Goals and the Informativeness of Prior Probabilities

Olav B. Vassend

August 4, 2017

Abstract

I argue that information is a goal-relative concept for Bayesians. More precisely, I argue that how much information (or confirmation) is provided by a piece of evidence depends on whether the goal is to learn the truth or to rank actions by their expected utility, and that different confirmation measures should therefore be used in different contexts. I then show how information measures may reasonably be derived from confirmation measures, and I show how to derive goal-relative non-informative and informative priors given background information. Finally, I argue that my arguments have important implications for both objective and subjective Bayesianism. In particular, the Uniqueness Thesis is either false or must be modified. Moreover, objective Bayesians must concede that pragmatic factors systematically influence which priors are rational, and subjective Bayesians must concede that pragmatic factors sometimes partly determine which prior distribution most accurately represents an agent's epistemic state.

1 Introduction

Suppose you are about to roll a six-sided die (with faces numbered one through six) and you want a probability distribution that represents how probable each of the six possible outcomes is.[1,2] I have rolled the die many times already, and I tell you that – on average – the die has landed on 5. Clearly, the die is strongly biased towards landing on high numbers, and it seems intuitively probable that the die will land on a high number on the next roll as well. But how do you come up with precise probabilities for each of the possible outcomes? This is an instance of the so-called "problem of the priors": how do you translate background information into a probability distribution, and – in the absence of background information – how do you represent a lack of information probabilistically? This paper argues that how you should answer these questions depends on what goals you have.

More precisely, I will consider two different situations, defined by two different goals that an agent may have. In the first situation, the goal of the agent is to learn which hypothesis in a partition of hypotheses is true. In the second situation, the agent instead intends to use the partition of hypotheses as a predictive tool in decision making. My arguments will show that these two situations call for different prior distributions. The implication in the die example is that you need to figure out why you are interested in the outcome of the die roll before you can figure out which

[1]I wish to thank Malcolm Forster, Jan Sprenger, Reuben Stern, and the FEW 2016 referees for reading a version of the paper. I also wish to thank the FEW audience, and especially mycommentator, Kenny Easwaran. I also thank the referees for helpful comments.

[2]This example is originally due to E. T. Jaynes (see, e.g., Jaynes (1989)). For an extended critical discussion, see Seidenfeld (1986). Although the example is clearly highly artificial, it is structurally similar to many real scientific examples.

prior probability you should use.[3]

The arguments of the paper have important implications for both objective and subjective Bayesians. In particular, the Uniqueness Thesis for priors, which is a prominent thesis among objective Bayesians according to which there is a uniquely rational prior given any background information, is either false or must be modified. Moreover, objective Bayesians must concede that pragmatic factors systematically influence which probability distribution is most rational. Subjective Bayesians, on the other hand, must concede that pragmatic factors sometimes in part determine which probability distribution most faithfully represents an agent's epistemic state.

2 Notation and the Basics of Bayesianism

A few notational remarks are in order. First, I will generally use $\mathbf{H}$ to refer to a partition of hypotheses (i.e. a set of mutually exclusive and exhaustive hypotheses), and I will use H_j to refer to some arbitrary member in the partition. Similarly, I will generally use $\mathbf{E}$ to refer to a partition of possible evidence and E_i to refer to some element in the partition. However, if I am explicitly discussing a *continuous* hypothesis space (i.e. a hypothesis space that is indexed by a real-valued parameter), then I will use Θ to refer to a partition of hypotheses, θ to refer to some hypothesis in the partition, X to refer to a partition of possible evidence, and x to refer to an element of the partition. Generally, sums over all the elements in a partition will be denoted by $\sum_i$ or $\sum_j$, unless the sum is over a continuous space, in which case integrals will be used instead.

[3] I return to the die example in Section 8.

