, and Kadane (2001). They seem to show that in certain partitions $\mathbf{A}$, there is an event $A$ with $p(A \mid A, \mathbf{A}) = 0$, which would seem to be quite bad. However, I claim that this problem only arises in cases where $\mathbf{A}$ is used in a mathematically improper way.

The mathematical result they show is that $p(B \mid \mathcal{A})(\omega) = 0$ even though $\omega \in B$. As an example, let $\Omega$ be the set $[0,1]$ of real numbers between 0 and 1, let $\mathcal{F}$ be the collection of all Borel subsets of this set, and let $p$ be the standard Lebesgue measure on $\mathcal{F}$. Let $\mathcal{A}$ be the collection of all countable subsets of $[0,1]$ and their complements. It is straightforward to check that $p(B \mid \mathcal{A})(\omega) = p(B)$ is a Regular Conditional Probability.[15] However, if $B = \{\omega\}$ (or any other countable set containing $\omega$) then $p(B \mid \mathcal{A})(\omega) = p(B) = 0$. Given my translation of $p(B \mid A, \mathbf{A})$, this would seem to mean that $p(\{\omega\} \mid \{\omega\}, \mathbf{A}) = 0$, where $\mathbf{A}$ is the partition into singletons.

However, this is the point at which the distinction between the $\sigma$-algebra *generated* by $\mathbf{A}$ and the *restriction* of $\mathcal{F}$ to $\mathbf{A}$ is important. The $\sigma$-algebra $\mathcal{A}$ above is the algebra generated by the partition into singletons, but it is *not* the restriction of $\mathcal{F}$ to the partition into singletons. The restriction of $\mathcal{F}$ to the partition into singletons just is $\mathcal{F}$ (as it is for any $\mathcal{F}$—recall that the restriction of $\mathcal{F}$ includes all elements of $\mathcal{F}$ that do not crosscut any element of the partition, and no set crosscuts a singleton). Although $p(B \mid \mathcal{A})(\omega) = p(B)$ is a Regular Conditional Probability, it is straightforward to show that the parallel does not work for $p(B \mid \mathcal{F})(\omega)$. In fact, any Regular Conditional

15 The first two conditions are trivial. The third condition requires that $\int_A p(B \mid \mathcal{A})(\omega)\, \mathrm{d}p = p(A \cap B)$ for all $A \in \mathcal{A}$. However, since $p(B \mid \mathcal{A})(\omega) = p(B)$ for all $\omega$, the left side of the integral just is $p(A)p(B)$. But if $A$ is countable, then $p(A) = 0$, as does $p(A \cap B)$, while if $A$'s complement is countable, then $p(A) = 1$ and $p(A \cap B) = p(B)$.

Probability for this conditioning algebra must have a set $C$ with $p(C) = 1$ such that whenever $\omega \in C$, $p(B \mid \mathcal{F})(\omega) = 1$ if $\omega \in B$ and 0 otherwise, as expected. And Theorem 2 of Blackwell and Dubins (1975) and Theorem 1 of Seidenfeld et al. (2001) show that this is quite general. Whenever $\mathcal{A}$ is countably generated, for any Regular Conditional Probability $p(B \mid \mathcal{A})(\omega)$, there is a set $C$ with $p(C) = 1$ such that whenever $\omega \in C$ and $B \in \mathcal{A}$, $p(B \mid \mathcal{A})(\omega) = 1$.[16] Thus, in my translation, $p(B \mid A, \mathbf{A}) = 1$ if $A \subseteq B$, as expected, whenever the restriction of $\mathcal{F}$ to $\mathbf{A}$ is countably generated. This will automatically be the case if $\mathbf{A}$ is the partition of possible values of a random variable. But I claim that it should hold generally for any partition that is graspable by a finite human mind.

### 2.3.3 *The Regular Conditional Probability is* Almost *Unique*

Now that we have established that Regular Conditional Probabilities exist and are well-behaved, it remains to see when they are uniquely determined by the unconditional probability space $(\Omega, \mathcal{F}, p)$. It turns out that the answer is *never* in any interesting case. However, the different Regular Conditional Probabilities that exist are *almost* identical in a natural sense. Furthermore, for some sets of niceness conditions, exactly one of them will be nice, and this can be designated as the correct one.

If $p(B \mid \mathcal{A})(\omega)$ is one Regular Conditional Probability, and $S \in \mathcal{A}$ is any set with $p(S) = 0$, then we can let $p'(B \mid \mathcal{A})(\omega) = p(B \mid \mathcal{A})(\omega)$ whenever $\omega \notin S$ and replace the function with any other probability function we like within $S$, and the result is also a Regular Conditional Probability. This is because the only constraint on the values of a Regular Conditional Probability are through its integrals, and changing a function on a set of probability 0 does not change any of its integrals. Translating to $p(B \mid A, \mathbf{A})$, this means that we can change the values of the conditional probability function on any collection of $A \in \mathbf{A}$ whose total probability is 0 and still satisfy Conglomerability.

Conversely, if $p(B \mid \mathcal{A})(\omega)$ and $p'(B \mid \mathcal{A})(\omega)$ are two Regular Conditional Probabilities for a given unconditional probability, then we can show that for any $B$ and $\mathcal{A}$, the set of $\omega$ for which they differ must have probability 0. If it had positive probability, then there would be some $\epsilon$ such that the set $C$ of $\omega$ on which they differ by at least $\epsilon$ would have positive probability, and would be a member of $\mathcal{A}$. But this would contradict the condition that $\int_C p(B \mid \mathcal{A})(\omega)\, \mathrm{d}p = p(B \cap C) = \int_C p'(B \mid \mathcal{A})(\omega)\, \mathrm{d}p$. Thus,

16 Of course, this assumes that a Regular Conditional Probability exists, which requires that $\mathcal{F}$ be a nice algebra, such as the algebra generated by a random variable. See Blackwell (1956) for more on these conditions. In fact, for these sorts of spaces, Yu (1990) proves that existence of the relevant function can be proven in the system "$\mathrm{ACA}_0$" of reverse mathematics, so that strong set-theoretic hypotheses like the Axiom of Choice are not required.

although the Regular Conditional Probability is not exactly unique, it is in a sense "almost" unique. These different Regular Conditional Probabilities are often called "versions" of the Regular Conditional Probability for the given unconditional probability.

This almost uniqueness is not quite enough to satisfy the idea that conditional probability is defined by the unconditional probability function. However, in some cases there is a prospect that by specifying a further condition, we can pick out a unique version of the Regular Conditional Probability. For instance, consider the case of the Borel paradox. As I showed in Section 2.1, one version of the Regular Conditional Probability for this example can be generated by integrals of a probability density that also generates the unconditional probability. In this case, there is a *continuous* density function that generates the unconditional probability (namely, the density function that was given there, with $d(x, y) = \cos y$). Furthermore, it is easy to see that no other continuous density generates the same unconditional probability function. (If two continuous density functions differ at some point, then they must differ on some neighborhood of that point, which would have non-zero probability.) Thus, if an unconditional probability function is generated by some continuous density on the values of some random variables, then we can require that the version of the Regular Conditional Probability used be the one that is generated by this integral calculation from the unique continuous density that generates the unconditional probability.[17]

17 Oddly, if we just consider the partitions into longitudes through various choices of poles, we may be able to take advantage of this non-uniqueness to find a *Coherent* Conditional Probability that satisfies Disintegrability. If we assume the Axiom of Choice and the Continuum Hypothesis (or Martin's Axiom—both assumptions entail that the union of any collection of fewer than continuum-many sets with probability 0 is also a set of probability 0), then we can do the following. Choose some well-ordering of the points on the sphere such that each has fewer than continuum-many predecessors. For any great circle $A$, find the point $x \in A$ that comes earliest in this ordering. Let $p(B \mid A)$ take the value given by integration with respect to the continuous density where $x$ is chosen as the north pole of the coordinate system.
Now if we consider any particular partition into longitudes with $x$ as a pole, we can see that each line of longitude will give rise to a conditional probability that agrees with the one required for Disintegrability in this partition iff there is no point on the line earlier than $x$ in the chosen ordering. However, because of the way the ordering was set up, there are fewer than continuum-many points earlier than $x$ in the ordering, so the union of all the lines of longitude that contain such a point has probability 0. Thus, enough of the conditional probabilities agree with integration with respect to the relevant continuous density that Disintegrability is satisfied in this partition.
Of course, this particular method only satisfies Disintegrability over partitions into lines of longitude, and not into lines of latitude, or other partitions. Furthermore, the particular Coherent Conditional Probability produced over these conditioning events is highly asymmetrical and requires the Axiom of Choice for its construction. But it is useful to observe that this sort of construction is at least sometimes possible.

