COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2016) 1–33

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12444

What Are the "True" Statistics of the Environment?

Jacob Feldman

Department of Psychology, Center for Cognitive Science, Rutgers University

Received 21 October 2015; received in revised form 19 June 2016; accepted 1 August 2016

Abstract

A widespread assumption in the contemporary discussion of probabilistic models of cognition, often attributed to the Bayesian program, is that inference is optimal when the observer's priors match the *true* priors in the world—the actual "statistics of the environment." But in fact the idea of a "true" prior plays no role in traditional Bayesian philosophy, which regards probability as a quantification of belief, not an objective characteristic of the world. In this paper I discuss the significance of the traditional Bayesian epistemic view of probability and its mismatch with the more objectivist assumptions about probability that are widely held in contemporary cognitive science. I then introduce a novel mathematical framework, the *observer lattice*, that aims to clarify this issue while avoiding philosophically tendentious assumptions. The mathematical argument shows that even if we assume that "ground truth" probabilities actually *do* exist, there is no objective way to tell what they are. Different observers, conditioning on different information, will inevitably have different probability estimates, and there is no general procedure to determine which one is right. The argument sheds light on the use of probabilistic models in cognitive science, and in particular on what exactly it means for the mind to be "tuned" to its environment.

Keywords: Bayesian inference; Probability; Subjectivism; Epistemology

1. The statistics of the natural world and the Conventional Wisdom

For cognitive scientists interested in probabilistic models of perception and cognition, the last decade has seen the emergence of a widespread viewpoint, which I will refer to as the Conventional Wisdom (CW). The CW includes the following premises:

(CW1) each event class X in the environment occurs with some true probability $p(X)$;
(CW2) an agent can achieve optimal performance when its prior on event X is $p(X)$;

Correspondence should be sent to Jacob Feldman, Department of Psychology, Center for Cognitive Science, 152 Frelinghuysen Rd., Piscataway, NJ 08854. Rutgers University. E-mail: jacob@ruccs.rutgers.edu

(CW3) natural selection applies adaptive pressure on organisms toward the true prior $p(X)$.

In this paper I will argue that each of these points CW1–CW3 is incorrect, or at least substantially misguided. The problems begin with CW1, which asserts the independent, objective existence of statistical regularities in the environment, meaning "true" probabilities of events or properties. CW2 and CW2 exploit this assumption by positing that knowledge of the true probabilities $p(X)$ will lead to more accurate inferences (CW2) and thus, in the long run, more adaptive behavior (CW3). In what follows, I will argue that CW1 is wrong—specifically, that the idea of "true probability" is essentially meaningless—and that CW2 and CW3 are wrong as well, in part because they depend on CW1, and also for several independent reasons.

To many readers the CW may sound essentially like a statement of the Bayesian program, which is often described as the use of statistical regularities to guide inference. But in order to help keep the sides clear in what follows, it is important to understand that my criticisms of the CW reflect a *Bayesian* perspective—that is, the problems with the CW arise because it actually conflicts with Bayesian views of probability. What follows is, in many ways, simply a defence of traditional Bayesian philosophy of probability. Before we begin, it is essential to correct the somewhat scrambled perception of which side is which that has become common in cognitive science (see Feldman, 2013). In fact, as I will argue below, the CW can be safely set aside without giving up what so many researchers find attractive about the Bayesian program.

Indeed, traditional Bayesians won't recognize their own views at all in the CW. To traditional Bayesians, probabilities are *degrees of belief*—that is, they are quantifications of epistemic or mental states and thus do not have objectively correct values in the world. Degrees of belief change when one's state of knowledge changes, because they are characteristics of observers rather than characteristics of the world. By contrast, much of traditional statistics assumes a *frequentist* view of probability, in which probability is the limiting value of relative frequency in a long sequence of repetitions of a random experiment.[1] For example, when flipping a fair coin, frequentists view the statement "the probability of heads is .5," as a statement about how often heads comes up in an infinitely long sequence of imaginary coin flips run under identical conditions. Epistemicists (usually called *subjectivists*, but see below about problems with this term) view it as a statement about the state of knowledge of an observer, namely in this case that he or she has no idea whether heads or tails will occur on the next flip. As Pearl (2000) put it, "The wheel of fortune is turned, not by the wisdom of God, but by the ignorance of man."[2] To frequentists, the epistemic definition is uncomfortably subjective. To epistemicists, the frequentist definition is based on an imaginary thought experiment that can ever be carried out.

The distinction between epistemic and frequentist views is sometimes dismissed as a semantic issue with few real consequences. But in the 19th century it became a central point of contention, because (as both sides recognized) the application of Bayes' rule to inference—the estimation of the state of nature based on observable data—almost always

requires adopting the epistemic view. To frequentists, probability can only be assigned to events that are intrinsically repeatable, like coin flips or other "random experiments," which can be run many times yielding different (random) results. But statements that have definite though unknown truth values, like scientific hypotheses, do not have "relative frequencies"—and thus to a frequentist do not have probabilities. This rules out assigning probabilities to scientific hypotheses, and for this reason frequentists such as Fisher (1925) insisted that it was meaningless to use Bayes' rule to determine the posterior probability that a particular hypothesis was true. (This restriction spurred the development of alternative methods of theory evaluation, such as maximum likelihood and null hypothesis significance testing.[3]) But to an epistemicist, probabilities can be assigned to any hypothesis whatsoever, simply based on a consideration of the observer's state of knowledge.

Thus, if one wants to use Bayes' rule to estimate the true state of nature that best explains observable facts—called *inverse probability*, originally a term of derision—one has little choice but to adopt an epistemic view in which the probability assigned to various possible states of nature represents our degrees of belief in them rather than their relative frequency. The relative frequency of rain tomorrow is meaningless, because tomorrow is only going to happen once. But a Bayesian is happy to assign a probability indicating the degree of belief in rain tomorrow based on our current state of knowledge. Of course this is exactly what the weather forecaster is doing when she pronounces "a 30% chance" of rain tomorrow (see Gigerenzer, Hertwig, van den Broek, Fasolo, & Katsikopoulos, 2005), and the same can be said for the probability of any definite, nonrepeatable property of the environment. In this sense, the epistemic view is essential to the entire Bayesian program and cannot be lightly set aside.

As a result, essentially all traditional Bayesians, and most contemporary Bayesians in cognitive science[4] adopt, and at times argue forcefully for, an epistemic attitude toward probability (see Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Oaksford & Chater, 2009, for overviews). In traditional Bayesian practice, one selects priors by seeking distributions that express the observer's state of knowledge (and ignorance) in as accurate or neutral way as possible. For example, one might adopt a Gaussian prior $p(X) \sim N(\mu, \sigma^2)$ on the value of an unknown parameter x, because it is the maximum-entropy (and thus maximally neutral) distribution with a given mean and variance (see Jaynes, 1982; Robert, 2007) or similar arguments (see Bernardo & Smith, 1994; Box & Tiao, 1973; Lee, 2004). That is, the Gaussian prior simply gives mathematical form to the observer's vague belief that the unknown parameter probably lies somewhere around μ with some uncertainty σ. It emphatically does *not* mean that the observer thinks the parameter is "actually" Gaussian distributed in the world—and in fact in the epistemic tradition such a claim is not really meaningful. Similar arguments apply to the medley of other standard prior distributions that populate Bayesian textbooks. These are not chosen just for mathematical convenience (as sometimes charged), but because by various arguments they effectively represent particular states of knowledge.

Indeed, such techniques work very well in practice, often giving better results than priors based on tabulations of "environmental statistics," which can be faulty or misleading in

various ways (see Feldman, 2013). Jaynes (2003) discusses this phenomenon in depth, referring to it as the "ubiquitous success" of Gaussian priors. This success is puzzling if you assume that the environmental frequency distribution is the objectively correct prior, but less so once one understands that from a Bayesian viewpoint there is actually no such thing, and that the goal of the prior is simply to express your knowledge prior to consideration of the data.

This perspective on probability is usually called *subjectivist*, but the terminology is a bit ambiguous because Bayesians themselves are sometimes divided into two camps, somewhat confusingly called *subjectivist* and *objectivist*. But both subjectivist and objectivist Bayesians traditionally hold an epistemic view of probability in which probability is a degree of belief, not a relative frequency, and not a characteristic of the outside world. De Finetti (1970/1974), an influential subjectivist whose position is sometimes called *personalistic*, denounced "objective" probabilities as a delusion akin to phlogiston, fairies, and witches. But even Jaynes, an ardent objectivist, viewed probability epistemically; in his influential (2003) book he repeatedly disparages the idea that probabilities are characteristics of the outside world as the "mind projection fallacy." As he put it, "probabilities change when we change our state of knowledge; frequencies do not."[5] For these reasons I will generally avoid the term *subjectivist*, which without clarification can either refer to personalistic Bayesians or to anyone who holds an epistemic view of probability, which includes nearly all traditional Bayesians, both subjectivists and objectivists alike.[6]

As a consequence of all this, traditional Bayesians generally do not believe in the objective reality of probabilities, meaning they do not think of them as having "true" values in the outside world. "The notion of a 'true prior distribution' is not part of the Bayesian vernacular" (Samaniego, 2010). This may come as a surprise to many contemporary readers in cognitive science, who are accustomed to the CW in which probabilistic inference yields valid results only if probabilistic assumptions match the "true" probabilities at work in the environment. Many contemporary authors routinely refer to environmental probabilities (CW1) and seem to assume that probabilistic inference is most effective when priors match them (CW2). In the elegant—but somewhat frequentistic—metaphor of Burge, Fowlkes and Banks (2010), adopting a statistical model that matches environmental statistics, "closes the loop" between mind and world: "A key idea in the probabilistic characterization of perception is that perceptual systems have internalized the statistical properties of their sensors and the natural environment, and that the systems use those properties efficiently to make optimal perceptual estimates."[7] Such rhetoric is very common (see more examples below), and usually the conflict with the traditional Bayesian stance goes completely unremarked.

This attitude is perhaps most salient in critics of the Bayesian program, who often fault Bayesian models for failing to adopt "realistic" priors, or as Marcus and Davis (2013a) put it, "real-world frequencies." As they comment: "[T]here are often multiple Bayesian models that offer differing predictions because they are based on differing assumptions; at most one of them can be optimal." Only one can be optimal, that is, because only one is "correct" in the environment. Similarly, in a critical assessment of the state of Bayesian modeling in neuroscience, Colombo and Seriès (2012) remark that "the criterion for

choosing the prior should include some 'ecological' consideration since neural processing is influenced by the statistical properties of the environment." Jones and Love (2011) similar disparage many Bayesian models for assuming priors *not* motivated by the true statistics of the environment, apparently presuming that is their only legitimate source ("[T]he prior can be a strong point of the model if it is derived from empirical statistics of real environments"). To be clear, there is no argument on the Bayesian side that beliefs should *not* be influenced by the environment—indeed, that is the entire point of Bayesian inference, to update beliefs in light of data via Bayes' rule. But these authors apparently assume that the environment defines a unique "correct" prior, imposing a frequentist criterion that is incongruent with the Bayesian framework.

