Pragmatic Encroachment in Bayesian Epistemology: The Interpretive Problem and Its Solutions

November 10, 2017

Abstract

Scientists and Bayesian statisticians often study hypotheses that they know to be false. This creates an interpretive problem because the Bayesian probability of a hypothesis is typically interpreted as a degree of belief that the hypothesis is true. I present and contrast two solutions to the interpretive problem. Both solutions involve giving a new interpretation of the Bayesian framework and both interpretations posit a complex kind of epistemic attitude that is distinct from either a degree of belief or a full belief. Importantly, according to both solutions, there is significant pragmatic encroachment on epistemic rationality.

1 Introduction

Scientists who use Bayesian methods often assign probabilities over sets of hypotheses that they are certain are false; yet, according to the standard Bayesian interpretation of probability, the probability assigned to a hypothesis is supposed to represent a degree of belief in the truth of the hypothesis. Thus, there is an inconsistency between the standard interpretation of the Bayesian framework and the way the framework is...
often applied—let’s call this inconsistency the “interpretive problem.”¹ I will argue that the only satisfactory solutions to the problem involve reinterpreting what it means to assign a probability to a hypothesis. According to one solution (originally proposed by Sprenger (2017)), probabilities are interpreted as counterfactual degrees of belief; according to a second solution, probabilities are interpreted as what I will refer to as “verisimilitude degrees of belief.”

Although the interpretive problem arises in applied statistics, the problem has important implications for epistemology. In particular, as we will see, both solutions to the interpretive problem posit a complex kind of epistemic attitude that is distinct from either a degree of belief or a full belief. Furthermore, according to both solutions, whether a given epistemic attitude is rational for an agent is in part determined by the goals of the agent, including practical goals. In recent years, there has been much debate over whether there is sometimes “pragmatic encroachment” on the epistemic, i.e. whether pragmatic factors can sometimes influence whether an agent, for instance, knows whether a proposition is true.² As Mark Schroeder (2017) point outs, it seems to be almost universally agreed among participants of this debate that although there may be pragmatic encroachment on knowledge or rational (full) belief, there is no pragmatic encroachment on agents’ rational credences—that is, there is no pragmatic encroachment in Bayesian epistemology. The arguments in this paper rebut this view. Whereas it may be true that there is no pragmatic encroachment on standard degrees of belief, there is significant pragmatic encroachment on both counterfactual and verisimilitude degrees of belief.

Here is the plan of the paper. Section 2 draws an important distinction between the syntax and semantics of Bayesian inference that seems often to be overlooked. Section 3 introduces the interpretive problem through an example and argues that the problem is pervasive in applied statistics. Section 4 discusses ways of trying to

¹This problem has been noted in the past, e.g. by Box (1980), Bernardo and Smith (1994), Forster and Sober (1994), Forster (1995), Key et al. (1999), Shaffer (2001), Sprenger (2009), Gelman and Shalizi (2013), and more recently Sprenger (2017) – indeed, Jan Sprenger (2017) calls the problem the “scandal of Bayesianism”– but in general the seriousness of the issue seems to be under-appreciated.

²See e.g. Stanley (2005), Fantl and McGrath (2002), Ross and Schroeder (2014), Rubin (2015), or Roeber (2016)
avoid or circumvent the interpretive problem and argues that these strategies are unsuccessful: solving the interpretive problem necessarily involves coming up with a new semantics for Bayesianism. Section 5 introduces the verisimilitude interpretation of probability and shows how this interpretation solves the interpretive problem. Section 6 uses an example to show why the rational status of a verisimilitude probability is dependent on pragmatic factors. Section 7 introduces the counterfactual interpretation and shows how this interpretation, too, solves the interpretive problem. Section 8 shows that there is a way of translating between the verisimilitude and counterfactual interpretations and argues that the two interpretations are really two sides of the same coin. Section 9 explores the relationship that the counterfactual and verisimilitude interpretations bear to the standard interpretation. Section 10 introduces the plausible conjecture that all adequate solutions to the interpretive problem will have the same broad features as the verisimilitude and counterfactual solutions, and uses this conjecture in conjunction with the conclusions established in the paper to argue that there is a pervasive and important kind of pragmatic encroachment in epistemology. Section 11 ends the paper with a few concluding thoughts and ideas for future directions.

2 The Semantics and Syntax of Bayesian Inference

To understand the interpretive problem, it is important to recognize that Bayesian inference can be characterized on two distinct levels: the purely formal (“syntactic”) level, and the interpretive (“semantic”) level. On a purely formal level, the basic objects of study in Bayesian statistical inference are statistical models. Given a set of candidate hypotheses indexed by a parameter, $\theta$ in $\Theta$, and given some particular context in which the possible observations or outcomes are $x_1$, $x_2$, etc. – or $X$, for short – a statistical model is a set of conditional probability (density) distributions.

From now on, I will for simplicity simply use “probability” although in practice probability densities are more common.
Given a statistical model or a set of statistical models, Bayesians do inference by following a three-step procedure:

In the first step, a probability is assigned to each $\theta \in \Theta$; these probabilities are supposed to be assigned before looking at the data and are therefore known as “prior” probabilities. If there are multiple candidate statistical models, then all of the models must be assigned prior probabilities as well.

In the second step, data $x$ are collected and the “likelihood” of each hypothesis is calculated. The likelihood of $\theta$ is the probability that $\theta$ assigns to the data, $p(x|\theta)$.

In the third and final step, the posterior probability of each parameter and each statistical model is calculated by combining the prior and the likelihood of each hypothesis using Bayes’s theorem, $p(\theta|x) = p(x|\theta) * p(\theta)/p(x)$.

The preceding three-step description of Bayesian inference is intentionally abstract; it is a completely formal description of how Bayesians apply the probabilistic machinery when they do statistical inference – in what follows, I will refer to the preceding three-step procedure as “standard Bayesian inference.” Of course, on the semantic level, Bayesians also have an interpretation of what it is that they are doing. According to the standard “subjective” interpretation, all the probabilities are the degrees of belief of some agent. Thus, in particular, $p(\theta_i)$ represents an agent’s degree of belief that $\theta_i$ is true. I will refer to this interpretation as the “standard interpretation” or the “degree-of-belief” interpretation. The interpretive problem arises whenever the Bayesian syntax is applied in cases where the standard Bayesian semantics is plainly not applicable. As I argue in the next section, this happens quite often.