However, while I think it is not that implausible to think that all realistic epistemic spaces are generated by some density on the values of some random variables, I don't see any good reason to believe that there must always be a *continuous* density function that generates the unconditional probability. Perhaps there is some similar requirement that could be used to find the "right" Regular Conditional Probability to go along with any unconditional probability function. But I have no idea what that requirement might be. So for now, we have some reason to believe that the existence of uncountable (though countably generated) probability spaces, together with Conglomerability, force us to use Regular Conditional Probabilities, which suggests that conditional probability is in some sense at least as fundamental as unconditional probability. However, if one is only given the unconditional probability function, then for any countably-generated partition $\mathbf{A}$ one can find *some* Regular Conditional Probability $p(B \mid A, \mathbf{A})$ for all propositions $B$ on the elements of $\mathbf{A}$, and one can be sure that *almost all* of the values given by this function will line up with the "correct" conditional probability function. The question is just whether this "almost all" can be turned into "all," or whether conditional probability needs to be specified along with unconditional probability in defining a probability space.

## 3 COHERENT CONDITIONAL PROBABILITIES

Recall that Coherent Conditional Probability is conditional probability defined as a function just of two events, with no dependence on a partition or sub-$\sigma$-algebra or anything else. If Additivity fails at some level (possibly beyond the countable), then Conglomerability and Disintegrability will also fail. There are several different formal theories of Coherent Conditional Probability that have been proposed by philosophers, mathematicians, and statisticians. In this section I will describe three of the most prominent ones.

### 3.1 *Popper*

The first, which is both oldest and probably most familiar to philosophers, was developed by Karl Popper in his (1955). Popper considered this formulation of conditional probability important enough that he included a revised and simplified version in new appendices *iv and *v to the second edition of *The Logic of Scientific Discovery* (1959a). Popper's axiom system is particularly well-suited to an interpretation of probability as a logical (or even semantic) relation. But I claim that it is not sufficient for general epistemological applications, particularly for scientific purposes.

In this section I will describe Popper's later version of the system, and the features it has.

Popper postulates a finite or countable set of sentence letters $A, B, C, \ldots$, and two uninterpreted connectives—a binary connective '$\wedge$' and a unary connective '$\neg$'. (I have replaced his notation with a more modern one.) He then postulates a two-place conditional probability function mapping pairs of formulas in the language generated by these letters and connectives to real numbers. He then postulates six conditions on the function expressible with these uninterpreted connectives. (I will discuss these conditions later.) Finally, he defines unconditional probability in terms of conditional probability.

One of the important things Popper does along the way is to develop a probabilistic notion of equivalence. He says that two formulas $\phi$ and $\psi$ of the language are probabilistically equivalent iff replacing $\phi$ with $\psi$ anywhere in any statement of probability will yield the same value. He then proves that if two formulas are classically logically equivalent, then they are probabilistically equivalent. He doesn't explicitly assume commutativity and associativity for $\wedge$, or the double negation rule, or anything of that sort, but is able to derive probabilistic equivalents of them from his probability axioms.

Popper's axioms entail that some elements $\psi$ are such that for all $\phi$, $p(\phi \mid \psi) = 1$. (Among other things, this means that $p(\neg\psi \mid \psi) = 1$!) Following van Fraassen (1976), we call such elements *abnormal* and all others *normal*. Popper's axioms entail that if $\chi$ is normal, then $0 \leq p(\phi \mid \chi) \leq 1$, and that $p(\phi \mid \chi) + p(\psi \mid \chi) = p(\neg(\neg\phi \wedge \neg\psi) \mid \chi) + p(\phi \wedge \psi \mid \chi)$, so that conditional on any normal event, we have a standard probability function. Furthermore, they entail that if $\psi$ is abnormal, then for any $\chi$, $p(\neg\psi \mid \chi) = 1$. Finally, they entail that whenever $\phi$ is a classical logical contradiction, $\phi$ is abnormal.

Importantly, this means that Popper's notion of conditional probability (like all the others I am aware of) is of no help in using conditionalization to represent belief *revision* rather than just update. Consider an update rule that says $p_{t'}(\phi \mid \psi) = p_t(\phi \mid \psi \wedge \chi)$, where $\chi$ is the conjunction of everything that one has learned between $t$ and $t'$. Now imagine a person who, between time 0 and time 1 learns $A$, and between time 1 and time 2 learns $\neg A$. If update can include revision of past learning (which implicitly means that learning is fallible), then this should result in something reasonable. However, what we see is that for any $\phi$ and $\psi$, $p_2(\phi \mid \psi) = p_1(\phi \mid \psi \wedge \neg A) = p_0(\phi \mid (\psi \wedge \neg A) \wedge A)$. But since $(\psi \wedge \neg A) \wedge A$ is a contradiction, it is abnormal. Thus, $p_0(\phi \mid (\psi \wedge \neg A) \wedge A) = 1$. So by updating on the negation of something that one previously learned, one's degrees of belief have become unusable, because all probabilities are equal to 1. This is why I focused in Section 1.3 on the role of infinity in generating events of

probability 0, rather than Hájek's examples of conditionalizing on the negation of something that has already been learned.

However, one important thing to note for Popper's system is that $p(\psi) = 0$ does *not* entail that $\psi$ is abnormal. However, if $p(\psi) = 0$ but $\psi$ is normal, then unconditional probabilities alone do not suffice to determine the probabilities conditional on $\psi$. Thus, conditional probability really is primitive in this system. For instance, consider models of Popper's axioms with sentence letters $A$ and $B$, with $p(A) = 1/2$ and $p(B) = 0$. Every formula of the language is classically equivalent to a contradiction, or to a disjunction of some of $A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B$. The stipulated values determine all the unconditional probabilities, and thus all the probabilities conditional on formulas of positive unconditional probability. However, it is consistent with these values that $A \wedge B$ and $\neg A \wedge B$ be either normal or abnormal. If both are abnormal, then so is $B$, and probabilities conditional on any of the three of them are all equal to 1. If one is abnormal and the other is normal, then probabilities of any formula conditional on the normal one are 1 or 0 depending on whether the formula is entailed by it or not. If both are normal, then any value for $p(A \mid B)$ is possible, but this value then suffices to determine the rest of the probabilities in the model.

And in fact, Kemeny (1955) proves that something like this holds fairly generally for finite languages. If we only have $n$ sentence letters, then there are $2^n$ "state descriptions" in the language (conjunctions of each sentence letter or its negation), and every formula is either a contradiction or equivalent to a disjunction of some of these. The Popper axioms are then equivalent to the following stipulation. There are $k$ functions $m_i$ for $i < k$, and each of these functions assign a non-negative real number to each state description. For each $m_i$, the sum of the values it assigns to the state descriptions is 1. For each state description $X$, there is at most one $m_i$ such that $m_i(X) > 0$. A proposition is abnormal iff it is either a contradiction, or it is a disjunction of state descriptions that are assigned value 0 by every $m_i$. If $\psi$ is normal, then let $i$ be the lowest number such that there is a state description $X$ with $m_i(X) > 0$ and $X$ entails $\psi$. Then

$$p(\phi \mid \psi) = \frac{\sum_{X \text{ entails } \phi \wedge \psi} m_i(X)}{\sum_{X \text{ entails } \psi} m_i(X)}.$$

In this system, unconditional probabilities are just equal to the sums of the values of $m_1$, but they put no constraints on the values of the succeeding functions, which are needed to define the full conditional probability function.

For infinite languages, things can be slightly more complicated. Consider a language with sentence letters $A_i$ for natural numbers $i$. Consider just the models $M_i$ where $M_i$ satisfies sentence $A_i$ and none of the others. It

is not hard to check that every formula of the language is either true in finitely many of these models and false in the rest, or false in finitely many of these models and true in the rest. If $\psi$ is true in infinitely many models, then let $p(\phi \mid \psi) = 0$ if $\phi$ is true in finitely many models and 1 otherwise. If $\psi$ is true in none of these models, then $\psi$ is abnormal. Otherwise, if $\psi$ is true in finitely many models, then define $p(\phi \mid \psi)$ as the ratio of the number of models in which $\phi \wedge \psi$ is true to the number of models in which $\psi$ is true. This definition satisfies Popper's axioms, but cannot be represented by a lexicographically ordered set of probability functions as Kemeny shows in the finite case. (This example is one that Halpern, 2009 attributes to Stalnaker.) Halpern also discusses a slight variant of this case where the probability function agrees with this one in all cases except where $\psi$ is true in finitely many models. In the variant, $p(\phi \mid \psi) = 1$ if $\phi$ is true in the *highest numbered* model in which $\psi$ is true, and 0 otherwise. This probability function also satisfies Popper's axioms but cannot be represented by a lexicographically ordered set of probability functions. But again, these functions have the same unconditional probabilities and the same abnormal propositions, but different conditional probabilities, so that conditional probability must be specified separately from unconditional probabilities.