The presumption of objective environmental probabilities is by no means limited to Bayesian detractors; it occasionally enters the rhetoric of Bayesians as well, sometimes even juxtaposed with routine recitations of the "degree of belief" definition of probability, and usually without recognition of the contradiction. This mixture of views is most common in the tradition of "natural image statistics," where the goal is (among other things) to set priors by tabulating frequencies in natural images or representative image databases. For example, Burge et al. (2010), in a study of figure/ground estimation, question whether "the estimated prior actually matches the distributions of speeds encountered in natural viewing," apparently assuming the environment specifies a unique distribution, and that it is the proper source for the prior. Purves (2010) asserts that (in the context of vision) "the needed priors are the frequency of occurrence in the world of surface reflectance values, illuminants, distances, object sizes, and so on." Geisler and Diehl (2003), in proposing a Bayesian framework in which to understand perceptual evolution, explicitly identify the prior with the actual frequency distribution of features in the environment: "In the Bayesian framework, natural scene statistics as well as known physical properties of the environment [. . .] are represented by prior probability and likelihood distributions" (see also Geisler & Diehl, 2002). The assumption is that the environment itself defines a unique correct prior, if only we knew what it was. This conflation of prior with objective frequencies in the environment represents a fairly radical departure from the original idea of inverse probability, implying for example that each environment (however "environments" might even be individuated) defines a *unique* rational observer. This is dangerously close to tautology: The maximum-fitness observer *conditioned on an environment defined by its own prior* has maximum fitness. And indeed Geisler and Diehl's conclusions have been sharply questioned by others unwilling to commit to this variety of objectivism (Hoffman & Singh, 2012; Mark, Marion, & Hoffman, 2010).

But the presumption that the environment defines objectively correct probabilities extends far beyond Bayesian advocates and detractors, permeating thinking about probability in the broader community of cognitive science. Several generations of psychologists have learned the idea of probability from introductory statistics classes, which until very recently relied exclusively on the frequentist notion of probability advocated by the inventors of null hypothesis significance testing (chiefly Fisher and Neyman and Pearson; see Gigerenzer & Regier, 1996). As mentioned above, in the frequentist orthodoxy that held sway in psychological statistics since the 1930s, "probability" was reserved for

quantities calculated from repeated samples from a population (i.e. sampling distributions), and the resulting distributions were thought of as characteristics of that population. This conception implies that probability is an objective characteristic of external events, and psychologists absorbed this assumption along with their *t*-tests and ANOVAS.

More specifically, many cognitive psychologists first encountered the notion of a Bayesian prior in the context of the "base rate fallacy," the tendency for decision makers to ignore the known underlying frequency of particular events. In Kahneman and Tversky's original papers on the subject (e.g., Tversky & Kahneman, 1974, 1982), as well as in countless textbooks and even popular media,[8] the terms *base rate* and *prior probability* are explicitly treated as synonymous. The general presumption is that the frequency in the population defines an objectively correct prior that ought to be adopted by any rational observer. Later, this notion of prior was popularized in evolutionary psychology (e.g., Cosmides & Tooby, 1995), in which our minds are understood to have adapted to an early environment in which certain priors objectively held sway.

But as is occasionally pointed out (e.g., Gigerenzer, 1994), but does not seem to be widely understood, in a Bayesian framework, the prior is *not* actually the same as the base rate. Imagine, as in Kahneman and Tversky's example about a car accident involving a taxicab, 15% of the city's taxis are from the blue company and 85% are from the green (Tversky & Kahneman, 1982), and that an observer induces a prior over taxi color by observing taxis drawn randomly (independently and identically distributed, or i.i.d.) from the city's taxi population (i.e., the observer develops his or her prior by observing taxis from this city, before encountering the test problem in Tversky and Kahneman's experiment). There are a number of ways of modeling this situation, but in most typical Bayesian analyses an observer would end up with a prior that is a beta distribution over the underlying green-taxi probability. For sufficiently large N, this distribution is approximately normal with mean around $p = .15$ and standard deviation that is a decreasing function of N, meaning that as more taxis are observed the prior gets more sharply peaked around $p = .15$.[9] For example, after observing 100 taxis, the prior would have a mean of about $p = .157$ with a standard deviation about .036. Such an observer would believe that the proportion of green taxis in the population was *about* 15%, not that it was *exactly* 15%—specifically, that it was about 68% likely to be between 12.1% and 19.3%, and 95% likely to be between 8.5% and 22.9%. A rational observer's prior on taxi color in this city might, after sufficient data, converge to approximately the base rate—but only after considerable experience observing taxis, and even then only approximately. Only after an *infinite* number of observations, and assuming perfectly invariant sampling conditions, would it converge exactly to the base rate. And if the observations are not i.i.d. as in this example, the prior and the base rate can be radically different (see examples below). But the broader point is that the prior is not *definitionally equivalent* to the base rate.

Elsewhere Gigerenzer (1991) has famously argued that the base rate fallacy can be reduced or eliminated if probabilities are presented more naturalistically including both numerator and denominator (e.g., "15 out of 100 taxis" rather than "15% of taxis"). However the point here goes beyond a difference in presentation format: The Bayesian prior for a rational observer of such a world is simply not the same thing as the base rate, even if the base rate defines the generative conditions from which evidence about the prior is drawn.

The almost universal conflation of the prior and base rate reflects a pervasive and usually unquestioned frequentist assumption that the prior is set objectively by the environment.[10]

2. Separating frequency from probability

For readers steeped in frequentist assumptions, it may take a few examples to fully appreciate how different *probability* and *frequency* can be when probability is viewed epistemically. Imagine that an observer encounters a stream of N instances of the binary variable $\mathbf{X}$:

$$\mathbf{X}: \overbrace{1\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 1}^{\text{trials } 1 \ldots N} \mid \overbrace{?}^{N+1}$$

The question is: What is the probability $p(X)$ (that is, the probability that $\mathbf{X}$ will take on value 1) on the $N + 1$-th trial? Data like this are often treated as samples from what statisticians call a *Bernoulli source*, meaning draws from a binary random variable (a "coin"). If so, the mean $\bar{X}$ tends to converge on the "true" value of $p(X)$ as N increases. A coin has come up heads on 62% of flips so far, its "true" heads probability is probably about 0.62. But as suggested in the taxi example above, this does not mean that the probability of heads on the next flip is necessarily .62, for a number of reasons. For one thing, that inference depends on our assumption (which we did make in the taxicab example) that the source is i.i.d., like a real coin. But this is only an assumption, and in some situations it may be a poor one. If, for example, the data were alternating,

$$\mathbf{X}: 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0 \mid ? \tag{1}$$

one might form a strong expectation (high probability) that the next datum will be a 1 (rather than $p(X = 1) \approx 0.5$ as suggested by the Bernoulli model). Or the data might spell out the first six words of the Gettysburg Address ("Four score and seven years ago") in Morse code (1 = dash, 0 = dot),

$$\mathbf{X}: \ 00101110010100001010111010001101000000000101010110010100000111011 \tag{2}$$

in which case the astute observer might form a strong belief that the next character will be the first character in the next word ("our," first code element 1). Concretely, the Bernoulli model would suggest $p(1) \approx 28/66 = .42$, while the Gettysburg model suggests $p(1) \approx 1$. Observers equipped with different models may form different expectations about the next symbol, and hence, different probabilities.[11]

An even simpler and yet more extreme example involves successive draws from an ordinary shuffled deck of cards. What is the probability that the next card will be the ace of spades? With every passing card, the fact that the ace of spades has *not* yet been observed

increases one's belief that it will appear next (because it has to appear eventually). In the limit, after 51 instances of *not* observing the ace of spades, one's belief that it will appear on the next draw rises to 100%. In this example, the *less* frequently something has been observed in the past, the *more* probable it is believed to be in the future—again, because in this case one has as a very non-i.i.d. (permutation-based) model of the data generating process. Indeed, many examples suggesting that probability and relative frequency are approximately the same thing presume an i.i.d. model (as in the taxi example above), because in such a model relative frequency does in fact asymptotically approach the probability assigned by almost any reasonable observer (this is essentially the "law of large numbers"). But this is not true in the general case. The larger point though is that (to a Bayesian) an agent's probabilistic belief that 1 will occur on the $N + 1$-th trial is based on the entire ensemble of beliefs he or she has about the data-generating process and the data observed so far, not just on the *rate* at which 1s have been observed so far.

From this point of view, probability and relative frequency need not be closely related, and of course from an epistemic perspective they are completely different ideas. These examples suggest that all probability estimates depend on a *model*, which is indeed a cliche among Bayesians. This suggests that no probability estimate is completely objective, any more so than any one model of the environment is objectively correct. As Goodman et al. (2015) remark in their rebuttal to Marcus and Davis (2013a), "*an* optimal analysis is not *the* optimal analysis for a task or domain" (emphasis theirs; see also Frank, 2013, for similar points). But while all of this may be self-evident to Bayesians, who already assume an epistemic view of probabilities, it is far from obvious to the larger community, who are accustomed to objective environmental probabilities. Even as Bayesian models have entered the textbooks, frequentism remains part of the zeitgeist.

The goal of this paper is to develop a novel argument against the objectivity of environmental probabilities which does *not* depend on assuming an epistemic view of probability. Rather, my argument only assumes things that all sides can agree on, in particular *conditionalization*, that is, the idea that probabilities change when they are conditionalized on different information. (Even frequentists agree that $p(X|A)$ is not necessarily the same quantity as $p(X)$.) In the mathematical framework, I assume an "objective" data source, and attempt to steer clear of any assumptions that would be considered tendentious. Nevertheless, the mathematical argument will establish that there is still no single, unique probability that can be regarded as the "correct" value of the probability of any particular feature; or, more precisely, that even if you assume a well-defined ground truth probability, there is no way for any real observer to tell what it is. As mentioned above, this certainly does not imply that a Bayesian observer can't be "tuned" to the environment (see Feldman, 2013). But it does contradict the Conventional Wisdom about exactly what this means.

3. Technical argument

The discussion so far has centered on the historical argument about the objectivity of probability, an essentially philosophical dispute. But the central question can be

understood in perfectly concrete mathematical terms, and it has very real ramifications for cognitive theory.