---

4The fact that $p(x|\theta)$ is a probability distribution over $X$ means that $p$ satisfies the three probability axioms: (1) $\sum_i p(x_i|\theta) = 1$, where the sum is over all $x_i$ in $X$. (2) $p(x_i|\theta) \geq 0$, for all $x_i$. (3) $p(x_i \lor x_j|\theta) = p(x_i|\theta) + p(x_j|\theta)$, when $x_i$ and $x_j$ are mutually exclusive (conditional on $\theta$).

5There is also an “objective” interpretation of Bayesian probability assignments, and everything I say in this paper also carries over if we instead adopt that interpretation. However, since most Bayesian epistemologists are subjective Bayesians, I will assume the subjective interpretation throughout the paper for ease of presentation.

6Or more precisely, a degree of belief that the hypothesis indexed by $\theta_i$ is true.
3 The Interpretive Problem in Bayesian Inference

Here’s a simple example\footnote{Discussed, for example, by Choi et al. (2016).} of how the interpretive problem arises in scientific practice. Suppose you are interested in the functional relationship between two quantities, \(X\) and \(Y\). For concreteness, suppose \(X\) represents some measurement of a complex system, e.g. the minimal pressure of a tropical storm, and \(Y\) represents some quantity of interest, e.g. the maximal windspeed of the storm. The true functional dependence of \(Y\) on \(X\) is unknown and complex, in part because the functional dependence is mediated by the geometry and rotation of the earth. Nonetheless, it is very common in such cases to restrict attention to classes of simple functional relationships, such as the set of lines, which models the relationship between \(Y\) and \(X\) as follows:

\[
Y = \alpha X + \beta + \epsilon
\]  

Here, \(\epsilon\) represents the (hypothesized) random fluctuation around the line \(Y = \alpha X + \beta\); \(\epsilon\) is generally taken to be a normal distribution with a mean of 0 and standard deviation \(\sigma\). \(\alpha\) and \(\beta\) are the parameters of interest while \(\sigma\) is a nuisance parameter (auxiliary assumption); they all need to be estimated from data.

Ideally, the statistical model should be justified on scientific grounds.\footnote{In the social sciences, the model is often not justified on scientific grounds, but is instead chosen based on convenience and tradition. In those cases, one can be very sure that the resulting model is false.} In the case of the relationship between max windspeed and min pressure, various idealized assumptions (see Knaff and Zehr (2007)) justify the model \(Y = \alpha X^n + \beta + \epsilon\),\footnote{Strictly speaking, the model is of the following form: \(Y = \alpha(1010 - X)^n + \epsilon\). I’ve changed it slightly for the sake of streamlining and simplifying the presentation.} and \(n = 1\) is a reasonable choice if the minimal pressure is in an intermediate range. Note, however, that the fact that the model is based on idealized assumptions (i.e. assumptions that are known to be violated in practice) implies that the model in fact is known to be false.

In order to perform the standard three-step Bayesian inference procedure on the preceding model, each value of \(\alpha\) must be assigned a prior probability, and according
to the standard interpretation each probability assignment represents a degree of belief. However, the fact that the statistical model is known to be false means that every straight line picked out by $\alpha$ is known to be false – none of them describe the true relationship between pressure and windspeed. If you know that all the lines are false and if the probabilities are supposed to represent your degrees of belief, then you should assign the minimal probability – i.e. 0 – to every value of $\alpha$. But this is not what Bayesian statisticians do—and for good reason, since assigning a hypothesis a probability of 0 is tantamount to excluding the hypothesis from any further consideration.

The above example is by no means unrepresentative. Indeed, it’s an instance of a much broader class of problems, namely regression problems. In regression analysis, the goal is to estimate the functional relationship between two or more quantities. Almost invariably, the hypotheses under consideration will be restricted to very simple functional relationships, such as the set of lines, parabolas, exponentials, etc. Most functional relationships in the world cannot realistically be expected to belong to one of these sets of simple functional relationships, and indeed the choice of functional class is usually justified on the basis of highly idealized scientific assumptions, if it is justified at all. Hence, scientists generally face the interpretive problem when they use Bayesian methods to solve regression problems.

Another major area of statistical inference where scientists face the interpretive problem is in Bayesian phylogenetics. Phylogeneticists in both biology and linguistics use trees to represent family relationships between species or between languages. In both cases, the trees investigated omit known relationships and introduce false idealizations. For example, a tree phylogeny for a language family is premised on the (false) idea that languages bifurcate instantaneously and are forever separated thereafter. Yet, even though all phylogenetic trees are clearly false, linguistic phylogeneticists often use Bayesian inference and are often interested in discovering which tree has the highest posterior probability (e.g. Bouckaert et al. (2012)). These prob-

10What exactly does it mean for a given value of $\alpha$ to be “true” or “false”? Well, $\alpha$ indexes a set of hypotheses, namely $Y = \alpha X + \beta + \epsilon$, so to say that $\alpha_0$ is “true” in this case is the same as saying that there exist values of $\beta$ and $\epsilon$ such that the hypothesis $Y = \alpha_0 X + \beta + \epsilon$ is the true functional relationship between $X$ and $Y$. 
abilities cannot comfortably be interpreted as degrees of belief that the trees are literally true, and thus scientists who use Bayesian inference on problems such as this are faced with the interpretive problem.

The widespread practice of assigning non-negative prior probabilities to hypotheses that are obviously false is what leads to the interpretive problem, which may be phrased in the form of a question: what does it mean to assign a model or hypothesis that is known to be false a non-zero probability? To more precisely diagnose the problem, it helps to state the probability axioms with the standard Bayesian interpretation made explicit:

Suppose $H$ is a set of hypotheses $\{H_1, H_2, \ldots, H_n\}$. Then

1S. $p(H) = 1$. Interpretation: you are certain that one of the hypotheses in $H$ is true.

2S. $p(H_i) \geq 0$ for all $H_i \in H$. Interpretation: degrees of belief are non-negative.

3S. $p(\bigvee H_i) = \sum p(H_i)$, when the $H_i$ are mutually exclusive. Interpretation: degrees of belief in a set of hypotheses are additive when it’s impossible for more than one of the hypotheses to be true.