Popper's six conditions are the following (Popper, 1959a, Appendix iv*).

1. For all $\phi, \psi$ there are $\chi, \theta$ with $p(\phi \mid \psi) \neq p(\chi \mid \theta)$.
2. If for all $\chi$, $p(\phi \mid \chi) = p(\psi \mid \chi)$, then for all $\theta$, $p(\theta \mid \phi) = p(\theta \mid \psi)$.
3. For all $\phi, \psi$, $p(\phi \mid \phi) = p(\psi \mid \psi)$.
4. $p(\phi \wedge \psi \mid \chi) \leq p(\phi \mid \chi)$.
5. $p(\phi \wedge \psi \mid \chi) = p(\phi \mid \psi \wedge \chi) p(\psi \mid \chi)$.
6. For all $\phi, \psi$, either $p(\phi \mid \psi) + p(\neg\phi \mid \psi) = p(\psi \mid \psi)$, or for all $\chi$, $p(\psi \mid \psi) = p(\chi \mid \psi)$.

In Appendix v* of Popper (1959a), he derives a sequence of consequences of these postulates. Importantly, he doesn't assume any logical features of $\wedge$ and $\neg$ in these derivations—he only uses the explicit probabilistic assumptions made above.

First, using condition 3, he defines $k = p(\phi \mid \phi)$ for any formula $\phi$. Using 4 and 5 he then proves that $k^2 \leq k$, so $0 \leq k \leq 1$. After a few more steps, he then proves that $0 \leq p(\phi \mid \psi) \leq k$ for any $\phi, \psi$. From this, he is then able to derive that $k = k^2$, so $k = 0$ or $k = 1$, but condition 1 rules out $k = 0$. Condition 4 then tells us that $1 = p(\phi \wedge \psi \mid \phi \wedge \psi) \leq p(\phi \mid \phi \wedge \psi)$, so $p(\phi \mid \phi \wedge \psi) = 1$. With condition 5 this then proves that $p(\phi \wedge \phi \mid \psi) = p(\phi \mid \psi)$. A bit more manipulation allows him to derive that

$p(\phi \wedge \psi \mid \chi) = p(\psi \wedge \phi \mid \chi)$, and that $p(\phi \wedge (\psi \wedge \chi) \mid (\phi \wedge \psi) \wedge \chi) = 1$, and after several more steps, that $p(\phi \wedge (\psi \wedge \chi) \mid \theta) = p((\phi \wedge \psi) \wedge \chi \mid \theta)$. Thus, he has derived that $\wedge$ is commutative and associative, up to probabilistic equivalence.

He then turns his attention to negation and derives several important results. First, he derives that $p(\neg(\phi \wedge \neg\phi) \mid \psi) = 1$. Then he derives that $p(\neg(\neg\phi \wedge \neg\psi) \mid \chi) = p(\phi \mid \chi) + p(\psi \mid \chi) - p(\phi \wedge \psi \mid \chi)$. If we introduce an abbreviation $\vee$ such that $\phi \vee \psi$ just stands for $\neg(\neg\phi \wedge \neg\psi)$, this becomes $p(\phi \vee \psi \mid \chi) = p(\phi \mid \chi) + p(\psi \mid \chi) - p(\phi \wedge \psi \mid \chi)$, which is a version of the standard law of Additivity. He then derives that $p(\phi \wedge (\psi \wedge \chi) \mid \theta) = p((\phi \wedge \psi) \wedge (\phi \wedge \chi) \mid \theta)$, and $p(\phi \wedge (\psi \vee \chi) \mid \theta) = p((\phi \wedge \psi) \vee (\phi \wedge \chi) \mid \theta)$. Using this, he derives that $p(\neg\neg\phi \wedge \psi \mid \chi) = p(\phi \wedge \psi \mid \chi)$ and that if $p(\phi \mid \chi) = p(\psi \mid \chi)$ then $p(\neg\phi \mid \chi) = p(\neg\psi \mid \chi)$. He then derives that $p(\phi \vee \phi \mid \psi) = p(\phi \mid \psi)$. And finally, he proves that if for all $\kappa$, $p(\phi \mid \kappa) = p(\psi \mid \kappa)$, and $p(\chi \mid \kappa) = p(\theta \mid \kappa)$, then for all $\kappa$, $p(\phi \wedge \psi \mid \kappa) = p(\chi \wedge \theta \mid \kappa)$.

With these conditions, he is then able to show that logically equivalent formulas are probabilistically equivalent, and derive the facts I mentioned above about abnormal formulas, and probabilities conditional on normal formulas.

For Popper, one of the important features of this characterization is that probability can play the role of giving the meanings of the logical symbols. This is quite a natural desideratum for a logical interpretation of probability, though it may not be as natural for other interpretations. This program is developed further by Field (1977), who gives a method for giving meanings to quantifiers (though this is substantially more clumsy than Popper's method for the connectives).

One thing to note about Popper's formalism is that infinitary versions of Additivity (and Conglomerability, and Disintegrability) can't even be *stated*, much less satisfied or violated. First, every formula is finite, so that even if the language is expanded by adding a disjunction symbol, there are no infinite disjunctions explicitly expressible in the language. Second, by the Compactness Theorem of propositional logic, no formula in this language is logically equivalent to an infinite disjunction of formulas expressible in the language unless it is also logically equivalent to a disjunction of finitely many of those disjuncts. One might wonder whether this holds for probabilistic equivalence, but probabilistic equivalence is only defined for formulas within the language, and infinite disjunctions aren't in the language, so the question doesn't arise.

While some might find this to be an advantage of the sentential formulation of probability, many have found it to be a limitation and have given what they call versions of Popper's system where the bearers of probability are sets rather than formulas of a language, and the operations are set intersection and complement rather than (uninterpreted) $\wedge$ and $\neg$

(Roeper & LeBlanc, 1999; Hájek & Fitelson, 2017). But since Popper's goal was at least partly to characterize the sentential operations in terms of probability, rather than using facts about sets to prove some results about probability, I think of these systems as significantly different.

Versions of these systems are given by van Fraassen (1976), Spohn (1986), McGee (1994), and Halpern (2009), among others. Because the bearers of probability are sets, these authors are able to prove more general characterizations than Kemeny. In particular, Spohn shows that if we add Countable Additivity to Popper's axioms, then these probabilities can always be represented as a lexicographically-ordered set of Countably Additive measures $m_i$. However, because of the results mentioned in Section 1.4.2, there must be failures of Conglomerability and Disintegrability in certain partitions, even if Countable Additivity is assumed. These authors also show several results relating these set-theoretic versions of Popper's system to probabilities involving infinitesimals (as discussed by Wenmackers, this volume). However, while McGee claims that the two systems are equivalent, Halpern shows that there are some subtleties to consider. But once we start looking at Countably Additive set-based systems that are like Popper's it is useful to consider a slightly more general formalization that includes all of the above as special cases.

## 3.2 *Rényi*

Alfréd Rényi gave the first English-language version of his system for conditional probability in his (1955), though it also appears briefly in the second chapter of the posthumous textbook (1970a) and is developed in somewhat greater detail in the second chapter of his (1970b). I will generally follow his (1955) in my discussion, though the structural requirements on $\mathcal{B}$ only appear in the later books. Some of the theory appears slightly earlier in publications in German or Hungarian.

Although philosophers often lump Popper and Rényi together, Rényi's early theory is much more flexible than Popper's. It does include a set-based version of Popper's system as a special case, but it also includes a version of Kolmogorov's Regular Conditional Probability as a special case as well. However, Rényi's major aim in developing his theory is to account for a very different application from either of these (and in fact, his later theory explicitly rules out non-trivial versions of Popper and Kolmogorov's systems in favor of these other applications). In statistical practice it is sometimes relevant to work with an "improper prior"—something much like a probability function, that can turn into a probability function by conditioning on some event, but for which the unconditional "probabilities" are infinite. This flexibility also allows Rényi's theory to include

actual relative frequencies, as a system where there is no unconditional probability and all probabilities are conditional.