3.1. Multiple channels

In the above examples (e.g. Eq. 1), we imagined an observer who has seen N instances of feature X, and asked what $p(X)$ is on the $N + 1$-the trial. Now, imagine that the observer has access to multiple "channels" or sources of information, which he or she may use to modify (conditionalize) the estimate of $p(X)$. Specifically, imagine that in addition to X the observer has an additional channel A originating from the same world as X, from which samples are drawn in conjunction with X:

$$\begin{aligned}\mathbf{X:}&\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ |\ ?\\ \mathbf{A:}&\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ |\ 1\end{aligned}$$

Clearly, A provides useful information about X—in this case, perhaps that the probability of X on trial $N + 1$ is slightly higher than $\bar{X}$, because X's seem to occur more often when A is 1 than when it is 0, and it is 1 in the $N + 1$-th trial. Hence, in light of A, we might revise our estimate of $p(X)$ upwards. For example, say we want to know the probability that it will rain tomorrow. First, it should be clear that this is nearly meaningless without at least *some* other information, like what planet we are on. (Bayesians usually think of the prior as the conditional probability of a particular state given a set of background information, not "unconditionally" as sometimes stated.) But even once we assume some basic background information, such as that we are on the surface of the earth in 2016, it is clear that additional information will change our estimate of the probability of rain. For example, learning that we are in London raises it, but Cairo decreases it. Such additional information is not limited to one additional channel. More broadly, we may have a second additional source B:

$$\begin{aligned}\mathbf{X:}&\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 1|?\\ \mathbf{A:}&\ \ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ |\ 1\\ \mathbf{B:}&\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ |\ 1\end{aligned}$$

or, more generally, K additional sources $A_1 \ldots A_K$:

$$\begin{aligned}\mathbf{X:}&\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0|?\\ \mathbf{A_1:}&\ \ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 1|1\\ &\ldots\\ \mathbf{A_K:}&1\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 1|0\end{aligned}$$

In what follows, the central question is the value of $p(X)$ on the $N + 1$-th trial, given the values of all channels except X on all trials $1, \ldots, N + 1$. To continue the weather example, the question is: What is the probability of rain given *all* the information we have (that we are in London, that it is September, that it is a Tuesday, etc.). Some factors will raise the apparent probability of rain, others will lower it, and others will seem irrelevant. When all factors are taken into account, what is the *true* value of the probability of rain?

This problem was debated vigorously by 19th century probability theorists, who were concerned that each additional potentially relevant factor narrows the set of comparable cases until, in the limit, only one individual actually fits the criteria, at which point probability ceases to have any meaning (Howie, 2004; Stigler, 1986). Each situation is unique, and it either rains or it doesn't. If we only consider situations *exactly like this one* to be comparable, then there are (by definition) no comparable cases, and hence no relative frequencies, so frequentists were troubled by this question (e.g., Venn, 1888). Generally they argued that probability can only be defined within a suitable class of "equivalent" instances (what von Mises, 1939, called an *ensemble* and Fisher, 1925, called a *population*). But from our perspective, the question is about the probability on the next trial, given the past trials, not the probability of the next trial given an imaginary ensemble of comparable trials. In other words, the cognitive agent needs to make estimates in light of experience and knowledge, not in light of an idealized experiment specially designed to ascertain the applicable probabilities.

In general, we imagine that the world contains a total of D distinct variables or channels (if finite) or an infinite number. Of these, any given observer has access to a finite set numbering $K + 1$, including the target variable X plus a set $\Sigma = \{A_1, A_2, \ldots, A_K\}$ of additional variables to use in estimating $p(X)$. (It should be understood that there is no fundamental distinction between X and the K variables in Σ; we simply notate X differently to indicate it is the variable whose probability we are trying to estimate.) Each distinct "alphabet" Σ of variables defines a distinct observer. An observer with alphabet Σ will have a probability assignment $p_\Sigma(X)$ concerning the target variable X. (Observers can also have different internal models by which they estimate $p_\Sigma(X)$, which is whole additional potential source of belief variation among observers, but to keep things simple the argument pursued below focuses solely on differences is available data.) The central question is how different observers, with different Σs, will compare to each other in terms of their probability assignments $p_\Sigma(X)$.

3.2. Determinism

Before we can answer this question, we need to make some assumptions about the nature of randomness in the system. Attitudes toward randomness, including such central questions as whether the universe is deterministic and why statistical regularities exist, have been vigorously debated for centuries (Hacking, 1990), and are of course tied up in the debates about probability discussed above. The Copenhagen interpretation of quantum theory holds that certain subatomic events are inherently random and do not reflect any

hidden variables, but this conclusion is notoriously controversial (Einstein famously rejected it), and in any case is not generally understood to apply to macroscopic events.[12] Among probability theorists, Bayesians have usually adopted a deterministic conception of the world, because as discussed above in Bayesian theory probability is understood to be epistemic: Uncertainty arises from ignorance, and it does not require the operation of inherently stochastic processes. By contrast, frequentists generally assumed that the environment is genuinely stochastic, because the basic repeated-trials conception of probability requires that actions can be repeated under identical conditions (e.g., coin flips) but yield randomly different outcomes. However, here my goal is to avoid making any assumptions that automatically tilt the scale toward the epistemic side, so as to avoid begging the main question. Hence, the mathematical setup below assumes that some observable variables have random values, without commitment as to whether that randomness is intrinsic or epistemic in nature.

Broadly speaking, most contemporary scientists adopt a deterministic conception of the world, but with the proviso that the number of interlocking variables, and the subtlety of their interconnections, is so enormous that the world is best understood *in practice* as a partly random system. We do not know the weather tomorrow, for example, because the meteorological system has too many interacting (and unknown) variables to model effectively. Unpredictability in observations reflects incomplete data and imperfect models. (Note in passing that this is approximately the Bayesian attitude toward randomness.) But again in what follows, it does not matter if we assume that the world is genuinely stochastic or only appears to be stochastic: We will simply assume some randomness in the data source.

A related issue is whether the real world contains a finite or infinite number of variables (or, more precisely, an infinite number of degrees of freedom, meaning the variation in the system cannot be modeled by a finite set of variables), an issue that is still hotly debated among physicists (see Greene, 2011, for a popular overview). But from the point of view of any finite observer, an infinite world would seem nondeterministic, because whether or not the system evolves deterministically, it will be at least partly unpredictable because some of the variables necessary to calculate its next state cannot be observed. (This point will be developed more explicitly below.) Hence, once again, we (as finite observers) are reduced to modeling the world as if it was finite but nondeterministic, simply because given our finite data the world remains incompletely predictable. This assumption, which I'll refer to as the *finite nondeterministic* (FND) world assumption, seems to be the most common metaphysical stance among working scientists in general. Again, these issues need to be squared away in order to address the central question of what "ground truth" probability is, and whether it exists.

Now, consider a finite world containing D variables. Technically, the world is deterministic if knowing the value of $D - 1$ of these variables exactly determines the value of the D-th. Such a system is completely specified by a finite amount of information; once all variables are known, uncertainty is reduced to zero. In contrast, if after $D - 1$ variables are known, the D-th variable is still unknown, then the system is inherently

nondeterministic; the uncertainty on the last channel is still not completely eliminated even when the rest of the system is fully specified.

With all this in mind, and assuming an FND model, consider what the ground truth probability of one variable X, denoted $p_G(X)$, means. A reasonable definition of "ground truth" in this context is "what would be believed by an omniscient observer."[13] That is, the ground truth value of $p(X_D)$ is, by definition, the probability that would be assigned by an observer who has access to the values on variables $X_1, \ldots, X_{D-1}$. Such an observer knows literally everything there is to know about the system except the actual value of X_D.

First, it should be immediately clear that in a finite deterministic system, ground truth probability *does not exist*, at least in the normal sense. In such a system, by definition, the observer who knows the value of $X_1, \ldots, X_{D-1}$ no longer has any uncertainty about the value of X_D. If you know the settings of all but one of the variables controlling a deterministic mechanism, for example, you can predict the last one too, so you can't, at least in the ordinary sense, assign probabilities to its values.[14]

The critical case, then, is the FND universe; here we can meaningful ask whether a ground truth probability exists. If different observers have different opinions about the probability $p(X)$, what can we say about how different observers' beliefs relate to each other? Can we meaningfully say which is right, or which is closer to the truth $p_G(X)$? In the next section, we explore these questions from a mathematical point of view.

3.3. The observer lattice

In this conception, each observer can be identified by the set of data channels Σ to which he or she has access. One observer, for example, might estimate the probability of rain by considering the set $\Sigma = \{\text{month, temperature}\}$, while another might consider $\Sigma = \{\text{month, humidity, day of week}\}$. (As above, we assume that all observers also have past information about the target variable, in this case the presence of rain; we characterize observers only by the variables other than the target.) The set of possible observers, then, corresponds to the set of possible subsets of the Σ_T, the total set of variables assumed in the universe at hand. As discussed above, each of these agents will, on the basis of the data available to them, have an estimate of $p(X)$. How do they compare?

First, notice that some observers are *subsets* of others, $\Sigma_1 \subseteq \Sigma_2$. In this case, Σ_2 has all the information that Σ_1 has and possibly more, in which case Σ_2 is said to be *better informed* than Σ_1 (*strictly better informed* if $\Sigma_1 \subsetneq \Sigma_2$). However, for other pairs of observers, neither is a subset of the other (neither $\Sigma_1 \subseteq \Sigma_2$ nor $\Sigma_2 \subseteq \Sigma_1$, as in the weather example above) in which case neither is better informed than the other, notated $\Sigma_1 \sim \Sigma_2$. In this case we say that Σ_1 and Σ_2 are *incommensurate*. Formally, the relation $\subseteq$ forms a *partial order*, specifically a lattice which I will refer to as the *observer lattice*. The observer lattice expresses which observers are better informed than others, and which pairs are unordered and thus incommensurate. The lattice can be plotted as a graph, drawing each observer as a node and drawing edges between commensurate observers (e.g., Fig. 1). Because each observer is simply a subset of Σ_T, this is simply the lattice of subsets of

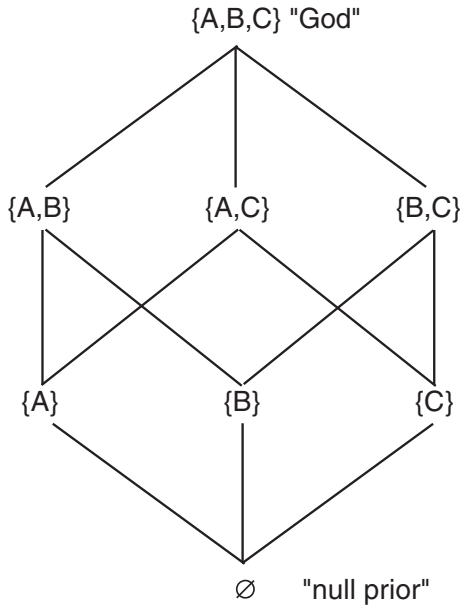

Fig. 1. The lattice of observers for $\Sigma_T = \{A, B, C\}$.

the set Σ_T (see Davey & Priestley, 1990). For example, Fig. 1 shows the lattice of observers for 4-variable universe $\{A, B, C, X\}$. Recall that X is not part of the Σ, so this lattice is simply the set of subsets of $\{X, Y, Z\}$. At the top of the lattice is the most informed observer ("God") whose beliefs about $p(X)$ define the ground truth. At the bottom of the lattice is the least informed observer $\varnothing$, who knows nothing at all about variables other than X. The least informed observer defines what could be thought of a "null prior" for X, because it quantifies the probability $p(X)$ integrating over (that is, assuming ignorance of) all other variables in the universe. Fig. 2 shows a larger ($D = 4$) lattice with examples of commensurate and incommensurate observer pairs marked.