Here we can see that the interpretive problem is really a problem with the standard interpretation of the first probability axiom. That is, for many of the hypothesis sets that scientists study, it will not be the case that they are certain that one of the hypotheses is true. Hence, strictly speaking, many hypothesis sets will not satisfy axiom 1S. Axioms 2S and 3S, on the other hand, will generally be satisfied by the kinds of hypothesis sets that Bayesian statisticians study.

4 Solutions to the Interpretive Problem

One response to the interpretive problem that initially strikes many philosophers as attractive is to try to change the algebra over which the probability function $p$ ranges. For example, some might be tempted to consider the algebra generated by
the associated propositions, \(<\theta_i \text{ is the best hypothesis}>\), for each \(\theta_i\), or something similar. The idea is that even though your degree of belief in \(\theta_i\) is 0, your degree of belief in a proposition such as \(<\theta_i \text{ is the best hypothesis}>\) can still reasonably be non-zero.

However, this proposal faces a fundamental difficulty. The problem is that the parameter \(\theta\) indexes a set of probability distributions in a statistical model and therefore entails probabilities for the various possible observations, but an expression such as \(<\theta_i \text{ is the best hypothesis}>\) does not. For example, in the example in Section 3, \(\alpha = 1\) picks out a particular class of lines that make testable predictions about the possible observations; but a proposition such as \(<\alpha = 1 \text{ is the best hypothesis}>\) does not make any testable predictions, and therefore cannot sensibly be incorporated in a statistical model.

To see the problem from a different perspective, consider Bayes’s formula:

\[
p(\theta|x) = \frac{p(x|\theta) \times p(\theta)}{p(x)}
\] (4.1)

Clearly, the likelihood and the prior have to range over the same set of hypotheses in order for Bayes’s formula to be applicable. If we change the algebra of hypotheses so that we instead assign probabilities to propositions of the form \(<\theta_i \text{ is the best hypothesis}>\), then we might avoid the interpretive problem of how to make sense of the prior and posterior probability of \(\theta\). However, now the likelihoods will be of the form \(p(x|<\theta_i \text{ is the best hypothesis}>)\), and it is hard to see how we are to evaluate these likelihoods. Again, \(\theta_i\) indexes a particular statistical hypothesis, and therefore entails probabilities for each possible value of \(x\)—i.e., \(\theta_i\) is a testable hypothesis. However, \(<\theta_i \text{ is the best hypothesis}>\) does not entail any probabilities for \(x\).\(^{11}\)

Of course, a radical subjectivist Bayesian might simply say that scientists are free to associate any probability value that they wish with \(p(x|<\theta_i \text{ is the best hypothesis}>)\). However, actual scientists are not radical subjectivist Bayesians, nor should they be. Again, \(\theta_i\) picks out a hypothesis in a (hopefully) scientifi-\(^{11}\)The problem here is similar to the problem that Bayesians have with assigning likelihoods to “catchall” hypotheses (Sober, 2009).
cally respectable and realistic model of the relevant phenomena, but \(<\theta_i \text{ is the best hypothesis}>\) is not a part of any scientific model, nor could it be. Going the radical subjectivist route would defeat the point of developing scientifically realistic models of phenomena.

One proposal that might seem attractive is to stipulate that, for each \(\theta_i\), \(p(x|\theta_i) = p(x|<\theta_i \text{ is the best hypothesis}>\)\). Let’s call this stipulation ‘(*)’. At least according to (*), you are not associating a completely arbitrary number with \(p(x|<\theta_i \text{ is the best hypothesis}>\)\). However, the proposal faces other difficulties. Suppose you are certain that \(\theta_1\) is the best hypothesis, so that \(p(<\theta_1 \text{ is the best hypothesis}>)=1\). Of course, the fact that \(\theta_1\) is best does not mean it’s any good. Suppose you know that \(\theta_1\) in fact is very bad; more precisely, suppose you know that \(\theta_1\) is extremely predictively inaccurate. If you accept (*), then, by the law of total probability, it follows that you have to also accept that \(p(x) = \sum_i p(x|<\theta_i \text{ is the best hypothesis}>\)\) \(\ast p(<\theta_i \text{ is the best hypothesis}>\)\) = \(\sum_i p(x|\theta_i) \ast p(<\theta_i \text{ is the best hypothesis}>\)\) = \(p(x|\theta_1)\). However, this means that you are forced to calibrate your actual degree of belief that \(x\) will happen to the prediction made by \(\theta_1\), even though you know that \(\theta_1\) is very predictively inaccurate! That is clearly irrational. In other words, if you accept (*), then you will be forced to have degrees of beliefs you yourself take to be irrational.\(^{12}\)

For all of the above reasons, avoiding the interpretive problem by changing the algebra over which \(p\) ranges is not a scientifically respectable or workable solution to the interpretive problem. Other ways of avoiding the interpretive problem also fail to deliver. For example, Morey et al. (2013) assert that “...scientific models, including statistical models, are neither true nor false” (p. 71). They then recommend assigning odds rather than probabilities to models because a “Bayesian who employs odds is silent on whether or not she is in possession of the true model, and, in fact, need not acknowledge the existence of a true model at all” (p. 71). It is, however, unclear how using odds rather than probabilities is supposed to avoid the interpretive problem. And it is not clear how refusing to assign truth values to models avoids the problem either. What does it mean to say that your odds are 5 to 1 in a model?

\(^{12}\)By a similar line of reasoning, you will also be forced to accept bets that you do not think are good, provided you use your degrees of belief as your betting ratios.
that is neither true nor false as against another model that is also neither true nor false? The interpretive problem seems to be just as severe here as before.

If we are to face interpretive problem head on, then we have to face up to the fact that it really is an interpretive problem—the problem is that the standard Bayesian interpretation of probability does not fit with how the Bayesian machinery is often applied in practice. To solve the problem, it follows that we will have to come up with a different interpretation of the Bayesian framework. The degree of belief interpretation will have to be discarded, at least in the cases where we face the interpretive problem. For the remainder of the paper, I will consider two solutions to the interpretive problem. One solution involves interpreting probabilities as counterfactual degrees of belief rather than standard degrees of belief, while the other interpretation involves interpreting probabilities as what I will refer to as a “verisimilitude degree of belief.” As we will see, both interpretations have the consequence that there is pragmatic encroachment on whether a probability distribution is rational. Of course, there may be other reinterpretations of the Bayesian framework that would also solve the interpretive problem, but my conjecture will be that all such reinterpretations would have the same broad features as the interpretations that I will discuss have.