### 3.2.1 *Overview*

The background theory for Rényi's conditional probabilities (just like for Regular Conditional Probabilities) is the traditional Kolmogorov axiomatization of probability. There is some set $\Omega$ of "possibilities," and the bearers of probability are subsets of this set. (Different interpretations of probability will interpret these possibilities and sets of them differently.) Not every subset of the space of possibilities is a bearer of probability, but there is some collection $\mathcal{A}$ of them that are. $\mathcal{A}$ is assumed to be a *$\sigma$-algebra* or *$\sigma$-field*, which means (as before) that the empty set is an element of $\mathcal{A}$, the complement of any element of $\mathcal{A}$ is an element of $\mathcal{A}$, and if $A_i$ for $i \in \mathbb{N}$ are any countable collection of elements of $\mathcal{A}$, then $\bigcup_{i\in\mathbb{N}} A_i$ is also an element of $\mathcal{A}$.[18]

$\mathcal{A}$ is the set of bearers of probability. But unlike in Popper's theory, not every bearer of probability can be conditioned on. Instead, Rényi considers a collection $\mathcal{B} \subseteq \mathcal{A}$ subject to the following conditions. For any $B_1$ and $B_2$ that are both in $\mathcal{B}$, $B_1 \cup B_2 \in \mathcal{B}$. There exists a countable sequence $B_i$ for $i \in \mathbb{N}$ of elements of $\mathcal{B}$ such that $\bigcup_{i\in\mathbb{N}} B_i = \Omega$. And $\emptyset \notin \mathcal{B}$. While $\mathcal{A}$ is a $\sigma$-algebra, $\mathcal{B}$ is a "bunch," that may lack complements and infinite unions, as well as $\Omega$, and definitely lacks the empty set.

He then defines a conditional probability function $p(A \mid B)$ for $A \in \mathcal{A}$ and $B \in \mathcal{B}$ to be any function satisfying the following conditions. For all $A \in \mathcal{A}$ and $B \in \mathcal{B}$, $p(A \mid B) \geq 0$ and $p(B \mid B) = 1$. For any countable sequence of disjoint sets $A_i \in \mathcal{A}$, $p(\bigcup_{i\in\mathbb{N}} A_i \mid B) = \sum_{i\in\mathbb{N}} p(A_i \mid B)$—conditional on any fixed element $B$, probability is Countably Additive. Finally, if $B, C, B \cap C \in \mathcal{B}$, then $p(A \cap B \mid C) = p(A \mid B \cap C)p(B \mid C)$. (In the later book he adds one more condition, which I will discuss later.) Although there is no official notion of unconditional probability, if $\Omega \in \mathcal{B}$, then we can use $p(A \mid \Omega)$ as a surrogate for $p(A)$. (The fact that $\mathcal{B}$ may lack $\Omega$ may make this formalism of particular interest for interpretations of probability where some positive amount of information is needed to generate any probabilities, like actual relative frequency, and perhaps logical and evidential probability. See Section 1.2.)

18 In the previous section, '$\mathcal{F}$' was used for the field of all bearers of probability and '$\mathcal{A}$' was used for the sub-field that we are conditioning on. In this section I follow Rényi in using '$\mathcal{A}$' for the field of all bearers of probability, and '$\mathcal{B}$' for the subset that can be conditioned on. I hope that the change in notation is not too confusing—readers should expect still other choices of letters in other sources on this topic.

### 3.2.2 *Simplest Examples*

Rényi gives several basic examples of conditional probability spaces satisfying these axioms. Many of these examples use the notion of a "measure," which is very much like a probability function. A measure is just a Countably Additive function $\mu$ assigning non-negative extended real numbers to elements of a $\sigma$-algebra $\mathcal{A}$ of subsets of some set $\Omega$. To say that the values are "extended real numbers" is just to say that in addition to all the non-negative real numbers, $+\infty$ is also a possible value of the function, with Countable Additivity defined to include this value in the obvious ways (as the sum of any non-convergent series of positive real numbers, or as the sum of any set including $+\infty$). The difference between a measure and a probability function is that for a standard probability function, $p(\Omega) = 1$, while for a measure, $\mu(\Omega)$ can be any non-negative extended real number. A measure is said to be *finite* if $\mu(\Omega)$ is a positive real number, and *$\sigma$-finite* if there is a countable collection of sets $S_i$ for $i \in \mathbb{N}$ with each $\mu(S_i)$ finite and $\Omega = \bigcup_{i\in\mathbb{N}} S_i$.

The most basic example of a Rényi conditional probability space is to let $\mu$ be any finite measure, and let $\mathcal{B}$ be the collection of all elements of $\mathcal{A}$ whose measure is positive. Then define $p(A \mid B) = \mu(A \cap B)/\mu(B)$, and it is straightforward to see that all axioms apply. Of course, this example is of no help to the problems discussed in Section 1.3, because it leaves probabilities conditional on many elements of $\mathcal{A}$ undefined, and in particular on any element whose measure is 0, which are exactly the elements that have unconditional probability 0.

A slightly more general example is to let $\mu$ be any measure at all on $\Omega$, and let $\mathcal{B}$ be the collection of all elements of $\mathcal{A}$ whose measure is positive and finite. Then define $p(A \mid B) = \mu(A \cap B)/\mu(B)$. Interestingly, if $\mu(\Omega) = +\infty$, then this means that there is no notion of unconditional probability—all probability is conditional probability. However, in addition to leaving out probabilities conditional on $\Omega$, this sort of example also still leaves out $p(A \mid B)$ when $\mu(B) = 0$. However, this sort of example is the one that motivated Rényi's development of the theory, and in his later books he adds an axiom that entails that every conditional probability space is of this type, with $\mu$ being $\sigma$-finite. I will come back to the features of this class of examples later.

### 3.2.3 *Popper and Kolmogorov*

In the slightly more general system defined in his earlier paper, he also gives several other interesting examples. Instead of a single measure $\mu$ we can consider a countable *set* of measures $\mu_i$ for $i \in \mathbb{N}$. Then we let $\mathcal{B}$ be the collection of all members of $\mathcal{A}$ such that there is exactly one $\alpha$ with $\mu_\alpha(B) > 0$, and no $\alpha$ such that $\mu_\alpha(B) = +\infty$. If we define

$p(A \mid B) = \mu_\alpha(A \cap B)/\mu_\alpha(B)$ for this unique $\alpha$, then we have another example of a Rényi conditional probability function. By Spohn's result mentioned in Section 3.1, this means that every Countably Additive Popper function is an example of a Rényi conditional probability function (where we leave probability conditional on abnormal sets undefined, rather than saying it is uniformly equal to 1).

Rényi also considers cases in which Disintegrability or Conglomerability might be satisfied. Starting on p. 307 of his (1955), he discusses both what he calls "Cavalieri spaces" and then "regular probability spaces." These are spaces in which $\mathcal{A}$ is the $\sigma$-algebra generated by a random variable $V$, and $\mathcal{B}$ contains all the sets of the form $x < V < y$ as well as the sets of the form $V = x$, and in which the probability function satisfies Conglomerability with respect to the partition in terms of $V = x$. As he notes, his basic definition of a conditional probability space allows for Conglomerability over $\mathcal{A}$ to fail. However, he gives several examples in which it holds, including an instance of the Borel paradox where $\mathcal{B}$ is the set of longitudes and wedges built up from longitudes. This shows a case where he allows for non-trivial probabilities conditional on some events of probability 0. But it leaves conditional probability undefined for *any* event that is not composed of longitudes.

As I discussed in Section 1.4.3, if we consider not just one conditional probability function, but have many, each with its own $\mathcal{B}$, such that every non-empty set is in one of the $\mathcal{B}$, then we can get an adequate notion of conditional probability that responds to the problem of conditioning on events of probability 0 (from Section 1.3) while satisfying Conglomerability. However, $p(A \mid B)$ will then depend on which probability function is being used, which corresponds to the question of which bunch $\mathcal{B}$ of sets is the base of conditioning. Regular Conditional Probability is a special case of Rényi's theory, where $\mathcal{B}$ ranges only over sub-$\sigma$-algebras and Conglomerability is required to hold.

Thus, Rényi's theory is mathematically more general than the theory of Regular Conditional Probability. However, this generality leaves many choices open to us. If the philosophical interest is in preserving a unique notion of conditional probability that doesn't depend on $\mathcal{B}$ at all, then most of this generality is unwanted. Restricting to the case where $\mathcal{B}$ just is the set of all non-empty sets is the subject of Section 3.3.