In an FND universe, it is meaningful to ask how well various observers approximate the ground truth (the top of the lattice). We might imagine, for example, that as observers climb the lattice from least informed to best informed, they progressively approach the ground truth. We now show that this is in fact not correct.

Consider two observers Σ_1 and Σ_2 corresponding to respective probability assignments p_{Σ_1} and p_{Σ_2}. Σ_2's beliefs are *more correct* than the Σ_1's, written $p_{\Sigma_1} \prec p_{\Sigma_2}$, iff $p_{\Sigma_2}(X)$ lies between $p_{\Sigma_1}(X)$ and ground truth $p_{\Sigma_T}(X)$ (that is, if $p_{\Sigma_1}(X) \leq p_{\Sigma_2}(X) \leq p_{\Sigma_T}(X)$ or $p_{\Sigma_T}(X) \leq p_{\Sigma_2}(X) \leq p_{\Sigma_1}(X)$). If $p_{\Sigma_1} \prec p_{\Sigma_2}$, then p_{Σ_2} is a "step in the right direction" relative to p_{Σ_1}. Two such observers have a definite ordering in terms of their proximity to the truth. What conditions allow this?

First, as the terminology suggests, it should be apparent that incommensurate observers cannot be compared in terms of relative truth. If $\Sigma_1 \sim \Sigma_2$, then neither is better informed than the other, and there is no general way of telling which will end up with a more accurate estimate. (See Appendix for a proof sketch.)

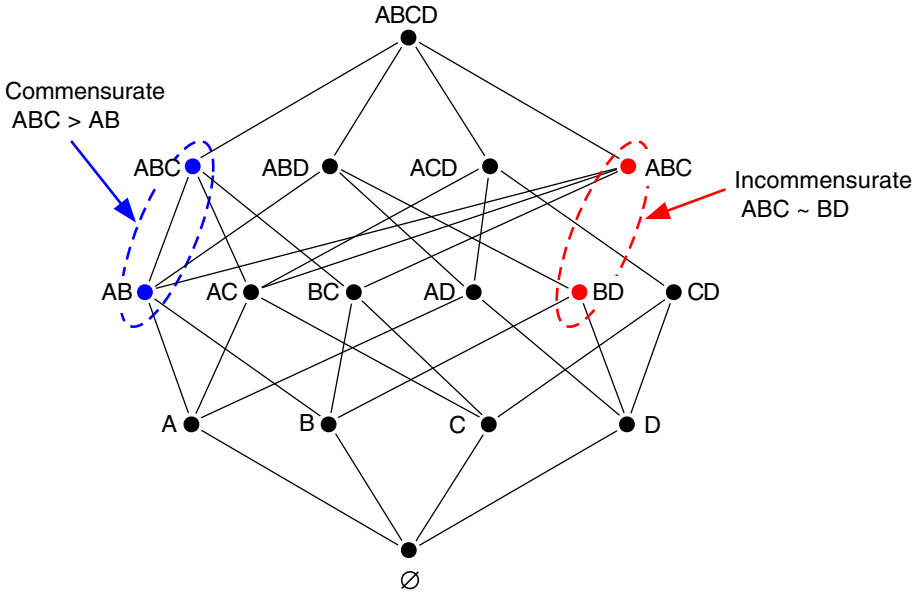

Fig. 2. The lattice of observers for $\Sigma_T = \{A, B, C, D\}$, showing a pair of commensurate observers (blue) and a pair of incommensurate ones (red). Note that *most* pairs of observers are incommensurate.

The more difficult case is that of commensurate observers. In this case, one observer *is* more informed than the other. Does this mean that its estimate will be closer to the truth? In other words, if $\Sigma_1 \subset \Sigma_2$, then does it necessarily follow that $p_{\Sigma_1} \prec p_{\Sigma_2}$?

Perhaps counter-intuitively, the answer is still no. More informed observers can, notwithstanding, have probabilities that are farther from the ground truth.[15] Fig. 3 gives a graphical depiction of an example (see Appendix for a proof sketch). The figure illustrates a $D = 2$ FND universe ($\Sigma_T = \{A, B\}$, plus target variable X), giving probability mass (red dots, all considered equiprobable) for all 2^{D+1} possible cases. Consider the case $A = 1$, $B = 1$. Ground truth $p(X)$ (defined by the observer $\{A, B\}$) is 1/2 (equal probability mass at $A = 1$, $B = 1$, $X = 1$ as at $A = 1$, $B = 1$, $X = 0$). An observer who knows neither A nor B (the null prior, $\Sigma = \varnothing$) will believe $p(X) = 2/5$), lower than the true value. Adding the knowledge that $A = 1$ pushes the probability $p(X = 1)$ *up* to 2/3, in the correct direction toward the truth (1/2) though actually "too far." But adding that knowledge that $B = 1$ (instead) pushes the probability $p(X = 1)$ *down* to 1/3, in the wrong direction away, from the true value. That is, the progression of the probability estimate as one climbs the observer lattice is *nonmonotonic*: adding information can improve one's estimates or degrade them, with no way (from the observer's point of view) to know which. Only adding information about *all* factors—that is, becoming omniscient—is guaranteed to bring us to the true value.

Fig. 3 is only an example, but it illustrates in principle that better informed observers are not necessarily closer to the ground truth. As one gains information (collects more information channels), one's probabilistic beliefs can fluctuate up and down, and are in

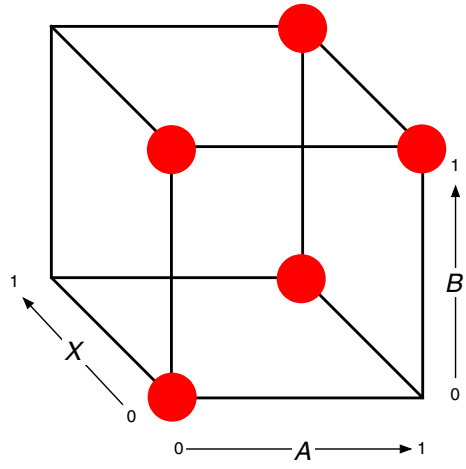

Fig. 3. Illustration of the non-monotonicity of probabilities as information increases. Red dots indicate probability mass for each of the 2^3 combinations of $\{A, B, X\}$ (each dot indicates an equal quantity of probability). As an agent progresses from the null prior $p(X = 1) = 2/5$ to the more informed observers $\{A\}$ or $\{B\}$, probability can either rise or fall (respectively to $p(X = 1|A = 1) = 2/3$ or $p(X = 1|B = 1) = 1/3$), giving no assurance that one is approaching the ground truth probability ($p(X = 1|A = 1, B = 1) = 1/2$). Acquiring more channels does not necessarily bring one closer to ground truth.

no sense guaranteed to progressively converge on the ground truth. Only when *all* extant information is included (which is only possible in a finite universe), and only if the universe is nondeterministic (or else ground truth probability is not defined), will the probability arrive at the true value. The intuition that additional information leads to an inexorable march toward objective truth is, at least without additional assumptions, unfounded.

A useful way to look at this is that one cannot, in general, approximate the ground truth probability by progressively acquiring more data channels. Additional data can move the observer's probability assignment either toward or away from the ground truth, and there is no way to tell (from the observer's point of view) which way is which. As discussed above, frequentists define probability as the limiting value of the relative frequency in an infinite series of trials. Bayesians like Jeffreys (1939/1961) and Jaynes (2003) objected to this definition because (among many other reasons) it imagines an infinite sequence, the sequence of relative frequencies as N increases, and assumes that this sequence has a limiting value *without demonstrating that the sequence converges*—an elementary mathematical mistake. Indeed, empirical data sequences cannot be assumed to converge to stable values, because the data-generating conditions may change in unknown ways (your coin can change properties). If the data are assumed to be generated by a fixed, infinitely repeating Bernoulli process, then the limit does exist (as shown by J. Bernoulli himself) and to be equal to the true probability. De Finetti generalized this result to what he termed *exchangeable* sequences, again setting conditions for convergence, but without requiring any notion of "true" probability as the limiting value. (Again De Finetti

was an extreme subjectivist, so he did not accept the notion of "true" probability. The importance of exchangeability will be discussed further below.) The current argument gives a comparable result, though a negative one, in a literally orthogonal dimension (vertically rather horizontally in the examples above—that is, in the D direction rather than in the N direction). What we have shown here is that increasing the number of *dimensions* (rather than trials) does not, necessarily, cause the probability to converge on the true probability. Independent of how many datapoints you have, gathering more variables sometimes helps and sometimes hinders. If a ground truth exists, you are not guaranteed to arrive at it (or even near it) until *all* dimensions have been acquired—at which point probability becomes degenerate anyway. In this sense, there is no guarantee of progressive approximation to the truth.

It is worth noting that most pairs of observers are, in any case, incommensurate. For a universe with D variables, each observer has a sublattice of less informed agents hanging below it, and a sublattice of more informed agents above it; it is commensurate only with agents in one of these two groups. By symmetry, these each number about $2^{D/2}$, for a total of about $2^{D/2+1}$ incommensurate observers, out of a total of about 2^D. Hence, the number of incommensurate observers in any given universe is roughly the square of the number of commensurate ones, and thus enormously outnumber them as D rises.

Several more mathematical definitions generalize the above situation. For any two observers Σ_1 and Σ_2, there is always a *least common superset* (*supremum*) $\Sigma_1 \cup \Sigma_2$, defined as the smallest observer S such that both $\Sigma_1 \subseteq S$ and $\Sigma_2 \subseteq S$. This is the smallest observer that is more informed than either Σ_1 or Σ_2 and can thus be thought of as their "relative ground truth." The relative ground truth is an observer who, like the omniscient observer who knows the values of all variables, knows all the variables that either Σ_1 or Σ_2 knows—but no more. Similarly, there also exists a *greatest common subset* (*infimum*) $\Sigma_1 \cap \Sigma_2$, which is the largest observer S such that both $S \subseteq \Sigma_1$ and $S \subseteq \Sigma_2$. (See Bennett, Hoffman, & Murthy, 1993, for related mathematical constructions.) The greatest common subset observer is the largest observer that is less informed than either, and thus can be thought of as the *relative null prior*. Like the (absolute) null prior, this observer is at least as ignorant as either observer, but no more so. More broadly, any set of observers $\{\Sigma_i\}$ has both a relative ground truth $\bigcup_i\{\Sigma_i\}$ and a relative null prior $\bigcap_i\{\Sigma_i\}$ (see Davey & Priestley, 1990; Grätzer, 1971). These notions are useful in thinking about the "common knowledge" connecting a set of agents (Richards, McKay, & Richards, 2002), defining respectively the union and intersection of what they all know. Moreover, the relative ground truth is a natural way to understand the apparent "ground truth" in any FND model of a complex deterministic universe (for example, the "true" probability of rain tomorrow). Rather than being truly objective as the real ground truth should be, it is simply the restriction of the notion of ground truth to a finite set of observers (the union of which incorporate all factors believed relevant to the weather). Of course, if some factors outside this set turn out to actually affect the weather, then the "ground truth" relative to this observer set will no longer be "true." And again you can't simply incorporate *all* factors in the universe, because at that point the weather is no longer stochastic and the rain no longer has a probability at all.