5 The Verisimilitude Interpretation of Probability

In cases where all the hypotheses under consideration are known to be false, the goal of Bayesian inference cannot reasonably be construed as discovering the hypothesis that most probably is true. A natural proposal is that the goal in such cases changes to discovering which hypothesis is—in some sense—closest to the truth. Indeed, scientific realists have long held that the real (achievable) goal of inference is closeness to the truth rather than truth itself.

The idea that the goal of inference is to identify the \( \theta \) that is closest to the truth leads to a natural reinterpretation of probability. Instead of interpreting \( p(\theta) \) as a degree of belief in \( \theta \), we interpret \( p(\theta) \) as a degree of belief that \( \theta \) is closest to the truth out of the hypotheses in \( \Theta \). I will call this interpretation of probability the “verisimilitude interpretation.”
The reader may wonder how this differs from the earlier rejected suggestion of changing the algebra of hypotheses. Does the verisimilitude interpretation not just say that we ought to assign probabilities to propositions of the form $<\theta$ is closest to the truth> rather than to $\theta$ itself? The answer is “no.” According to the verisimilitude interpretation, $p(\theta)$ is a probability that is assigned to $\theta$ itself, not to $<\theta$ is closest to the truth>. Thus, according to the verisimilitude interpretation:

$$p(\theta) = \text{a degree of belief that } \theta \text{ is closest to the truth out of the hypotheses in } \Theta.$$  

In other words, according to the the verisimilitude interpretation, a probability assignment to $\theta$ represents a complex epistemic attitude taken towards $\theta$; it does not represent a simple attitude taken towards a complex proposition.\(^{13}\) This is important, because as we saw earlier, $\theta$ is a parameter in a statistical (and ideally scientifically respectable) model, whereas $<\theta$ is closest to the truth> is not.\(^{14}\)

Of course, we could define an associated probability distribution $q$ ranging over the set of associated propositions of the form $<\theta_i$ is closest to the truth>, and stipulate that, for each $\theta_i$, $p(\theta_i) = q(<\theta_i$ is closest to the truth>). But the fact that we can define such a probability function $q$ does not imply that we do not also need the original probability function $p$. We still need $p$, since only $p$ ranges over hypotheses that are part of a statistical model.

Similarly, standard subjective Bayesians posit that there are such things as subjective degrees of beliefs. If your degree of belief in $\theta_i$ is $p(\theta_i) = d$, then – Bayesians say – you have a certain epistemic attitude towards $\theta_i$, and this epistemic attitudes is associated with the real number $d$. Note, however, that just as with the verisimilitude interpretation, it is possible to define an associated set of propositions of the form $<\theta_i$ is true with probability $d>$ and a full belief function $B$ such that $B(m) = 1$ iff the agent has a full belief in $m$ and $B(m) = 0$ otherwise. Next, we may stipulate that, for each $\theta_i$, $p(\theta) = d$ iff $B(<\theta_i$ is true with probability $d>) = 1$. Because this is

\(^{13}\)Cf. the point made by Moss (2018), although the the lesson drawn here is different.

\(^{14}\)Also, on the verisimilitude interpretation of probability, $p(x) = \sum_i p(x|\theta_i)p(\theta_i)$ no longer represents your actual degree of belief that $x$ will happen.
possible, some may be tempted to claim that degrees of belief are dispensable, since
instead of representing agents as having degrees of belief, we can instead represent
them as having full beliefs in propositions that have probabilistic content. However,
this is a mistake, because \( p \) plays a functional role that \( B \) is not able to play: in
particular, only \( p \) can enter into Bayesian calculations that allow the agent to update
on evidence. Thus, even though it is possible to define \( B \), \( p \) is still indispensable.

What I am arguing is that considerations very similar to the considerations that
justify the positing of standard degrees of beliefs also justify the positing of verisimil-
itude degrees of beliefs. In particular, Bayesian statistical inference requires the use
of a probability distribution over the hypotheses under consideration, but the inter-
pretive problem prevents this probability distribution from having its standard
semantics. According to the verisimilitude interpretation, the probability distribu-
tion instead represents a complex kind of epistemic attitude, distinct from either full
beliefs or degrees of beliefs, namely verisimilitude degrees of beliefs. The fact that \( p \),
on the verisimilitude interpretation, is construed as representing a complex epistemic
attitude is what opens the door to pragmatic encroachment. But before we can see
why that is, a bit more needs to be said about verisimilitude.

### 5.1 Solving the Interpretive problem with the Verisimilitude
Interpretation

The study of verisimilitude was initiated by Popper (1963) and has by now accumu-
lated a large literature.\(^{15}\) The most influential contemporary approach in the study
of verisimilitude – known in the literature as the “similarity approach” – understands
verisimilitude as a particular kind of approximation. To say that something is a good
approximation of something else is to say that the two things are similar in some
relevant respect. Thus, to say that a hypothesis is close to the truth is to say that
the hypothesis is similar to the true hypothesis.

This idea can be formalized if we suppose that there is a (context-appropriate\(^{16}\))

\(^{15}\) See Niiniluoto (1998) for a survey.
\(^{16}\) In general I agree with Northcott (2013) that there is little reason to assume a priori that there
verisimilitude measure, \( v \), that ranks hypotheses by how similar they are to the true hypotheses. If we presume that such functions are available, we can say that \( \theta_1 \) is closer to the truth than \( \theta_2 \) if and only if \( v(\theta_1) > v(\theta_2) \).

As a concrete example, a probabilistic divergence measure that has been suggested as a verisimilitude measure in a statistical context is the Kullback-Leibler divergence (Forster and Sober, 1994). Supposing that \( q \) is the “true” probability distribution that governs the distribution of the data, then the verisimilitude (according to the K-L divergence) of some hypothesis \( \theta \) (that does not contain adjustable parameters) is \( KL(\theta) = - \int q(x) \log \frac{q(x)}{p(x|\theta)} \, dx \). Later in the paper I will suggest a simple verisimilitude measure that makes sense in the earlier example concerning the relationship between windspeed and pressure.