### 3.2.4 *Infinite Measure*

Despite the interest of these sorts of conditional probability spaces, Rényi's primary interest is in the second example from Section 3.2.2, where the conditional probability is defined from a single measure $\mu$ that is $\sigma$-finite but not finite. This is made clear by the discussion in the first two pages

of his (1955) of the importance of unbounded measures in statistical practice. In his (1970a) he adds an extra axiom to the definition of a conditional probability space, requiring that for any $B, C \in \mathcal{B}$ with $B \subseteq C$, $p(B \mid C) > 0$.[19] And in his (1955), most of his discussion is confined to spaces that satisfy it.

As Theorem 8 in his (1955), and as Theorem 2.2.1 of his (1970a), he proves that for every conditional probability space satisfying this further condition, there is a $\sigma$-finite measure $\mu$ such that $p(A \mid B) = \mu(A \cap B)/\mu(B)$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$, and that this measure is unique up to constant multiple.

The proof is not terribly difficult. Recall that there is a countable sequence $B_i \in \mathcal{B}$, for $i \in \mathbb{N}$ with $\bigcup_{i \in \mathbb{N}} B_i = \Omega$. Without loss of generality, we can assume that $B_i \subseteq B_j$ for any $i \leq j$. (If they don't already satisfy this condition, just replace $B_j$ with the finite union $\bigcup_{i \leq j} B_i$.) Now we can define $\mu(B_1) = 1$, and $\mu(B_n) = 1/p(B_1 \mid B_n)$. Then, for any $A \in \mathcal{A}$, we can define $\mu(A) = \lim_{n \to \infty} \mu(B_n) p(A \mid B_n)$. Verifying that this definition of $\mu$ is well-defined and gives a measure is somewhat tedious, but not terribly difficult. It is substantially easier to verify that any other measure giving the same conditional probability function must be a constant multiple of this one, and that this one is $\sigma$-finite.

By restricting consideration to this sort of probability space, Rényi eliminates all of the non-trivial Popper functions. This is because under this new characterization, whenever $p(A \mid B)$ is defined, $p(B \mid C)$ will be positive whenever it is also defined, unless $C \cap B = \emptyset$. However, Popper's notion of conditional probability was intended to capture cases where $p(B) = 0$ and yet $B$ is normal.

Some philosophers have grouped Popper and Rényi together as giving similar notions of primitive conditional probability. However, Rényi requires Countable Additivity where Popper can't even state it, and Rényi's mature theory rules out all interesting Popper functions, as well as ruling out any resolution to the problem of conditioning on events of probability 0. Although Rényi's theory even more so than Popper's makes conditional probability the basic notion (because $\Omega$ can fail to be in $\mathcal{B}$), it addresses only the motivating problem from Section 1.2 (the conceptual requirement that all probabilities are conditional) and not the one from Section 1.3 (conditioning on events of probability 0).

This mature theory works well for the actual relative frequency interpretation of probability. In fact, one of the standard examples that Rényi considers has exactly this form. Let $\Omega$ be some countable set, let $\mathcal{A}$ be the collection of all subsets of this set, and let $\mu(A)$ be the number of

19 He appears to have this same restriction in mind in his (1970b), though he writes the requirement in a way that is *conditional* on $p(B \mid C) > 0$ rather than requiring it. But that book develops very little of the theory.

elements of $A$. (Since $\Omega$ is countable, we see that $\mu$ is $\sigma$-finite, since $\Omega$ is the union of countably many sets with finitely many elements each.) If we let $\mathcal{B}$ be the set of all non-empty finite subsets of $\Omega$, and define $p(A \mid B) = \mu(A \cap B)/\mu(B)$, then this just is the definition of finite relative frequency.

### 3.2.5 *Improper Priors*

Another more characteristic example lets $\Omega$ be the set $\mathbb{R}^2$ of pairs of real numbers. Let $\mathcal{A}$ be the collection of all Lebesgue measurable subsets of this set, and let $\mu$ be standard Lebesgue measure. Then let $\mathcal{B}$ be the set of all Lebesgue measurable subsets of this set with positive finite measure. The resulting probability measure is uniform conditional on any finite region, and undefined on infinite or null regions.

If we return to the generality of the early theory (so that we allow $\mathcal{B}$ to contain elements whose probability is 0 conditional on large elements of $\mathcal{B}$), we can generalize to a slightly more interesting set $\mathcal{B}$ as follows. Let $R^{x_2,y_2}_{x_1,y_1}$ be the rectangle of points $\{(x,y)\colon x_1 \leq x \leq x_2, y_1 \leq y \leq y_2\}$. Let $\mathcal{B}$ be the set of all such rectangles. When $x_1 < x_2$ and $y_1 < y_2$, we define $p(A \mid R^{x_2,y_2}_{x_1,y_1})$ as before, as the ratio of the standard two-dimensional Lebesgue measure of $A \cap R^{x_2,y_2}_{x_1,y_1}$ to the measure of $R^{x_2,y_2}_{x_1,y_1}$, which is just $(x_2 - x_1)(y_2 - y_1)$. However, when $x_1 = x_2$ or $y_1 = y_2$, the "rectangle" is actually a line segment. In such a case we use the relevant *one*-dimensional Lebesgue measure to define the conditional probability. (This is effectively an example where we have a sequence of three measures—two-dimensional Lebesgue measure $\mu_{x,y}$, one-dimensional Lebesgue measure $\mu_x$ with respect to $x$, and one-dimensional Lebesgue measure $\mu_y$ with respect to $y$.) Again, our probability is uniform conditional on finite rectangles of positive size, but it is also uniform conditional on finite line segments parallel to the $x$ or $y$ axis. But again, there is no unconditional probability, because the space as a whole has infinite measure.

The motivation for this sort of example comes when we generalize it still further. Instead of using Lebesgue measure, we use a measure with a non-uniform density. Then the formulas for calculating conditional probabilities are exactly those given in Section 2.1 for Kolmogorov's Regular Conditional Probabilities, except that some of the integrals might be infinite, and we only officially allow for probabilities conditional on sets where the integrals are finite. In that section, since there was an unconditional probability function, the integrals were always guaranteed to be finite, but here we allow for them to be infinite. When they are infinite, it is standard to say that the conditional probability function arises from an "improper prior," which is not itself a probability function.

This is the foundation of much Bayesian statistical practice. For instance, one might be interested in estimating the distribution of values of $V$ in some population. One might antecedently be sure that, over the relevant population, $V$ is distributed according to a normal distribution with some unknown mean $\mu$ and variance $\sigma^2$. In the absence of information one wants an "uninformative prior," which should be invariant under changes of measuring scale of $V$. (For instance, we might convert feet to meters, or Fahrenheit to Celsius.) It turns out that the only such prior is one where the probability that $x_1 < \mu < x_2$ and $0 < y_1 < \sigma^2 < y_2$ is proportional to $(x_2 - x_1) \log \frac{y_2}{y_1}$. But without antecedent bounds on how large $\mu$ and $\sigma^2$ might be, this gives rise to an improper prior. In particular, since

$$\int_{x_1}^{x_2} \int_{y_1}^{y_2} \frac{1}{y} \, \mathrm{d}y \, \mathrm{d}x = (x_2 - x_1) \log \frac{y_2}{y_1},$$

this means that we can do the calculations with a density given by $d(\mu, \sigma^2) = 1/\sigma^2$.

In this case, in addition to the population mean and variance, there are further random variables given by the observed values of $V$ on samples from the population. We have assumed that each of these samples is taken from the same normal distribution with mean $\mu$ and variance $\sigma^2$. If we represent the density of the normal distribution by $N_{\mu,\sigma^2}(x)$, then our overall density is given by $d(x, \mu, \sigma^2) = N_{\mu,\sigma^2}(x)/\sigma^2$. Interestingly, although this density yields an improper prior, it turns out that conditional on any possible observed value of $x$, the integral over all values of $\mu$ and $\sigma^2$ is finite (because the normal distribution dies off fast enough in each direction). It is a classic result of Bayesian statistics that the posterior distribution of $\mu$ conditional on observed $x$ values is given by Student's $t$-distribution. There are many other cases like this, where a density function over some parameters gives rise to an improper prior, but the natural likelihood function for some observable evidence yields a proper posterior conditional on any possible observation.