Summarizing, an FND universe includes a variety of observers, each with a different set of data upon which to base their estimate of $p(X)$, and no way to tell which one is "right." Most pairs of observers are intrinsically incommensurate, with no way to tell whose probability assignment about X is closer to the truth. Even when one observer is strictly better informed than another, it is still impossible to tell which is closer to the truth. Hence, even in the limited sense in which ground truth exists in an FND universe, it is impossible to tell which observer has access to it.

All of this leads inevitably to the conclusion that CW1—the assumption that events or conditions occur in the environment with objectively true probabilities—is deeply flawed. Even though we explicitly assumed the existence of a ground truth probability p_G, bending over backwards to accommodate an ontological stance somewhat foreign to Bayesianism, the conclusion is that p_G is in a deep sense inaccessible to real observers. Even though it exists, gathering information from the world does not help figure out what it is. Of course, this conclusion is perfectly consistent with the epistemic view of probability, which does not accept that p_G exists in the first place. But the point is that while Bayesians usually accept this assumption on philosophical grounds, and because it comports with the goals of inverse probability, here we have arrived at it starting from neutral premises.

In other words, even if you do not accept in principle that probability values are subjective characteristics of observers, you cannot avoid the conclusion that they depend in a very concrete way on the observer's epistemic characteristics, that is, on what information they have available to them, formalized in the above by Σ. More specifically, given two observers with different epistemic status (different Σs), the generic situation is that they will disagree about the probability of any given feature of the world, and that neither can be meaningfully said to be more right than the other. More precisely, one will be closer to the ground truth (if it exists) than the other, but there is no principled way to say *which*. In this sense, probability is subjective, and CW1 is substantively false.

3.3.1. Exchangeability and the possibility of prediction

The above conclusion is, admittedly, extremely counter-intuitive. Anyone experienced with probabilistic prediction knows that some predictors provide more information about the target variable than others. Statistical regression, for example, is a technique for determining which factors provide more accurate predictions of a target variable. Surely a meteorologist who has hundreds of weather-related measurements will make a more accurate prediction about the rain tomorrow than an amateur with a barometer?

Perhaps, but not without making more assumptions than we have made so far. The key limitation so far is that we assume the various data streams in Σ can take arbitrary values; the future is not constrained to resemble the past in any particular way (cf., Goodman, 1954's grue/bleen problem). One cannot simply assume that future data will bear a statistical resemblance with past data, because that can mean literally anything without making some assumptions about what "statistical resemblance" means. Recall the discussion above of decks of cards; the ace of spades becomes gradually more probable because it does *not* resemble past cards.

One way to see this more clearly is to introduce an additional assumption that does, in fact, allow evidence to vary in predictive power in the ordinary way. The key assumption is De Finetti's notion of *exchangeability* (de Finetti, 1964 2008), which for this very reason is sometimes regarded in Bayesian theory as foundational—because in a sense it licenses statistical prediction itself. A set of observations is said to be exchangeable if they can be freely permuted without changing the resulting expectations (see Lindley & Phillips, 1976; Zabell, 2005, for discussions). In the examples given above, we assumed a particular sequence of N observations. Are these exchangeable? If we assume that the order in which they occur is immaterial, like coin flips, then by definition they are exchangeable. But if they have inherently sequential properties, like draws from a deck of cards, or words in a sentence, then they are not exchangeable, because permuting their order would change expectations about what is going to follow.

De Finetti's representation theorem shows, essentially, that if a series of observations is exchangeable, then there exists a probability distribution describing its output. In our context, that means that if the sequence of observations of our data channels Σ and target variable X is regarded as exchangeable, then there exists a joint probability distribution $p(\Sigma \cup X)$ giving the joint distribution over all the variables, and in particular a conditional distribution $p(X|\Sigma)$ giving the probability of each possible value of the target variable X conditioned on observing the data channels Σ. The degree to which X is (un) predictable given Σ is quantified by the conditional uncertainty $-\log p(X|\Sigma)$, which in turn depends on the mutual information $H(X, \Sigma)$ that X shares with the other channels (see Cover & Thomas, 1991). Clearly, some choices of Σ will have more mutual information with X than others, and as a direct consequence, some observers will make predictions about future observations of X that are more likely to be correct. Once we assume exchangeability, a probability distribution governing X—including its covariation with other variables—can be said to exist, and all of these intuitive features of ordinary prediction follow (see Jordan, 2010). Of course, the attribution of exchangeability still does not guarantee that future data will resemble past data, but it does guarantee the existence of an "i.i.d.-like" probabilistic model of the data source that can serve as the basis for prediction.

But *are* the observations exchangeable? It depends on your model, and specifically on what information you have available, in a way the observer lattice setup makes clear. For an observer with alphabet Σ, it is reasonable to assume exchangeability on the entire ensemble of data channels, because by definition there are *no other data channels* that might distinguish the observations from each other, making them ipso facto exchangeable. That is, by assumption, the observer is blind to any additional channels not in Σ that might distinguish them. However, a superset observer $\Sigma \cup A$ might not consider Σ to be exchangeable, because the data in the additional channel A distinguish different observations of Σ. The additional channel breaks, or at least has the potential to break, the symmetry entailed by exchangeability.

The conclusion is that whether a given data sequence appears exchangeable, and thus whether or not we can make meaningful distinctions between better and worse predictions, depends on our epistemic state. Exchangeability is essentially an expression of

ignorance about features that might distinguish cases. But observations that seem interchangeable to one observer will seem distinct to a better informed observer, and therefore by de Finetti's theorem incapable of being rolled into a single probability distribution. All of our usual intuitions about the relative value of different predictors depend on this highly observer-dependent assumption and cannot be justified universally.

4. Frequently asked questions

Before moving on to discuss CW2 and CW3, we should pause to acknowledge that the subjectivity of probabilities is a very counter-intuitive conclusion to anyone accustomed to thinking about probabilities as characteristics of external events like coin tosses. With apologies for the informality, I will address some natural objections in the form of a list of "Frequently asked questions" (FAQ).

Q. But isn't it *objectively true* that an ordinary coin has heads probability .5? Are you seriously suggesting it might be .9 or something?

A. No. That may be what you think, and I think, and all of us think, but that does not make it *objectively* true. The probability .5 is really just a statement about your state of knowledge—that you have no idea which of the two sides will come up. That is your epistemic state, and it seems objective only because it is the nearly universal epistemic state of observers watching coin tosses—we normally expect the coin to be physically symmetric, and we almost never have any information before the coin is tossed that would allow us to predict which side will come up. But if another observer watching the same coin toss sees, say, a glint of light on the coin as it spins through the air, which she is savvy enough to understand means that the ensuing toss is actually most likely to be heads, then she might think the probability is .9. At the same time, another savvy observer on the other side of the room thinks the coin tosser is a con man and the coin has two tails. Who is right? Everybody. All probabilities simply reflect their believers' respective states of knowledge.

Q. But surely the world is characterized by consistent statistical regularities—robins are usually red, it is usually hot in summer, bread usually falls butter side down. Are you denying that these tendencies truly hold?

A. No, but the world exhibits consistencies not because its governing probabilities have external existence, but because (a) it's *physically* consistent—its physical properties tend not to change very much over time—and, equally importantly (b) the epistemic conditions under which we observe it are also generally consistent. Almost all of us observing a coin toss have the same very limited information about the throwing process, and thus no basis for predicting which way it will fall. But if we change the epistemic conditions, the probabilities change, even if the world does not. The probability of rain changes when we see a cloud, feel the humidity rise, or for that matter see a weather forecast, any of which change our epistemic status. Thus, even though the world has stable physical characteristics, the probabilities of particular characteristics will be different for different observers, and it is not really meaningful to say that some are right and some are wrong.

Q. Are you just saying that phenomena in the world aren't real? If one is a "realist" in the philosophical sense, meaning that one accepts the actual reality of things outside the head, does that negate the argument here?

A. No, that is not the issue. There is legitimate debate about the objective existence of physical qualities (see Eddington, 1928; Hoffman, 2009; Hoffman & Prakash, 2014; Koenderink, 2011, 2012, for skeptical views), but it is not taken up here. In the current paper there is no question of the actual reality of the *data*—just of the trend or pattern describing the data's behavior, as quantified by a specific probability value. The question is whether this pattern has objective existence *apart from the data itself*. If it does, this would imply that probability has a definite value no matter what the observer thinks (as frequentists traditionally argue). Alternatively, this pattern can be viewed as an induction from the data by an observer, meaning that different observers can have different probabilities and be equally "right" (as Bayesians traditionally argue). This suggests a kind of "radical relativism," in which different observers are free to draw strikingly different conclusions from the same world. But they are not disagreeing about the facts—they are disagreeing about the *propensities implied by the facts*.

Q. So this is just a fancy kind of skepticism? You can believe anything you want?

A. No. In Bayesian theory there is still a rational connection between the evidence available to you, the models you have chosen to consider, and the conclusions you draw —namely, Bayes' rule. But there is no "right" set of evidence and models. Different observers have different models and different sources of evidence, and thus draw different conclusions. One is not better than the other; they are just different.

5. Optimality

The second element of the conventional wisdom, CW2, holds that inference will be optimal in a given environment if the observer's assumptions match the objective statistical characteristics of that environment. Our estimates of tomorrow's weather, for example, are best when our prior for rain matches the objective probability of rain. But there cannot be an optimal observer with respect to the objective statistical characteristics of the environment if, as per the above argument, the environment doesn't *have* objective statistical characteristics. How, if at all, can we reconcile the notion of optimal inference with the subjectivity of probabilities?

The idea of optimal inference is often expressed in terms of an *ideal observer* (Geisler, 2003; Green & Swets, 1966), meaning a decision-making procedure that is provably optimal relative to a particular data-generating model. For example, an ideal observer attempting to distinguish samples of parameter x drawn from a unit normal centered at 0 ($x \sim N(0, 1^2)$) versus a unit normal centered at 1 ($x \sim N(1, 1^2)$) will decide based on a threshold at $x = 0.5$. Such a strategy minimizes the expected rate of errors.