Given a measure of verisimilitude, \( v \), I will use \( p_v \) with a \( v \) subscript to indicate that the intended interpretation of \( p_v \) is the verisimilitude interpretation with measure \( v \). Recall that it was the first axiom that was the source of the interpretive problem. Here is the first axiom with the verisimilitude interpretation made explicit:

1V. \( p_v(H) = 1 \). Interpretation: You are certain that one of the hypotheses in \( H \) is a maximum of \( v \).

It is clear that the verisimilitude interpretation avoids the interpretive problem because 1V is easily satisfied and, moreover, will almost always be satisfied in practice. First, it is reasonable to demand that \( v \) be continuous; secondly, in applied Bayesian inference it is almost universally assumed that the hypothesis space is compact.\(^{17}\) If both the preceding conditions are satisfied, then \( v \) is mathematically guaranteed to have a maximum over the hypothesis space. Hence you will generally be (justifiably) certain that one of the hypotheses is maximally close to the truth.

Note that, on the verisimilitude interpretation, the probability assigned to a hypothesis is relative to a given way of measuring verisimilitude. The verisimilitude probability of a hypothesis is therefore not an absolute number; in particular, as

\(^{17}\)“compactness” is a mathematical concept that will not be defined here. The interested reader may consult any textbook on point set topology.
we will see in the next section, it can be rationally influenced by pragmatic factors such as why you are interested in the hypothesis. This is in sharp contrast to the standard degree-of-belief interpretation, according to which the probability assigned to a hypothesis represents your confidence in the hypothesis, which is standardly taken not to vary with pragmatic factors.

6 Goal-relativity of Probability Assignments under the Verisimilitude Interpretation of Probability

The purpose of the present section is two-fold. On the one hand, I want to show, through an example, how the verisimilitude interpretation may actually be useful. More precisely, the goal is to show how it’s possible to combine background information with a verisimilitude measure in a principled manner in order to arrive at rational constraints on the verisimilitude degree of belief function. Thus, verisimilitude degrees of belief can play the same role in inference – and have the same advantages – as standard degrees of belief have (whenever standard degrees of belief are applicable). On the other hand, the example will also serve to show how pragmatic factors influence what the rational constraints on the prior probability function turn out to be.

In order to get a sense of how this will work, it is helpful to first look at a simple example of how your background knowledge can lead to rational constraints on your prior distribution in the standard case where there is no interpretive problem. Suppose you are estimating the mass of a small cup of water, and suppose you model the outcome of your measurement as a likelihood function $p(x|m)$, where $x$ is the outcome of your measurement and $m$ is a possible value of the cup’s mass. In order to do calculate the posterior probability of $m$ given $x$, you need to assign a prior probability distribution over the possible values of $m$. Obviously, $m$ cannot take on any negative values (the mass of an object cannot be a negative number). Furthermore, you can be certain that a small cup of water will not weigh more
than, say, 1kg. This background knowledge straightforwardly leads to a rational constraint on the prior distribution that you ought to use, namely that any such prior distribution should assign a probability (density) of 0 to any value of \( m \) less than 0 or greater than 1. Thus, in the standard case – i.e. the case in which you know one of the hypotheses under consideration is true – background information straightforwardly leads to rational constraints on the prior. When we are faced with the interpretive problem, things are somewhat more complicated.

Consider again the example concerning the relationship between barometric pressure (\( X \)) and maximum windspeed (\( Y \)). Let’s use \( f \) to denote the true (unknown) functional dependency of \( Y \) on \( X \). Now, suppose one of the things you know about the relationship between barometric pressure and windspeed is that changes in maximum windspeed are relatively insensitive to changes in barometric pressure.

So far, this is background knowledge about the actual, unknown function relating barometric pressure and windspeed. What consequences does this background knowledge about \( f \) have for inferences about the hypothesis set actually under consideration? Suppose, as before, that the hypothesis set you are considering is the set of lines and that you know that \( f \) is not in this set. Can you use your background knowledge about \( f \) to discriminate between the various false lines in a principled way? The answer is yes, but how your background knowledge affects the inferences you are entitled to make will depend on how you measure verisimilitude.

Suppose that your ultimate goal is to build a structure that will be able to withstand strong winds.\(^{18}\) In that case, it is very important that the maximal error you make when you estimate windspeed is as small as possible. In other words, Figure 1 is a natural measure of closeness to the truth given your goal; this is not to say that this is an appropriate way to measure closeness to the truth given other goals.

Mathematically, the verisimilitude of some straight line \( L \) is given by the formula

\[
v(L) = -Max|f-L|,\]

where the maximum is taken over a range of relevant possible

\(^{18}\)I thank A for suggesting this example to me.

\(^{19}\)The minus sign is there to make sure that higher verisimilitude scores correspond to a smaller maximum difference between the line and true function.
pressures. Given that you use \( v \) to measure verisimilitude, and given that you have restricted the analysis to the class of lines, your more immediate goal is to identify lines that are close to the truth according to \( v \). It is in fact easy to show that, under the given conditions, some (identifiable) lines will be further from the truth than others, given the way you measure verisimilitude and given your background knowledge—in particular, lines that have a very steep slope.\(^{20}\) Such lines should therefore, rationally, be assigned a prior probability (density) of 0 since they do not have any chance of being closest to the truth.

A prior distribution \( p \) that assigns probability 0 to all lines that cannot possibly be closest to the truth will be more rational than – say – a flat prior that assigns the same probability to every logically possible line in the sense that a Bayesian inference that uses \( p \) as the prior can be expected to more quickly converge on the line that is closest to the truth in the relevant sense. This can be expected because \( p \) excludes from consideration hypotheses that cannot possibly be closest to the truth, whereas

\(^{20}\)To keep the paper relatively non-technical, I have omitted the proof, which requires a few concepts from calculus. However, I hope it’s intuitive why it has to be true: a line that has a very steep slope will almost certainly have a large maximal distance from the true curve, and will therefore not be plausibly closest to the truth in the relevant sense.
the flat prior excludes no hypotheses from consideration. The role played by \( p \) is therefore exactly analogous to the role played by a prior over the possible values of the mass of a cup of water that assigns, say, 0 to all negative values.