Of course, all of this Bayesian analysis only works when it is possible to calculate probabilities by integrating densities. This only works when the conditional distributions satisfy Conglomerability (and thus Countable Additivity) wherever they are defined. Thus, this sort of statistical application requires both Rényi's idea that "unconditional probabilities" can be unbounded, and Kolmogorov's idea that conditional probabilities might be relativized to a partition.

However, the notion of an improper prior is also in some ways closely conceptually related to *failures* of Countable Additivity. This can be seen by looking back at the first example we gave of an improper prior. This was the conditional probability space given by finite counting over a countable set. There is some sense in which this conditional probability

space is aiming to represent a uniform unconditional probability over the countable set, like the de Finetti lottery that (for some) motivates rejection of Countable Additivity. By the technique of improper priors, Rényi is able to represent this distribution in a way that captures much that is important, though it does not give any notion of unconditional probability. Because the total space is $\sigma$-finite, there is a countable sequence of sets $B_i \in \mathcal{B}$ for $i \in \mathbb{N}$ such that $\Omega = \bigcup_{i \in \mathbb{N}} B_i$. We can define a merely Finitely Additive probability function over $\Omega$ by defining $p(A) = \lim_{i \to \infty} p(A \mid B_i)$, though for many sets $A$ this limit is undefined, and in general the limit will depend on the specific choice of the sequence $B_i$.

### 3.3 *De Finetti/Dubins—Full Coherent Conditional Probabilities*

The final theory of Coherent Conditional Probabilities to be considered here takes seriously the motivation in these cases to have well-defined unconditional probabilities while giving up on Countable Additivity. This theory arises from de Finetti (1974) and Dubins (1975, section 3). However, it may be useful for many readers to also consult the expositions of this theory by Seidenfeld (2001), Seidenfeld et al. (2013), or the book length treatment by Coletti and Scozzafava (2002).

The basic background system is the same as that of Kolmogorov and Rényi, but I repeat the definitions here so that readers don't have to flip back. There is a set $\Omega$ of possibilities, and we consider some collection $\mathcal{A}$ of subsets of $\Omega$. If $\mathcal{A}$ contains the empty set, as well as complements and pairwise unions of its members, then $\mathcal{A}$ is said to be an *algebra*. If it also contains unions of any countable set of its elements, then it is said to be a *$\sigma$-algebra*. An algebra $\mathcal{B}$ is said to be a *sub-algebra* of $\mathcal{A}$ iff every member of $\mathcal{B}$ is a member of $\mathcal{A}$, and a *sub-$\sigma$-algebra* of $\mathcal{A}$ if $\mathcal{B}$ is a $\sigma$-algebra.

Unconditional probability for an algebra $\mathcal{A}$ is assumed to be given by a function $p(A)$ defined for $A \in \mathcal{A}$ subject to the three basic principles. $p(\Omega) = 1$, $p(A) \geq 0$ for all $A \in \mathcal{A}$, and $p(A \cup B) = p(A) + p(B)$ when $A$ and $B$ are disjoint members of $\mathcal{A}$. If $\mathcal{B}$ is a sub-algebra of $\mathcal{A}$, then a conditional probability for $(\mathcal{A}, \mathcal{B})$ is a two-place function $p(A \mid B)$ defined for $A \in \mathcal{A}$ and non-empty $B \in \mathcal{B}$ subject to the following constraints. For any $A \in \mathcal{A}$ and non-empty $B \in \mathcal{B}$, $p(A \mid B) \geq 0$ and $p(B \mid B) = 1$. For any $A_1, A_2 \in \mathcal{A}$ and non-empty $B \in \mathcal{B}$, if $A_1 \cap A_2 \cap B$ is empty, then $p(A_1 \mid B) + p(A_2 \mid B) = p(A_1 \cup A_2 \mid B)$. For any $B, C \in \mathcal{B}$ with $B \cap C$ non-empty, and any $A \in \mathcal{A}$, $p(A \cap B \mid C) = p(A \mid B \cap C) p(B \mid C)$.

These axioms are much like Popper's axioms, but formulated in terms of sets rather than sentences of a language. They are much more like Rényi's axioms, but without Countable Additivity (and without the requirement that $\mathcal{A}$ be a $\sigma$-algebra), and with the additional requirement that $p(A \mid \Omega)$ be defined (since $\Omega$ is a member of any algebra $\mathcal{B}$).

One further notion is of great interest here. If $\mathcal{B} = \mathcal{A}$, then the Coherent Conditional Probability is said to be *Full*. The central results in the relevant section of Dubins' paper show that for any probability function on an algebra $\mathcal{A}$ there is a Full Coherent Conditional Probability agreeing with it, and that for any conditional probability function on $(\mathcal{A}, \mathcal{B})$ there is an extension to a Full Coherent Conditional Probability. In fact, he shows that the same is true for any partial function, each of whose finite fragments can be extended to a Full Coherent Conditional Probability function on its finite algebra. In particular, this applies to any Rényi conditional probability function, and even allows us to extend to the case in which $\mathcal{A}$ is the full power set of $\Omega$. Thus, we are able to get what Popper was after—a notion of conditional probability that is defined for every non-empty set.

However, the techniques for proving that these Full Coherent Conditional Probabilities exist are non-constructive. Dubins uses Tychonov's theorem (which is equivalent to the Axiom of Choice), and cites similar results by Krauss (1968) arrived at using non-principal ultrafilters (whose existence is proven using the Axiom of Choice). Similar results extending linear (i.e., finitely additive) functions on subspaces to full spaces often appeal to the Hahn-Banach Theorem, which is also independent of Zermelo-Fraenkel set theory without the Axiom of Choice. Given a Full Coherent Conditional Probability on the surface of a sphere, one can generate the paradoxical Banach-Tarski sets (Pawlikowski, 1991). Thus, we are not usually able to work with these Full Coherent Conditional Probabilities in any explicit way, if we really want them to be defined on *all* subsets of a reasonably-sized probability space. I have argued elsewhere (Easwaran, 2014) that mathematical structures depending on the Axiom of Choice in this way cannot be of epistemic or physical relevance, though they are surely of mathematical interest.

Given the results of Section 1.4.3, Full Coherent Conditional Probabilities fail to satisfy Conglomerability when some Additivity fails. For instance, let $\Omega$ be the set of pairs $(m, n)$ of natural numbers. Let $S_m$ be the set of all pairs whose first coordinate is $m$ and let $T_n$ be the set of all pairs whose second coordinate is $n$. Let $p$ be any probability function such that $p(S_m \mid T_n) = p(T_n \mid S_m) = 0$ for all $m$ and $n$. (We can think of this probability function as describing two independent de Finetti lotteries.) Let $E$ be the event that $m > n$. Then we can see that for any $m$, $p(E \mid S_m) = 0$ (since, conditional on $S_m$, only finitely many values of $n$ will satisfy $E$), but for any $n$, $p(E \mid T_n) = 1$ (since, conditional on $T_n$, only finitely many values of $m$ will *fail* to satisfy $E$). Since the $S_m$ and the $T_n$ are both partitions, *any* value of $p(E)$ will fail to satisfy Conglomerability in at least one of these partitions. This sort of failure of Conglomerability is inevitable if one allows failures of Countable Additivity and requires that sets like $E$ nevertheless have both unconditional and conditional probabilities.

However, these Finitely Additive Full Coherent Conditional Probabilities have the advantage of existing even for algebras that are not countably generated, avoiding the problems for Regular Conditional Probabilities mentioned in Section 2.3.1. They also always satisfy $p(A \mid A) = 1$, even in the bad algebras where Countably Additive conditional probabilities are forced to allow for $p(A \mid A) = 0$, as mentioned in Section 2.3.2 (Seidenfeld et al., 2001). In particular, in addition to the case where one adds a non-measurable set to the collection of Borel sets, one might also consider the algebra of "tail events," defined as follows.

Let $\Omega$ be the set of all countable sequences $(a_0, a_1, a_2, \dots)$ of 0s and 1s (which can be taken to represent the set of all countable sequence of coin flips). Let $\mathcal{A}$ be the $\sigma$-algebra generated by the sets of the form

$$A_i = \{(a_0, a_1, a_2, \dots) : a_i = 1\}.$$

Say that an element $A \in \mathcal{A}$ is a "tail event" if, for any element of $A$, changing any finitely many places in the sequence results in another element of $A$. (The tail events are exactly those that depend only on the long-run behavior of the sequence and not on any short-term behavior.) Let $\mathcal{B}$ be the set of all tail events. It is clear that $\mathcal{B}$ is a sub-$\sigma$-algebra of $\mathcal{A}$.