Ideal observers are widely identified with Bayesian inference, because the use of Bayes' rule is demonstrably optimal relative to some set of prior assumptions (Cox, 1961; Jaynes, 2003). Informally, we often say that an ideal observer is an optimal

decision-maker in a given environment, meaning, in an environment in which certain probabilities hold. But if, as per the above argument, it is not really true that certain probabilities hold in a given environment, then is the notion of ideal observers invalid?

A slightly more careful consideration of exactly what is "ideal" about ideal observers gives a fairly simple resolution to this conundrum. The mathematics whereby one proves that the probability of error is minimal in an ideal observer necessarily makes assumptions about the data-generating processes at work in that environment. These assumptions are, of course, only a model. Equally importantly, any model makes assumptions, whether implicit or explicit, about the *epistemic* state of the observer: exactly which variables are observable and which ones are random from the observer's point of view. As discussed above, assuming that a given variable is "random" either means that it is intrinsically stochastic (if one assumes nondeterminism, as in some versions of quantum mechanics) or (as is more common) that it is the result of a complex system not all of whose variables are observable. In any real situation, of course, one cannot guarantee that variables hidden from one observer are not visible to another. This is the situation, for example, when one observer thinks the probability of rain tomorrow is 30% because (say) he knows only that rain occurs on 30% of days in his locale, while another thinks that it is 80% because she has seen dark clouds on the horizon. The two observers have different probabilities for rain in the *same* environment, because they have different epistemic states (different Σs). As argued above, neither is necessarily more right than the other. And lest one be tempted to conclude that the second observer has a better estimate than the first because she has more information, note that the mathematical argument above shows this to be unwarranted; additional information can be misleading.

All this suggests a simple conclusion: Ideal observers are optimal with respect to a *model*, including the epistemic assumptions in the model, not with respect to an environment per se. Any real environment may or may not obey the model, and any real observer may or may not satisfy the epistemic limitations it assumes. This is why, for example, better-than-ideal performance is perfectly possible in real environments: because the observer may have access to information assumed unobservable in the model from which the ideal was calculated (cf., Liu, Knill, & Kersten, 1995).

The confusion of model with reality is particularly tempting in laboratory experiments, where we actually program the computer that generates the stimuli, thus fully controlling the "environment" of the subject. Again, this determines not only the data-generating model, but also the epistemic conditions, such as which variables the subject can observe and which are hidden. Random parameters in an experiment (e.g., the order of trials) are determined by a random-number generator inside the computer. We know the subject cannot derive any useful information from such a parameter, because the parameter is chosen randomly independent of stimulus properties and thus does not contain any useful information about those properties. But an analogous parameter in the corresponding natural environment (say, the time at which a given stimulus is observed) might very well contain useful information, because it has *not* been chosen by a random-number generator, but rather is simply another variable in the complex unseen data-generating mechanism and may well correlate with variables of interest. In an experiment, the environment

necessarily obeys the model, making the distinction between environment and model moot. But in a real environment, there is no principled limit on what potentially relevant information the observer may glean from the environment, meaning that the observer can exceed ideal performance relative to any particular fixed model. That is, in an experiment, we control the *epistemic* conditions of the subject—what information he or she has access to—whereas in the real world, the observer is free to tap additional information sources, with concomitant changes to the probabilities.

All the above is, I think, fairly obvious to theorists who use ideal observers, but it is not quite reflected in the way we usually speak of probabilities. "An ideal observer for environment X" is simply *shorthand* for the more cumbersome "An ideal observer for an environment that obeys model X." But in the commodification of Bayesian models to the wider cognitive science community, the more careful version is often elided to create what is really a category error. The idea that "statistical properties of the environment" exist independent of the observer is an instance of Jaynes's "mind projection fallacy"—like thinking that *north* in the world is really *up* because it always points up in maps.

6. Evolution

If CW1 is false (objective probabilities generally do not exist) and CW2 is false (because one's priors cannot match a probability that does not objectively exist), it seems to follow immediately that CW3 is false as well. One cannot be guided by natural selection toward assumptions that are true in the environment if there is no such thing. But surely natural selection *has* guided the assumptions underlying our perceptual and cognitive faculties. If our assumptions are not steered toward the truth, what are they steered toward?

The answer is readily apparent when one considers how natural selection applies to statistical decisions. Evolution steers us toward the strategies that bring us the greatest reward, not toward ones that help us arrive at "truth" per se (Gigerenzer, Todd, & the ABC Research Group, 1999; Hoffman & Singh, 2012; Mark et al., 2010; Singh & Hoffman, 2013). Perceptual inference has consequences for the observer only via some notion of utilities or payoffs, or as Bayesians usually call it, a loss function. Evolution steers us toward the actions that bring the greatest benefit (minimum loss) and thus increase our genetic representation in the next generation. We are not rewarded for having a prior that is "accurate" about reality, nor are we rewarded for settling on a posterior that expresses the truth—whether or not those conditions actually mean anything. We are simply rewarded for taking actions that lead to reproductive success.[16]

The difference between optimizing belief and maximizing payoff is well developed in the vast literature on decision making (Maloney, 2002; Savage, 1954). The CW in effect conflates these two notions by assuming, tacitly and without careful consideration, that true beliefs inevitably maximize utility. Indeed, it makes no sense to assume that the environment will reward us for having true beliefs, when (a) as argued above, the very notion of "true" probabilistic beliefs is ill defined, and (b) the environment has no way of knowing the content of our beliefs—only our actions. The bottom line is that this

argument does not contradict the fundamental idea that the mind is tuned to the world (Feldman, 2012; Richards & Bobick, 1988; Shepard, 1989)—it just forces us to think about the "tuning" in a more subtle way, stripped of the idea that the statistics of the real world provide a "ground truth" to validate the prior.

7. Discussion

In summary: notwithstanding widespread intuition to the contrary, probabilities don't have objective values in the environment. Probabilities are mental states, not mind-independent characteristics of the outside world; they are characteristics of models, not reality. This means that CW1 is substantively false, and CW2 and CW3 concomitantly incoherent. This is not a technicality! While it is true that some models of the environment may be *almost* true, in that the predictions they generate tend to be supported by further data, that is true of a wide variety of models, and no one model is *generally* superior to all others.

7.1. If the CW is wrong, what's right?

So if the CW is wrong, what's right? The alternative position implied by the above argument boils down to approximately the following summary:

(A1) The probability $p(X)$ depends on the epistemic state (including Σ) of the observer.
(A2) No observer (except an omniscient one) is generally superior to all others.
(A3) Natural selection applies adaptive pressure on organisms to maximize utility, which is generally unrelated to the "truth" of their probabilistic beliefs—which is ill defined anyway.

This alternative view, and in particular A3, implies that observers—including ourselves—can benefit maximally from beliefs that are not in any meaningful sense "true," but are rather, one might say, "helpful fictions." Such a position has indeed been articulated by Hoffman and others under the label "user interface theory" (Hoffman, 2009; Hoffman & Prakash, 2014; Hoffman, Singh, & Prakash, 2015; Koenderink, 2011, 2015). This possibility is inherently counterintuitive, because it implies that our intuitions about the actual state of the world are, in a deep sense, illusory. But conflict with intuition is certainly not fatal for a scientific theory, especially when the intuitions themselves are what we are trying to explain (see Feldman, 2015), so that objection can, in my view, easily be set aside.

7.2. Takeaway

The ancient debate about the nature of probability hinges on whether probabilities have objective values or whether they inherently depend on the epistemic characteristics

of the observer. In this paper, I have attempted to clear some of the philosophical brush away from this question by proposing a simple technical definition of the "epistemic characteristics" of an observer (namely, its ensemble of data channels Σ); and then asking what consequences variation in epistemic status has on the probabilities that the observer will adopt. The conclusion was that, in a concrete technical sense, there is no one observer-independent value of a probability—no one "true statistics of the environment"—but rather a diverse array of beliefs of incommensurate validity. This points to a kind of radical epistemic relativism, in which distinct observers have distinct models of probabilistic reality, all of which are internally coherent and equally justifiable.

So why does this matter for working cognitive scientists? I suggest two broad answers, one negative and one positive.

7.2.1. Negative

Given the argument laid out above, the objectivity of probability is a myth that has enormous intuitive sticking power. Humans, like any observer, inherently cannot distinguish their own *models* of reality from reality itself. Thus, we confuse probabilities derived from our models with probabilities inherent in the outside world. And as in any other science, cognitive science cannot move forward without casting aside false premises, especially those that derive from persistent but misleading intuitions.

In this sense, the idea of objective environmental probabilities is like the "great chain of being" which dominated thinking about biological diversity before Darwin. The presumption, widely shared by religious and scientific thinkers alike, was that biological forms inhabited a scale of perfection, with "lower" forms inherently inferior to "higher," and humanity (naturally) at the top. Even after Darwin, a mistaken intuition persists that evolution represents a progression toward objectively superior forms. But as modern evolutionary theorists are often at pains to clarify, contrary to this intuition, each biological form is simply adapted to its own environment (modulo the limitations of adaptation), with no form generally superior to any other. Similarly, the CW implies that the environment is objectively characterized by one particular set of probabilities (the true priors), and that observers correctly equipped with these probabilities —presumably humans—are in some sense "more rational" than all the others. The idea that adaptation pushes us toward these objective priors is a natural corollary. In light of the arguments in this paper, this idea is simply wrong and gets in the way of a more rigorous understanding of the relationship between observers' beliefs and their environments.

Putting this another way: the idea that the mind is tuned to the world is one of the most profound in modern cognitive science and neuroscience, but exactly what this means and how it works is a critical unsolved problem. The simple answer suggested by the CW—that this process is complete when the observer's priors match "the statistics of the environment"—short-circuits this question and thus forestalls real progress in answering it.

7.2.2. Positive

At the same time, the framework introduced above opens the door to some fruitful avenues of research. First, viewing probabilities simply as beliefs conditioned by data and internal models sheds light on the sometimes enormous diversity of viewpoints and "mental sets" that different individuals bring to perceptual situations both in and out of the laboratory—not to mention different cultures and species. In the CW, each individual observer's priors and likelihood models (etc.) are either ideal with respect to the environment, or fail to be ideal in some way, since only one can match the environment exactly. Adopting a more epistemic stance makes it clear that even highly discrepant observers (say, a cat and its human companion viewing the same television program) simply embody different arrays of knowledge, beliefs, and goals. This suggests a more "democratic" viewpoint in which all these observers model the world in distinct ways, all of which work on their own terms—or, more accurately, might or might not work, depending on their tangible consequences for the observer, and not on their truth per se. (Cf. von Uexküll's notion of *Umwelt* as discussed in Koenderink, 2015.) Recall from the mathematical argument above that most pairs of observers are incommensurate—meaning that, even stipulating the existence of a well-defined ground truth probability, neither of them is consistently closer to it than the other. This pluralistic perspective is surely more conducive to understanding the amazing diversity of cognitive strategies employed by distinct individuals, cultures, and species.