However, crucially, if you measure closeness to the truth differently, you do not necessarily get the same rational requirement on the prior distribution. Suppose, for example, that you are instead very concerned with the minimal rather than maximal distance of each line from the truth. That is, you use \( w(L) = \text{Min}|f - L| \) to measure the verisimilitude of each line (see Figure 2).

According to \( w \), any line that intersects \( f \) will be maximally close to the truth, and so your goal is now to identify the lines that intersect \( f \). Clearly, lines that have a very steep slope will stand a better chance of intersecting \( f \) than lines that do not, and thus if you use \( w \) to measure verisimilitude, then it is rational for you to use a prior distribution that assigns more probability to lines that have a steep slope than to lines that have a non-steep slope; this is opposite of the result you get when you use the verisimilitude measure in Figure 1.

In general, how background knowledge interacts with a given measure of verisimilitude measure in order to induce rational requirements on the prior distribution is a
subtle and complex question. My goal in this section is not, however, to demonstrate in full generality how to best translate background information into reasonable requirements on prior distributions over sets of known false hypotheses. My goal is rather to show how, in principle, background knowledge can be used to discriminate between multiple false hypotheses. As we have seen, the way verisimilitude is measured plays a crucial role in shaping the rational constraints on the prior; moreover, we have also seen that the way verisimilitude ought to be measured is reasonably influenced by the goal that we have. Hence, the example in this section shows how there is pragmatic encroachment on Bayesian inference, given that we use the verisimilitude interpretation of probability.

7 The Counterfactual Interpretation of Probability

The verisimilitude interpretation has the feature that the prior probability distribution incorporates not just your background information, but also what you hope to accomplish, formalized by way of a verisimilitude measure. The result is that an agent’s verisimilitude degree of belief represents a complex epistemic attitude that can be influenced rationally by the agent’s goal. In a very recent paper, Jan Sprenger (2017) proposes an alternative solution the interpretive problem. Sprenger’s solution also involves reinterpreting the probability axioms, but he offers a reinterpretation that appears to be quite different from the verisimilitude interpretation. However, as we will soon see, given certain plausible assumptions, the verisimilitude solution and Sprenger’s solution share many features in common and are even formally intertranslatable.

Sprenger’s suggestion is that the probability of a false hypothesis can sensibly be interpreted as a counterfactual degree of belief. More precisely, suppose $\mathbf{H}$ is a set of hypotheses, all of which are known to be false. Then any probability assigned to some particular $H$ should be construed as a degree of belief in $H$ that is conditional on the (false) supposition that one of the hypotheses in $\mathbf{H}$ is true. In other words, the
probability of \( H \) is really the counterfactual conditional probability \( p(H|H) \), where the condition \( H \) is construed as the (false) claim that one of the hypotheses in \( H \) is true.

Note that \( p(H|H) \) cannot simply be replaced with \( p(H \rightarrow H) \), i.e. with a probability distribution defined over counterfactual propositions—the discussion on p. 11 applies equally here. \( H \) is a hypothesis in a scientific and statistical model that makes testable probabilistic predictions, but \( H \rightarrow H \) is not. As before, we could of course define an associated probability distribution \( q \) that ranges over all the counterfactuals of the form \( H \rightarrow H \), and simply stipulate that, for each \( i \), \( p(H|H) = q(H \rightarrow H) \), but this would not change the fact that \( p \) and not \( q \) is the distribution that we need to use if we are to do a Bayesian inference with the statistical model that ranges over the hypotheses in \( H \).

Let’s use the notation \( p_c \) to signify that the intended interpretation of \( p_c \) is the counterfactual interpretation. Recall that it was the first probability axiom that was the cause of the interpretive problem. On the counterfactual interpretation of probability, the first probability axiom should be written in something like the following way:

\[
1C. \quad p_c(H|H) = 1. \text{ Interpretation: You are certain that one of the hypotheses in } H \text{ would be true if it were the case that one of the hypotheses in } H \text{ were true.}
\]

It may look like \( 1C \) has to be obviously true. However, this impression is deceptive. How are we supposed to understand a claim such as that “you are certain that one of the hypotheses would be true”? Something more substantive needs to be said about how we are supposed to understand and evaluate counterfactual degrees of belief in order for \( 1C \) to really have content. Unfortunately, Sprenger does not offer us any guidance. However, a natural thought is that counterfactual probabilities should be evaluated in a way that is analogous to the way counterfactual conditionals are evaluated. According to (a simplified version of) the standard analysis of counterfactuals due to Lewis (1973), evaluating a counterfactual such as “If \( A \) were the case, then \( B \) would be the case,” involves considering the closest possible world
in which $A$ is true, and then checking whether $B$ is true in that world.\textsuperscript{21} Crucially, Lewis’s analysis depends on a ranking of possible worlds, where worlds are ranked by how similar they are to the actual world.

Presumably counterfactual probabilities should be assessed in a similar manner. It is not hard to imagine very strange and fanciful possible worlds in which pressure and windspeed are linearly related, but presumably most of those possible worlds are not interesting or relevant. As is the case in the counterfactual analysis of conditionals, it is presumably the closest possible worlds that are the interesting ones. But which possible worlds are those? To answer this question, you need to be able to rank worlds in terms of their closeness or similarity to the actual world.

If we do have available a similarity ranking over possible worlds, $s$, then there are two different ways in which we can make more precise what a counterfactual degree of belief in $H$ is.\textsuperscript{22} The first proposal is that $p(H|H)$ should be understood as the actual degree of belief you would have in $H$ if you were in the closest possible world in which $H$ is true. According to the second proposal, $p(H|H)$ should be understood as your degree of belief that $H$ is true in the closest possible world in which $H$ is true. Of these two proposals, the latter is arguably more appealing than the former. For note that, on the first reading, the first probability axiom with the counterfactual interpretation made explicit reads as follows:

\begin{equation}
1C. \quad p_c(H|H) = 1. \quad \text{Interpretation: In the closest possible worlds in which } H \text{ is true, you are certain that one of the hypotheses in } H \text{ is true.}
\end{equation}

This reading does not look right; why should you necessarily be certain that one of the hypotheses in $H$ is true just because you happen to be in a world in which it is the case that one of the hypotheses in $H$ is true? On the second proposal, the first probability axiom instead reads as follows:

\begin{equation}
1C. \quad p_c(H|H) = 1. \quad \text{Interpretation: In the closest possible worlds in which } H \text{ is true, you are certain that one of the hypotheses in } H \text{ is true.}
\end{equation}

\textsuperscript{21}To simplify the presentation, my discussion in this and the next section will assume the so-called “Uniqueness Assumption” according to which, for every $A$, there is a unique closest possible world in which $A$ is true. This is a strong and implausible assumption. However, nothing in the discussion will hinge on whether this assumption is true.