A classic result of Kolmogorov shows that if the unconditional probability is that on which each event $A_i$ ("the $i$-th flip results in heads") is independent with probability $1/2$, then every event in $\mathcal{B}$ has probability 1 or 0. A further generalization by Hewitt and Savage shows that if the unconditional probability is *any* "exchangeable" probability (in the sense of de Finetti), then the events in $\mathcal{B}$ all have probability 1 or 0. As a consequence of these results, and a theorem about algebras in which all probabilities are 1 and 0, it turns out that any element $B \in \mathcal{B}$ whose unconditional probability is 0 must also have $p(B \mid B) = 0$, if conditional probability is Countably Additive. (See Blackwell and Dubins, 1975, or Seidenfeld et al., 2001. This is possible because the algebra of tail events is not countably generated.) But if conditional probability is allowed to be merely Finitely Additive, then we can have $p(B \mid B) = 1$ for these tail events. Dubins and Heath (1983) show how to construct such a Full Coherent Conditional Probability. However, this construction assumes a particular merely Finitely Additive probability distribution over all subsets of the natural numbers, and thus indirectly appeals to the Hahn-Banach Theorem, and thus the Axiom of Choice.

Since these functions are defined on the full power set, there is a sense in which we no longer need to limit ourselves to an algebra $\mathcal{A}$ of "measurable" sets. Even the unmeasurable sets are assigned some probability. We aren't able to pin down precisely what the probability is of any such set, but since the non-measurable sets themselves are only proved to exist by non-constructive means using the Axiom of Choice, this may not be

such a problem. The Banach-Tarski Paradox shows that if $\Omega$ contains 3-dimensional (or higher) Euclidean space, then any such Finitely Additive probability function must fail to be invariant under rotations and translations. But again, the sets under which these invariances must fail are only proven to exist by means of the Axiom of Choice.[20]

Thus, provided that one is not worried about working with non-constructive methods, Full Coherent Conditional Probabilities can be of interest when dealing with algebras that aren't countably generated.

## 4 CONCLUSION

There are two main families of arguments that conditional probability should be taken as the basic notion of probability, or at least as equally fundamental to unconditional probability. One set of arguments (Section 1.2) is based on conceptual grounds, but apart from the interpretation of probability as actual frequency, it doesn't appear to be decisive. For logical, evidential, and perhaps even subjective probabilities (if we follow Levi), we may be able to argue that nearly all probabilities are conditional. But if we can make sense of conditioning on a tautology, then again the argument is not decisive. Instead, this argument points out that many probability functions depend on some background condition that is of a different type than the events that have probabilities.

The other set of arguments (Section 1.3) is based on mathematical grounds. Depending on how we treat vague or indeterminate probabilities (if there even are any), these problem cases may not motivate anything beyond a supervaluational treatment. I believe that supposed cases of conditioning on an event with undefined unconditional probability are either cases of maximally vague probability, cases where the "event" is actually part of the background for a probability function rather than a condition, or are cases where the conditional probability also does not exist.

Instead, it is cases of probability 0 (and particularly those where the 0 arises from an infinite partition) that motivate a reconsideration of the mathematics of probability theory the most strongly. To deny that these cases exist is to assume something much stronger than Finite Additivity or Countable Additivity—it is either to assume Full Additivity for all cardinalities (and thus discrete probability, distributed only over countably many possibilities) or else the even stronger assumption that there are only

20 If we replace the Axiom of Choice by the Axiom of Determinacy, then we lose the Hahn-Banach theorem and the other means by which these Finitely Additive functions were proven to exist, but Lebesgue measure turns out to already be defined—and Countably Additive!—over all subsets of Euclidean space. See Bingham (2010, Section 8).

finitely many possibilities. This seems to go against the meaningfulness of scientific vocabulary discussing numerical parameters in the world.

I have discussed four different mathematical theories for conditioning on events of probability 0. Regular Conditional Probabilities may allow us to say that unconditional probability is prior to conditional probability, while Popper's theory, Full Coherent Conditional Probabilities, and the most general version of Rényi's theory require conditional probability to be prior.

Popper's theory is the one most familiar to philosophers. This theory has the advantage of deriving the relations of deductive propositional logic as special consequences of the probability axioms, so it may be particularly well-suited to the logical interpretation of probability. But because the bearers of probability are sentences in a language rather than sets of possibilities, it can't even express the circumstances that give rise to the problem of probability 0, much less say anything useful about them. In any case, it is effectively an instance of the more general Dubins/de Finetti Full Coherent Conditional Probability.

Rényi's theory is the most general, having versions of the others as special cases (though some require dropping Countable Additivity). Rényi's theory is particularly well-suited to the account of probability as actual relative frequency, and may well be particularly suited to interpretations of probability where not every proposition can be conditionalized upon, particularly if the tautology is one of these propositions (so that there is no such thing as unconditional probability). It also has advantages for certain calculations in a Bayesian statistical framework that depend on the use of "improper priors."

The Dubins/de Finetti Full Coherent Conditional Probabilities, and the Regular Conditional Probabilities descending from Kolmogorov, have competing mathematical virtues. Regular Conditional Probabilities can satisfy Conglomerability in each partition, as well as Countable Additivity, which appears to be the most well-motivated level of Additivity. However, Full Coherent Conditional Probabilities allow each conditional probability to be defined in a unified and coherent way (rather than one depending on a partition in addition to a conditioning event). I suggested in Section 1.1 that actual applications of conditional probability always come with some clear sense of the partition that is relevant, so this is not a cost of the theory of Regular Conditional Probabilities. Full Coherent Conditional Probabilities avoid some problem cases that arise on badly behaved algebras. However, I claim these algebras are too complicated for a finite human mind to grasp, so I think they don't arise in epistemic application in any case. Regardless, Full Coherent Conditional Probabilities are themselves so complex that they can't be proved to exist without some version of the

Axiom of Choice, while Regular Conditional Probabilities can be given constructively when the unconditional probability is defined by a density.

The Regular Conditional Probabilities associated with an unconditional probability are generally only unique up to measure 0. Perhaps there could be some constraint like continuity, or computability, that might uniquely define conditional probabilities for each partition given unconditional probabilities on countably generated algebras. If this is right, then we may be able to say that unconditional probability is basic after all, and conditional probability defined in terms of it. But otherwise, there must be some sense in which conditional probability is either primitive, or at least equally fundamental to unconditional probability. Or else we can follow Myrvold (2015) and allow that we can't always get what we want in a theory of conditional probability.

Rényi's fully general theory must be used in a few situations where conditional probability is required to be independent of unconditional probability (namely, for actual relative frequency in infinite worlds, and in applications requiring "improper priors"). For other applications, the situation is summarized in Table 1 (page 193).

## REFERENCES

Arntzenius, F., Elga, A., & Hawthorne, J. (2004). Bayesianism, infinite decisions, and binding. *Mind*, *113*(450), 251–283.

Bacon, A. (2015). Stalnaker's thesis in context. *The Review of Symbolic Logic*, *8*(1), 131–163.

Bartha, P. (2004). Countable additivity and the de Finetti lottery. *British Journal for the Philosophy of Science*, *55*, 301–321.

Bertrand, J. (1889). *Calcul des probabilités*. Gauthier-Villars.

Billingsley, P. (1995). *Probability and measure*. Wiley.

Bingham, N. H. (2010). Finite additivity versus countable additivity: De Finetti and Savage. *Electronic Journal of the History of Probability and Statistics*, *6*(1), 1–33.

Blackwell, D. (1956). On a class of probability spaces. In *Proceedings of the third Berkeley symposium on mathematics, statistics, and probability* (Vol. 2, pp. 1–6).

Blackwell, D. & Dubins, L. (1975). On existence and non-existence of proper, regular, conditional distributions. *The Annals of Probability*, *3*(5), 741–752.

Brandenburger, A. (2007). The power of paradox: Some recent developments in interactive epistemology. *International Journal of Game Theory*, *35*, 465–492.

Briggs, R. A. (2009). Distorted reflection. *Philosophical Review*, *118*(1), 59–85.

| | Finite/Discrete Probability | Regular Conditional Probabilities | Full Coherent Conditional Probabilities |
|---|---|---|---|
| Pros | ratio definition<br>Full Additivity<br>mathematically simple | Conglomerability<br>Countable Additivity<br>mathematically standard | two-place $p(B \mid A)$<br>defined for all non-empty $A$<br>exist for all algebras |
| Cons | only countable spaces<br>no continuous random variables | three-place $p(B \mid A, \mathbf{A})$<br>problems in bad algebras<br>"almost" uniqueness | only Finitely Additive<br>require Axiom of Choice<br>non-Conglomerability |
| Priority | unconditional probability | uncond. probability (almost) | conditional probability |

Table 1: Summary of views and their features

Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press.