Moreover, some of the technical ideas introduced above are potentially useful tools for modeling relations among rational observers. For example, the relative null prior $\bigcap_i\{\Sigma_i\}$ and the relative ground truth $\bigcup_i\{\Sigma_i\}$ introduced above, which refer respectively to observers' upper and lower bounds in the observer lattice, are potentially helpful mathematical reference points when considering a set of epistemically distinct agents observing a common environment. These constructs are not quite intelligible if one assumes (as in the CW) a unique prior specified by the environment. But once this idea is cast aside such, they can be appreciated and potentially incorporated into modeling efforts.

Finally, the perspective developed above allows a potentially more productive way of looking at another fundamental unanswered question: How exactly is the mind tuned to the world? As argued above, this question is given a facile answer by the CW—the mind is properly tuned when it adopts empirically correct priors. A better answer is about adaptive consequences: Beliefs are validated not by their truth but by their effects on reproduction. This is the common motivation of both the "interface theory" of Hoffman et al. and the "ecological rationality" framework of Gigerenzer et al. (though beyond this issue these frameworks differ in many respects). There are a number of legitimate controversies about exactly how probabilistic accounts of cognition should be grounded, for example pitting a "pure" Bayesianism based on coherence (e.g., Lindley, 1972) against a consequentialist view based on adaptation (e.g., see Arkes, Gigerenzer, & Hertwig, 2015). But this essential debate cannot be advanced, nor even really understood, until the field sheds the CW's fixation on the literal empirical truth of probabilities.

8. Conclusion

In Conventional Wisdom, cognitive agents can achieve optimal inference by adopting a statistical model that is close to the true probabilities governing the environment as possible, and they are relentlessly driven by evolution toward such a model. In the subjectivist framework advocated here, distinct observers form an interconnected network of partially overlapping but distinguishable belief systems, none of whom has special claim to the truth. On this view—as in traditional Bayesian philosophy—"true" probabilities are not accessible and play no role. To speak of certain environmental probabilities as objectively true—no matter how accustomed many of us are to speaking that way—is a fallacy.

Acknowledgments

I am grateful to Manish Singh, Melchi Michel, Amy Perfors, Michael Lee, David Danks, and an anonymous reviewer for extremely helpful discussions. The author was supported by NIH EY021494.

Notes

1. Needless to say, there are many other views of probability other than epistemic and frequentist—indeed, there are probably as many views as there are writers on the subject. Of particular note is the *propensity* view, introduced by Karl Popper (e.g., see Popper, 1959), intended to be a middle way that avoids both the problems of the frequentist view and the subjectivity of the epistemic one. However, the goal of this paper is not to settle the endless controversy about the nature of probability, but merely to address one critical aspect of that debate in connection with cognitive science. See elsewhere (e.g., Earman, 1992; Mellor, 2005) for more comprehensive treatments of the many views of probability, and Howie, 2004; Stigler, 1986; von Plato, 1994; Zabell, 2005, for historical perspectives on the debate.
2. "All propositions are true or false, but the knowledge we have of them depends on our circumstances; and while it is often convenient to speak of propositions as certain or probable, this expresses strictly a relationship in which they stand to a corpus of knowledge, actual or hypothetical, and not a characteristic of the propositions in themselves."—J. M. Keynes (1921)
3. And this is why in a traditional statistics class, following the shibboleths of frequentism, you are not allowed to say "the null hypothesis is probably false"; you have to say the "null hypothesis is rejected." Hypotheses do not have probabilities.
4. It should be noted that Bayesian probability theory is used in a number of distinct ways in cognitive science. Some researchers use it as a theoretical model for

human inference, while others use it as a statistical tool for analyzing data (and many do both; see Lee, 2010, for discussion). The argument in this paper relates most directly to the former. The epistemic view of probability is certainly no less important in the latter, statistical estimation, but the widespread misconstrual of Bayesian philosophy of probability described below is much less common in that context.

5. What *does* divide subjectivist from objectivist Bayesians is the degree of individual freedom afforded observers in choosing priors (and thus posteriors). Objectivists contend that all rational observers, given the same data, should converge on the same beliefs. Subjectivists emphasize the individual nature of the prior and allow distinct observers more freedom in choosing it (see, e.g., de Finetti, 1970/1974; Jeffrey, 2004; Savage, 1954). However, both sides agree that probabilities are not "objective" characteristics of the world. Jaynes (2003) explained his notion of objectivism thus:

> (A) The prior probabilities represent our prior information, and are to be determined, not by introspection, but by logical analysis of that information. (B) Since the final conclusions depend necessarily on both the prior information and the data, it follows that, in formulating a problem, one must specify the prior information to be used just as fully as one specifies the data. (C) Our goal is that inferences are to be completely "objective" in the sense that two persons with the same prior information must assign the same prior probabilities. (Jaynes, 2003, p. 373)

But elsewhere in the same book he reaffirms the common view of subjectivist and objectivist Bayesians that probabilities represent states of belief (or ignorance) on the part of the observer, and they do not have real existence outside the head. "Probabilities change when we change our state of knowledge; frequencies do not" (Jaynes, 2003, p. 292) and later: "It seems to us that the belief that probabilities are realities existing in Nature is pure mind projection fallacy" (Jaynes, 2003, p. 411). Elsewhere he elaborates:

> The Old Sermon Still Another Time: [. . .] In orthodox statistics, a sampling distribution is always referred to as if it represent an "objectively real" fact, the frequency distribution of the errors. But we doubt whether anybody has ever seen a real problem in which one had prior knowledge of any such frequency distribution, or indeed prior knowledge that any limiting frequency distribution exists. How could one ever acquire information about the long run results of an experiment that has never been performed? That is part of the Mind Projecting Mythology that we discard. (Jaynes, 2003, p. 608)

6. Another ambiguity is that the term *subjective probability* is sometimes used to mean simply a subjective estimate of probability, as contrasted with the objective

value it actually holds in the environment. Though this distinction is obviously related to the difference between subjective and frequentist conceptions, in that one is in the head and the other the world, it is quite different because it presumes that subjective probabilities are simply *incorrect estimates* of objective ones. In some contexts, this usage seems to arise from a misunderstanding of the classical dichotomy. The classical debate is not about *what value* is placed on the probability, but what probability itself *means*—an objectively stochastic value measurable through repeated experiment (frequentist) versus a quantification of belief (epistemic).

7. Note there is thriving school of sometimes called Empirical Bayesianism (Efron, 2010) that deliberately incorporates an objectivist view of probabilities into an otherwise Bayesian framework. My argument here is not primarily aimed at empirical Bayesians, who are aware of both sides of the debate and intentionally adopt a more frequentist view, but rather on cognitive scientists who may have absorbed implicit frequentism without fully appreciating the consequences. Note however that for this reason Empirical Bayesians' procedures are sometimes criticized as "non-Bayesian" (Lindley, 1972; Robert, 2007).
8. From the *New Yorker:* "The fraction of women in their forties who have breast cancer is 0.014, which is about one in seventy. The fraction who do not have breast cancer is therefore $1 - 0.014 = 0.986$. These fractions are known as the prior probabilities" (Marcus & Davis, 2013b).
9. More specifically: typically one would assume a beta(α, β), e.g. beta(1, 1) which is a uniform distribution. If so, after N taxi of which pN were green, one would arrive at a posterior that was beta distributed with parameters $\alpha' = \alpha + .15N$, $\beta' = \beta + .85N$). This in turn is approximately normal with mean $\alpha'/(\alpha' + \beta')$ and variance $\frac{\alpha'\beta'}{(\alpha'+\beta')^2(1+\alpha'+\beta')}$. Assuming a uniform prior ($\alpha = \beta = 1$) and $N = 100$ taxis observed of which 15 were green yields the numbers given in the text. See Lee (2004) or any other standard Bayesian text.
10. A further complexity to this debate is that while Gigerenzer's emphasis on frequency format may suggest a frequentist philosophical position about probability (and is sometimes taken that way; e.g., Cosmides & Tooby, 1995), he himself is not a frequentist (see Vranas, 2000), though perhaps also not a Bayesian (Gigerenzer & Brighton, 2009). Nevertheless, it is reasonable to say that an observer drawing probabilities from data—which is essentially what happens when subjects are given information in frequency format, and need to infer probability—is a fairly canonic example of Bayesian inference.
11. Frequentists might argue that probability refers, not to relative frequency in the *actual* world, but to relative frequency in an imaginary infinite series of repetitions of identical conditions—in the example, the relative frequency of observing a 1 next in an infinite series of replays of the situation in which one has just received a Morse code transmission of the first six words of the Gettysburg Address. To Bayesians, though, this simply confirms the essentially *epistemic* nature of the definition. What we mean by "the same situation" is inescapably defined with respect to the observer's state of

knowledge, conjuring up an imaginary sequence of repetitions in which we have exactly the same information every time, but the outcome varies. This thought experiment contrasts with real-world repetitions, where each case involves a somewhat different combination of observable conditions, any of which could potentially change the probability via conditionalization.

12. Recently, Busemeyer and collaborators have proposed applying models of quantum probability to cognition (e.g., see Busemeyer & Bruza, 2012; Pothos & Busemeyer, 2013), but this remains highly controversial (Vanpaemel & Lee, 2013).
13. The subscript *G* is for *ground* or *God*, the Omniscient Observer. Note though that while in our context God knows the past values of all other variables, He does not know the future.
14. The assignment of probabilities to singular events is a subtle issue. To frequentists, individual events do not have probabilities, since by definition probability is assigned only to outcomes of long runs of events drawn from a class. But to Bayesians, probability can reasonably be assigned to individual events because (as usual) probabilities simply reflect the uncertainty in the mind of the observer, which applies to individual events as well as to any other. Hence, when (as discussed in the text) the event becomes deterministically predictable, this uncertainty goes to zero, and probability becomes either zero or one. Hence, from a Bayesian point of view, in an FND universe one *can* speak of objective probabilities, as long as it is understood that they always have values of zero or one—a notion of probability somewhat removed from the ordinary sense of the term.
15. The situation is very similar to Simpson's paradox, in which adding a new variable to a statistical analysis reverses the direction of the effects of other variables. This is not really a paradox, but simply an unavoidable consequence of the acquisition of new information. It also relates to an ongoing debate among statisticians about whether marginalizing over all covariates is desirable (Rubin, 2009) or undesirable (Pearl, 2009) in order to optimize estimates of probability.
16. "Perception is not about truth, it's about having kids."—Hoffman and Prakash (2014).