\textsuperscript{22}I thank B for pressing me on this point.
1C. $p_c(H) = 1$. Interpretation: You are certain that in the closest possible worlds in which $H$ is true, one of the hypotheses in $H$ is true.

This reading of 1C looks more plausible. Indeed, 1C now looks like it’s positing that you are certain that a tautology is true, but in fact that’s not quite the case. 1C assumes that there exists a set of closest possible worlds, according to the possible worlds ranking. But if the ranking of possible worlds is pathological or if the set of possible worlds is pathological, then this may not be the case. As in the case of the verisimilitude interpretation, if we assume that the ranking over possible worlds is continuous and that the set of possible worlds is compact, then we will be assured that there is a set of closest possible worlds. Since this assumption seems reasonable, 1C is plausibly true. Hence, the counterfactual interpretation, like the verisimilitude interpretation, solves the interpretive problem.

Note that it’s now clear that the counterfactual interpretation has the same broad features as the verisimilitude interpretation. In particular, on the counterfactual interpretation understood in the above Lewisian way, every probability assignment becomes relative to the way you measure similarity between worlds. Moreover, there are many ways of measuring similarity between worlds, but the way in which you ought to measure similarity between worlds is presumably relative to the features of the world that you care about, and what you care about is partly determined by the goals that you have. Indeed, in the next section we will see that the counterfactual and verisimilitude frameworks are plausibly inter-translatable, so that if verisimilitude probabilities are goal-relative, then so are counterfactual probabilities. Thus, the counterfactual interpretation, like the verisimilitude interpretation, makes the rationality of agents’ probability distributions goal-relative—there is exactly the same kind of pragmatic encroachment on Bayesian inference according to both solutions.
8 Relationship Between the Verisimilitude and Counterfactual Interpretations

At this point, we apparently have two viable reinterpretations of the Bayesian framework, both of which solve the interpretive problem. Many philosophers will be tempted to ask which of the two solutions is the better one. My contention is that neither solution is better than the other, and that in fact there is a sense in which the two solutions are two sides of the same coin.

Indeed, note that, in general, any similarity ranking of possible worlds straightforwardly induces a natural verisimilitude ranking of hypotheses, and vice versa. More precisely, suppose we are given a similarity ranking on worlds \( w_\alpha \geq w_1 \geq w_2 \geq \ldots \), where \( w_\alpha \) is the actual world. Then we can define a verisimilitude ranking on hypotheses as follows: suppose \( w \) is the closest world in which \( H \) is true and \( w' \) is the closest world in which \( H' \) is true, then \( v(H) \geq v(H') \) if and only if \( w \geq w' \).

Conversely, any verisimilitude ranking induces an ordering over possible worlds. Suppose \( v(H_0) > v(H_1) > v(H_2) > \ldots \) is a verisimilitude ranking of hypotheses, and for any hypothesis \( H \), let \( S_H \) denote the set of worlds in which \( H \) is true. Then we can define an ordering of possible worlds in the following way: suppose \( H \) is the hypothesis with the highest verisimilitude such that \( w \in S_H \) and suppose \( H' \) is the hypothesis with the highest verisimilitude such that \( w' \in S'_{H'} \), then \( w \geq w' \) if and only if \( v(H) \geq v(H') \).

According to the verisimilitude interpretation, agents have to evaluate which hypothesis is plausibly closest to the truth out of the hypotheses under consideration. According to the counterfactual interpretation, agents must instead evaluate which hypothesis is plausibly true in the closest possible world in which one of the hypotheses under consideration is true—in other words, they must evaluate what the closest possible world is plausibly like. Since any verisimilitude ranking may be translated into a ranking of worlds, and vice versa, it’s now clear that these two tasks are really one and the same.

\(^{23}\)Hilpinen (1976) uses a similar approach to define a specific verisimilitude measure.
None of the above should really be that surprising since a very similar fact is true of standard Bayesianism. There is a well known duality between propositions and possible worlds: a proposition is often construed as a set of possible worlds, and a possible world is often construed as a conjunction of propositions. Hence, an agent who has a degree of belief in a certain proposition may be regarded as implicitly having a degree of belief that the actual world is in a certain set of possible worlds, and vice versa. The correspondence between verisimilitude rankings and possible worlds rankings shown in this section arguably demonstrates that the same is true of counterfactual and verisimilitude degrees of belief: any counterfactual degree of belief may be regarded as an implicit verisimilitude degree of belief, and vice versa.

Thus, although they may appear different, the verisimilitude interpretation and the counterfactual interpretation of probability are, in a sense, two sides of the same coin. At the very least they are straightforwardly inter-translatable. This means that if there is pragmatic encroachment in the verisimilitude framework, there will also be pragmatic encroachment in the counterfactual framework. In particular, if the reader agrees that the example in Section 6 plausibly shows that verisimilitude rankings are sometimes goal-relative, then the same example will also show that rankings of worlds are sometimes goal-relative, since the verisimilitude ranking may simply be translated into a ranking of possible worlds using the recipe provided in this section. It follows that the rational status of counterfactual degrees of belief—like the rational status of verisimilitude degrees of belief—is sometimes goal-relative.

9 Relationship Between the Verisimilitude, Counterfactual, and Standard Interpretations

The preceding section investigated how the counterfactual and verisimilitude interpretations of probability relate to each other—how do either of these interpretations relate to the standard interpretation? According to the standard interpretation, \( p(H) \) is your degree of belief in \( H \). This interpretation fails when all the hypotheses under consideration are known to be false. I have suggested that in those cases we need to
move to either the verisimilitude interpretation or the counterfactual interpretation. Ideally, the verisimilitude and counterfactual interpretations should both be generalizations of the standard interpretation, so that both are extensionally equivalent to the standard interpretation in cases where the standard interpretation is applicable; i.e. in cases where we are certain that one of the hypotheses under consideration is true, so that 1S is satisfied. Is that the case?  