Chang, J. & Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, *51*(3), 287–317.

Coletti, G. & Scozzafava, R. (2002). *Probabilistic logic in a coherent setting*. Kluwer.

de Finetti, B. (1974). *Theory of probability*. Wiley.

Dubins, L. (1975). Finitely additive conditional probabilities, conglomerability, and disintegrations. *The Annals of Probability*, *3*(1), 89–99.

Dubins, L. & Heath, D. (1983). With respect to tail sigma fields, standard measures possess measurable disintegrations. *Proceedings of the American Mathematical Society*, *88*(3), 416–418.

Easwaran, K. (2011a). Bayesianism I: Introduction and arguments in favor. *Philosophy Compass*, *6*(5), 312–320.

Easwaran, K. (2011b). Bayesianism II: Applications and criticisms. *Philosophy Compass*, *6*(5), 321–332.

Easwaran, K. (2013a). Expected accuracy supports conditionalization — and conglomerability and reflection. *Philosophy of Science*, *80*(1), 119–142.

Easwaran, K. (2013b). Why countable additivity? *Thought*, *1*(4), 53–61.

Easwaran, K. (2014). Regularity and hyperreal credences. *The Philosophical Review*, *123*(1).

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Eells, E. (1985). Problems of old evidence. *Pacific Philosophical Quarterly*, *66*, 283–302.

Erdős, P. (1947). Some remarks on the theory of graphs. *Bulletin of the American Mathematical Society*, *53*, 292–294.

Faden, A. M. (1985). The existence of regular conditional probabilities: Necessary and sufficient conditions. *The Annals of Probability*, *13*(1), 288–298.

Field, H. (1977). Logic, meaning, and conceptual role. *The Journal of Philosophy*, *74*(7), 379–409.

Fitelson, B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66*(3), S362–S378.

Gallow, J. D. (2016). *Diachronic Dutch books and evidential import*. ms.

Glymour, C. (1980). *Theory and evidence*. Princeton University Press.

Greaves, H. & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*(459), 607–632.

Gyenis, Z., Hofer-Szabó, G., & Rédei, M. (2016). *Conditioning using conditional expectations: The Borel-Kolmogorov paradox*. ms.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, *137*, 273–323.

Hájek, A. (2007). Interpretations of probability. *Stanford Encyclopedia of Philosophy*.

Hájek, A. & Fitelson, B. (2017). Declarations of independence. *Synthese*, *194*(10), 3979–3995.

Halpern, J. (2009). Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*.

Hill, B. M. & Lane, D. (1985). Conglomerability and countable additivity. *Sankhyā: The Indian Journal of Statistics, Series A*, *47*(3), 366–379.

Hitchcock, C. (2010). Probabilistic causation. *Stanford Encyclopedia of Philosophy*.

Hoffmann-Jørgensen, J. (1971). Existence of conditional probabilities. *Mathematica Scandinavica*, 257–264.

Horowitz, S. & Dogramaci, S. (2016). Uniqueness: A new argument. *Philosophical Issues*, *26*.

Hosiasson-Lindenbaum, J. (1940). On confirmation. *Journal of Symbolic Logic*, *5*(4), 133–148.

Howson, C. (2008). De Finetti, countable additivity, consistency and coherence. *The British Journal for the Philosophy of Science*, *59*(1), 1–23.

Humphreys, P. (1985). Why propensities cannot be probabilities. *The Philosophical Review*, *94*(4), 557–570.

Humphreys, P. (2004). Some considerations on conditional chances. *British Journal for the Philosophy of Science*, *55*, 667–680.

Jeffrey, R. (1965). *The logic of decision*. McGraw-Hill.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*(4), 575–603.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1986). Statistical implications of finitely additive probability. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti*. North-Holland.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, *91*(435), 1228–1235.

Kemeny, J. (1955). Fair bets and inductive probabilities. *The Journal of Symbolic Logic*, *20*(3), 263–273.

Keynes, J. M. (1921). *A treatise on probability*. Macmillan and co.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. Chelsea.

Kopec, M. & Titelbaum, M. (2016). The uniqueness thesis. *Philosophy Compass*.

Krauss, P. H. (1968). Representation of conditional probability measures on Boolean algebras. *Acta Mathematica Hungarica*, *19*(3-4), 229–241.

Levi, I. (1980). *The enterprise of knowledge*. MIT Press.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2). University of California Press.

Maher, P. (2006). A conception of inductive logic. *Philosophy of Science*, *73*, 518–523.

Mahtani, A. (2019). Imprecise probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

Mayo, D. & Cox, D. (2006). Frequentist statistics as a theory of inductive inference. *IMS Lecture Notes–Monograph Series 2nd Lehmann Symposium — Optimality*, *49*, 77–97.

McGee, V. (1994). Learning the impossible. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals*. Cambridge University Press.

Meek, C. & Glymour, C. (1994). Conditioning and intervening. *British Journal for the Philosophy of Science*, *45*, 1001–1021.

Myrvold, W. (2015). You can't always get what you want: Some considerations regarding conditional probabilities. *Erkenntnis*, *80*, 573–603.

Pachl, J. K. (1978). Disintegration and compact measures. *Mathematica Scandinavica*, 157–168.

Pawlikowski. (1991). The Hahn-Banach theorem implies the Banach-Tarski paradox. *Fundamenta Mathematicae*, *138*(1), 21–22.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.

Popper, K. (1955). Two autonomous axiom systems for the calculus of probabilities. *The British Journal for the Philosophy of Science*, *6*(21), 51–57.

Popper, K. (1959a). *The logic of scientific discovery*. Harper & Row.

Popper, K. (1959b). The propensity interpretation of probability. *The British Journal for the Philosophy of Science*, 25–42.

Ramsey, F. P. (1930). On a problem of formal logic. *Proceedings of the London Mathematical Society*, *30*(1), 264–286.

Rényi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Hungarica*, 285–335.

Rényi, A. (1970a). *Foundations of probability*. Holden-Day.

Rényi, A. (1970b). *Probability theory*. North-Holland.

Rescorla, M. (2015). Some epistemological ramifications of the Borel-Kolmogorov paradox. *Synthese*, *192*, 735–767.

Rescorla, M. (2018). A Dutch book theorem and converse Dutch book theorem for Kolmogorov conditionalization. *The Review of Symbolic Logic*, *11*(4), 705–735.

Roeper, P. & LeBlanc, H. (1999). *Probability theory and probability logic*. University of Toronto.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Chapman and Hall.

Savage, L. J. (1954). *The foundations of statistics*. Dover.

Schervish, M. J., Seidenfeld, T., & Kadane, J. B. (1984). The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *66*, 205–226.

Schoenfield, M. (2016). *Conditionalization does not (in general) maximize expected accuracy*. ms.

Seidenfeld, T. (2001). Remarks on the theory of conditional probability: Some issues of finite versus countable additivity. In V. Hendricks (Ed.), *Probability theory: Philosophy, recent history and relations to science* (pp. 167–178). Kluwer.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2001). Improper regular conditional distributions. *The Annals of Probability*, *29*(4), 1612–1624.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2013). Two theories of conditional probability and non-conglomerability. In *8th international symposium on imprecise probability: Theories and applications, compiègne, france*.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2014). *Non-conglomerability for countably additive measures that are not κ-additive*. Carnegie Mellon University.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search*. MIT Press.

Spohn, W. (1986). The representation of Popper measures. *Topoi*, *5*, 69–74.

Titelbaum, M. G. (2019). Precise credences. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

van Fraassen, B. (1976). Representation of conditional probabilities. *Journal of Philosophical Logic*, *5*(3), 417–430.

van Fraassen, B. (1984). Belief and the will. *The Journal of Philosophy*, *81*(5), 235–256.

von Neumann, J. & Morgenstern, O. (1947). *Theory of games and economic behavior, second edition*. Princeton University Press.

Wenmackers, S. (2019). Infinitesimal probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

Williamson, T. (2002). *Knowledge and its limits*. Oxford.

Yu, X. (1990). Radon-Nikodym theorem is equivalent to arithmetical comprehension. In W. Sieg (Ed.), *Logic and computation, proceedings of*

*a workshop held at Carnegie Mellon University, June 30-July 2, 1987* (pp. 289–297). American Mathematical Society.