References

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2015). How bad is incoherence? *Decision*, *3*(1), 20–39.

Bennett, B. M., Hoffman, D. D., & Murthy, P. (1993). Lebesgue logic for probabilistic reasoning and some applications to perception. *Journal of Mathematical Psychology*, *37*(1), 63–103.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Chichester, UK: John Wiley & Sons.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Burge, J., Fowlkes, C. C., & Banks, M. S. (2010). Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. *The Journal of Neuroscience*, *30*(21), 7269–7280.

Busemeyer, J. M., & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge, UK: Cambridge University Press.

Colombo, M., & Seriès, P. (2012). Bayes in the brain—On Bayesian modelling in neuroscience. *British Journal for the Philosophy Science*, *63*(3), 697–723.

Cosmides, L., & Tooby, J. (1995). Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley.

Cox, R. T. (1961). *The algebra of probable inference*. London: Oxford Universitiy Press.

Davey, B., & Priestley, H. (1990). *Introduction to lattices and order*. Cambridge, UK: Cambridge University Press.

de Finetti, B. (1964). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokier (Eds.), *Studies in subjective probability* (pp. 93–158). New York: Wiley.

de Finetti, B. (1970/1974). *Theory of probability*. Torino: Giulio Einaudi. (Translation 1990 by A. Machi and A. Smith, John Wiley and Sons)

de Finetti, B. (2008). *Philosophical lectures on probability*. New York: Springer. (Collected, edited, and annotated by Alberto Mura)

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

Eddington, A. S. (1928). *The nature of the physical world*. New York: Macmillan.

Efron, B. (2010). *Large scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge, UK: Cambridge University Press.

Feldman, J. (2012). Symbolic representation of probabilistic worlds. *Cognition*, *123*, 61–83.

Feldman, J. (2013). Tuning your priors to the world. *Topics in Cognitive Science*, *5*(1), 13–34.

Feldman, J. (2015). Bayesian inference and "truth": A comment on Hoffman, Singh, and Prakash. *Psychonomic Bulletin and Review*, *22*(6), 1523–1525.

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Frank, M. C. (2013). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). *Cognition*, *128*(3), 417–423.

Geisler, Wilson S. (2003). "Ideal observer analysis". In L. M. Chalupa & J. S. Werner (Eds.), *The Visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London B*, *357*, 419–448.

Geisler, W. S., & Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, *27*, 379–402.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond heuristics and biases. *European Review of Social Psychology*, *2*, 83–115.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). Chichester, UK: Wiley.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin*, *119*(1), 23–26.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis*, *25*(3), 623–629.

Goodman, N. (1954). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P.W., & Hamrick, J. B. (2015). Relevant and robust: A response to Marcus and Davis (2013). *Psychological Science*, *26*(4), 539–541.

Grätzer, G. (1971). *Lattice theory: First concepts and distributive lattices*. San Francisco, CA: W. H. Freeman.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Greene, B. (2011). *The hidden reality: Parallel universes and the deep laws of the cosmos*. New York: Vintage.
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Science*, *14*(8), 357–364.
Hacking, I. (1990). *The taming of chance*. Cambridge, UK: Cambridge University Press.
Hoffman, D. D. (2009). The user-interface theory of perception: Natural selection drives true perception to swift extinction. In S. Dickinson, M. Tarr, A. Leonardis, & B. Schiele (Eds.), *Object categorization: Computer and human vision perspectives* (pp. 148–165). Cambridge, UK: Cambridge University Press.
Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. *Frontiers in Psychology*, *5*, 1–22.
Hoffman, D. D., & Singh, M. (2012). Computational evolutionary perception. *Perception*, *41*, 1073–1091.
Hoffman, D. D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic Bulletin and Review*, *22*(6), 1480–1506.
Howie, D. (2004). *Interpreting probability: Controversies and developments in the early twentieth century*. Cambridge, UK: Cambridge University Press.
Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, *70*(9), 939–952.
Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.
Jeffrey, R. (2004). *Subjective probability: The real thing*. Cambridge, UK: Cambridge University Press.
Jeffreys, H. (1939/1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press.
Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Science*, *34*, 169–188.
Jordan, M. I. (2010). Bayesian nonparametric learning: Expressive priors for intelligent systems. In R. Dechter, H. Geffner, & J. Halpern (Eds.), *Heuristics, probability and causality: A tribute to Judea Pearl* (pp. 167–186). College Publications.
Keynes, J. M. (1921). *Treatise on probability*. London: Macmillan & Co.
Koenderink, J. J. (2011). Vision as a user interface. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human vision and electronic imaging XVI 7865*, (pp. 1–13).
Koenderink, J. J. (2012). *Visual awareness*. Clootrans Press.
Koenderink, J. J. (2015). Gestalts as ecological templates. In J. Wagemans (Ed.), *Handbook of perceptual organization* (pp. 1046–1062). Oxford, UK: Oxford University Press.
Lee, P. (2004). *Bayesian statistics: An introduction* (3rd ed.). Hoboken, NJ: Wiley.
Lee, M. D. (2010). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.
Lindley, D. V. (1972). *Bayesian statistics: A review*. Philadelphia: SIAM.
Lindley, D. V., & Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, *30*(3), 112–119.
Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, *35*(4), 549–568.
Maloney, L. T. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfeld (Eds.), *Perception and the physical world: Psychological and philosophical issues in perception* (pp. 145–189). New York: Wiley.
Marcus, G. F., & Davis, E. (2013a). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*(12), 2351–2360.
Marcus, G. F., & Davis, E. (2013b). What Nate Silver gets wrong. *The New Yorker*.
Mark, J. T., Marion, B. B., & Hoffman, D. D. (2010). Natural selection and veridical perceptions. *Journal of Theoretical Biology*, *266*, 504–515.
Mellor, D. H. (2005). *Probability: A philosophical introduction*. London: Routledge.
Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Science*, *32*, 69–84.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, J. (2009). Myth, confusion, and science in causal analysis. *Tech. Rep. R- 348*, http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf, University of California, Los Angeles

Popper, K. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science*, *10*(37), 25–42.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Science*, *36*(3), 255–327.

Purves, D. (2010). *Brains: How they seem to work*. Saddle River, NJ: FT Press.

Richards, W. A., & Bobick, A. (1988). Playing twenty questions with nature. In Z. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary perspective* (pp. 3–26). Norwood, NJ: Ablex Publishing Corporation.

Richards, W. A., McKay, B. D., & Richards, D. (2002). Probability of collective choice with shared knowledge structures. *Journal of Mathematical Psychology*, *46*, 338–351.

Robert, C. (2007). *The Bayesian choice* (2nd ed.). New York: Springer.

Rubin, D. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine 28*(9), 1420–1423.

Samaniego, F. J. (2010). *A comparison of the bayesian and frequentist approaches to estimation*. New York: Springer.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Shepard, R. N. (1989). Internal representation of universal regularities: A challenge for connectionism. In L. Nadel, L. A. Cooper, P. Culicover, & R. M. Harnish (Eds.), *Neural connections, mental computation* (pp. 104–134). Cambridge, MA: MIT Press.

Singh, M., & Hoffman, D. D. (2013). Natural selection and shape perception. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision: An interdisciplinary perspective* (pp. 171–185). New York: Springer Verlag.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, UK: Cambridge University Press.

Vanpaemel, W., & Lee, M. (2013). Quantum models of cognition as Orwellian newspeak. *Behavioral and Brain Science*, *36*(3), 255–327.

Venn, J. (1888). *The logic of chance: An essay on the foundation and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics*. London: MacMillan.

von Mises, R. (1939). *Probability, statistics and truth*. New York: Macmillan.

von Plato, J. (1994). *Creating modern probability: Its mathematics, physics and philosophy in historical perspective*. Cambridge: Cambridge University Press.

Vranas, P. B. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition*, *76*(3), 179–193.

Zabell, S. L. (2005). *Symmetry and its discontents: Essays on the history of inductive probability (Cambridge studies in probability, induction and decision theory)*. Cambridge, UK: Cambridge University Press.

Appendix: Proof sketches

Given two observers, Σ_1 and Σ_2, it is in general not possible to say, based on the subset relations between them, whether $p_{\Sigma_1}(X) \prec p_{\Sigma_2}(X)$ or $p_{\Sigma_2}(X) \prec p_{\Sigma_1}(X)$.

Proof sketch.: Assume an FND model entailing a ground truth $p_G(X)$, defined by p_{Σ_T}.

Case 1: $\Sigma_1 \subseteq \Sigma_2$. There is in general no way to guarantee that either $p_{\Sigma_1}(X) \prec p_{\Sigma_2}(X)$ or $p_{\Sigma_2}(X) \prec p_{\Sigma_1}(X)$. Fig. 3 provides an example where $\Sigma_1 \subseteq \Sigma_2$, $\Sigma_1 \subseteq \Sigma_3$, yet $p_{\Sigma_1}(X) \prec p_{\Sigma_2}(X)$ while $p_{\Sigma_3}(X) \prec p_{\Sigma_1}(X)$. (In the figure, take Σ_1 to be $\varnothing$, Σ_2 to be $\{A\}$, and Σ_3 to be $\{B\}$). A superset observer can move probability either up or down relative to a subset, without regard to ground truth.

Case 2: $\Sigma_2 \subseteq \Sigma_1$. As in Case 1, by symmetry.

Case 3: $\Sigma_1 \sim \Sigma_2$ (incommensurate observers). For any two observers Σ_1 and Σ_2, there exists a least common superset observer $S = \Sigma_1 \cap \Sigma_2$ such that $\Sigma_1 \subseteq S$ and $\Sigma_2 \subseteq S$ (see text). Consider the chain of precedence as we move from Σ_1 to Σ_2 via S. As per the argument in Case 1, there are four subcases:

(i) $p_{\Sigma_1} \prec p_S, p_{\Sigma_2} \prec p_S$;
(ii) $p_{\Sigma_1} \prec p_S, p_S \prec p_{\Sigma_2}$;
(iii) $p_S \prec p_{\Sigma_1}, p_{\Sigma_2} \prec p_S$;
(iv) $p_S \prec p_{\Sigma_1}, p_S \prec p_{\Sigma_2}$.

Because $\prec$ is transitive, in case (ii) we can infer that $p_{\Sigma_1} \prec p_{\Sigma_2}$ and in case (iii) that $p_{\Sigma_2} \prec p_{\Sigma_1}$. However in cases (i) and (iv), there is no transitive chain and thus either $p_{\Sigma_1} \prec p_{\Sigma_2}$ or $p_{\Sigma_1} \prec p_{\Sigma_2}$ could hold. In other words, in (ii) we go from Σ_1 to Σ_2 by taking two steps up, and in (iii) two steps down; but in (i) and (iv) we take one step up and one step down, thus leaving us in an unknown position relative to the starting point.