The answer is that it depends on characteristics of the verisimilitude and counterfactual ranking measures. Let’s first consider the verisimilitude interpretation. Suppose one of the hypotheses under consideration is true, and let $p(H)$ be your degree of belief that $H$ is true. Let’s call the true – but unknown – hypothesis $t$. Suppose $v$ is such that it has a unique maximum, and that the unique maximum is $t$. Then, according to the verisimilitude interpretation, $p_v(H)$ is your degree of belief that $H$ is a maximum of $v$, which, under the conditions specified, means that $p_v(H)$ is your degree of belief that $H$ is $t$; in other words, $p_v(H)$ is your degree of belief that $H$ is true, so we have $p_v(H) = p(H)$. Hence, the verisimilitude interpretation is extensionally equivalent to the standard interpretation under the specified conditions in the sense that the the verisimilitude and standard probability distributions assign the same probabilities to all hypotheses. However, if $v$ has several maxima or if the truth is not among the maxima of $v$, then clearly $p_v(H)$ will not necessarily equal $p(H)$. Hence, the verisimilitude interpretation is extensionally equivalent to the standard interpretation just in case the following conditions are all met: (1) one of the hypotheses under consideration is true, (2) $v$ has a unique maximum, (3) the truth is a maximum of $v$.  

Now let’s consider the counterfactual interpretation of probability. Suppose the similarity ranking over possible worlds satisfies the following conditions: (1) there is a unique world that is closest to the actual world, (2) the actual world is closest to itself. Then, by essentially the same reasoning as above, it follows that we will have $p_c(H) = p(H)$. Hence, the counterfactual interpretation is extensionally equivalent to the standard interpretation just in case one of the hypotheses under consideration is true and the similarity ranking over possible worlds satisfies the constraint known

\footnote{I thank C for pressing me on this issue.}
in the counterfactuals literature as *strong centering*. If the similarity ranking does not satisfy strong centering, then the counterfactual interpretation and the standard interpretation will not be equivalent in cases where both are applicable.

## 10 Pragmatic Encroachment in Epistemology

Here is the argument so far. Bayesian inference faces a fundamental and pervasive problem, namely the interpretive problem. The only adequate solutions to the interpretive problem involve reinterpreting the Bayesian framework—we can keep the Bayesian syntactic framework, but only by replacing the semantics. Both of the solutions to the interpretive problem that I have considered have the upshot that we need to posit a new kind of complex epistemic attitude. Moreover, both the counterfactual and verisimilitude interpretation have the following two important features: (1) they both depend on a ranking over some sort of object (either hypotheses or possible worlds), (2) the ranking that it is rational for an agent to have is influenced by pragmatic factors, such as what the agent’s goals are. The upshot is that whether a given probability assignment (i.e. verisimilitude or counterfactual degree of belief) is rational is influenced by pragmatic factors.

Of course, the standard Bayesian interpretation also allows for pragmatic factors to play a role. According to standard Bayesian decision theory, agents have both a probability function and a utility function; any pragmatic factor – such as what the agent wants or is interested in – is relegated to the utility function. This neat separation between the purely epistemic and the pragmatic fails in cases where we face the interpretive problem. In those cases, pragmatic factors directly influence the agent’s probability function, not just the utility function.

It is conceivable that there may be solutions to the interpretive problem that do not have features (1) or (2), but I do not know of any such solutions (not for a lack of trying).\(^{25}\) My conjecture is that all adequate solutions to the interpretive problem will have features (1) and (2). If this conjecture is correct, it follows that the prag-

\(^{25}\)In conversation, D suggested to me a possibly different possible solution, which upon further analysis turned out also to have features (1) and (2).
matic encroachment that we have seen for both the verisimilitude and counterfactual interpretations is inevitable.

On the basis of the preceding conjecture and the various conclusions established so far in the paper, the following argument may be formulated:

P1: The interpretive problem is a pervasive problem in Bayesian applied inference.

P2: All satisfactory solutions to the interpretive problem involve reinterpreting what it means to assign a probability to a hypothesis.

P3: All satisfactory reinterpretations that solve the interpretive problem will have the following two features: (1) it will depend on a ranking over some sort of object, (2) whether a given ranking is rational will in part be determined by pragmatic factors.

P4: Bayesianism is an important framework in epistemology.

P5: If P1, P2, P3, and P4 are all true, then there is a pervasive and important kind of pragmatic encroachment on epistemology.

C: There is a pervasive and important kind of pragmatic encroachment on epistemology.

The support for P1 is given in Section 3 and Section 4. The support for P2 is given in section 5. P3 is the conjecture raised in this section. Finally, I take P4 to be obvious, and P5 is a bridge premise that also seems very plausible. The conclusion is that there is an important kind of pragmatic encroachment on epistemology. Even if the conjecture I have raised in this section is mistaken, the narrower upshot of the paper is still significant: both the counterfactual and verisimilitude interpretations of probability are intuitive and natural, and so it is significant that they both imply that there is pragmatic encroachment on epistemic rationality.
11 Concluding Thoughts

This paper has mainly been concerned with the implications of the interpretive problem for epistemology—in particular, I’ve argued that both solutions to the problem imply that there is an important and pervasive sort of pragmatic encroachment in Bayesian epistemology. However, the interpretive problem also has major implications for the philosophy and practice of Bayesian statistics.\(^{26}\) For one thing, once you change the way you interpret what you are doing, what you are doing will also sometimes have to change. We have already seen this in Section 6, where it was shown that whether a prior probability function is rational depends on the use to which the probability function will be put and on what you take yourself to be doing.

But the interpretive problem arguably has even greater implications for how we are to interpret, and use, the likelihood function and associated principles such as the Law of Likelihood.\(^{27}\) In particular, although I will not argue this here, the counterfactual and verisimilitude interpretations open the door to the possibility that it may sometimes be rational to use a measure of evidential impact other than the likelihood. Thus, although this paper has been concerned with showing that we sometimes need to change the standard Bayesian semantics, once we have a new semantics, it becomes apparent that we may sometimes be justified in also changing the standard Bayesian syntax.

References


Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard,\(^{26}\) As Gelman and Shalizi (2013) also point out.

\(^{27}\) According to the law of likelihood, evidence \(E\) favors \(H_1\) over \(H_2\) if and only if \(p(E|H_1) > p(E|H_2)\).