The basic problem in Bayesian inference is to infer the true, predictively accurate, or otherwise useful hypotheses in $\mathbf{H}$ given some particular observation E. Bayesians solve this problem by using Bayes's theorem. Bayes's theorem requires two ingredients: a prior probability function and a likelihood function. A *prior* probability function is a probability distribution p over $\mathbf{H}$ that is supposed to represent how probable each H_j is prior to any evidence. For ease of notation, I will sometimes use $p(\mathbf{H})$ to refer to the *set* of probabilities $p(H_1)$, $p(H_2)$, ..., $p(H_n)$ over the partition $\mathbf{H}$. The fact that p is a probability distribution means that we require that $\sum_j p(H_j) = 1$. A *likelihood* function, $p(E|H_j)$, is a function that says how probable each H_j makes the observation E.

Once we have a prior and a likelihood, Bayes's theorem says that $p(H_j|E) = cp(H_j)p(E|H_j)$.[4] To a Bayesian, $p(H|E)$ is the new probability that we ought to assign to H in light of having observed E.

3 The Importance of Information Measures for Bayesianism

An "information measure" is a quantitative measure of how "informative" or "opinionated" a probability distribution is.[5] The most well known information measure is the Shannon entropy, which says that the information content in $p(\mathbf{H})$ is given by $-\sum_j p(H_j) \log p(H_j)$. The higher the Shannon entropy, the less informative and less opinionated is the probability distribution. The probability distribution with

[4]Here $c = 1/\sum_j p(E|H_j)p(H_j)$.

[5]In this paper I will use "informative" and "opinionated" interchangeably.

the highest Shannon entropy is the "flat" distribution that assigns the same probability to every hypothesis in **H**. Intuitively the flat distribution is indeed the least informative and least opinionated probability distribution since it does not favor any hypothesis in **H** over any other. On the other hand, the distribution over **H** that has the lowest Shannon entropy and is therefore the most informative is the distribution that assigns all its probability mass to one of the hypotheses. This also seems intuitively reasonable. Indeed, we may view it as a sanity check on any proposed information measure that the measure deem a probability distribution that assigns all its probability to a single hypothesis maximally opinionated, and that it deem the flat probability distribution minimally opinionated.[6]

But what about all the other probability distributions in between the maximally and minimally opinionated ones? Here intuition often comes up short. Let's say we are considering distributions over a partition of three hypotheses. Is a distribution that assigns probabilities of 0.2, 0.3 and 0.5 to the three hypotheses more or less opinionated than a distribution that assigns 0.15, 0.4, and 0.45? This may seem like an esoteric question, but the answer to the question is of crucial importance, and is sensitive to the choice of information measure.

The reason why this question is of crucial importance to so-called "objective Bayesians" is clear. According to most objective Bayesians a probability distribution is rational for an agent if and only if the distribution is maximally non-informative relative to the agent's background knowledge; thus, objective Bayesians explicitly need an information measure in order to evaluate how informative various candidate

[6]This sanity check only makes sense when the hypothesis space is finite. Matters are subtler when the hypothesis space is continuous, as we shall see later.

probability distributions are.

That information measures are also crucially important to subjective Bayesians is probably a more contentious claim. I defer a more thorough discussion of this issue to Section 9.2, since my discussion will rely on developments made in the paper. However, the reason why information measures are also important to subjective Bayesians can be put briefly as follows. Subjective Bayesians hold that an agent's probability distribution should accurately represent the agent's epistemic state. Since most of us do not have numerical probabilities in our heads, this introduces a kind translation problem, because agents' subjective degrees of confidence must somehow be translated into numbers. How this translation problem should be solved will sometimes depend on what the goals of the agent are and what the correspondingly suitable information measure is.

4 Other Approaches to Measuring Information

Many information measures have been proposed in the statistical and information theory literatures.[7] Which of these many information measures is appropriate for Bayesian purposes? Most Bayesians who have thought about this issue have endorsed the aforementioned Shannon information measure. As was pointed out previously, the Shannon entropy has the intuitively appealing feature of declaring the flat distribution maximally uninformative and the distribution that assigns all its probability to a single hypothesis maximally informative. However, there are many other infor-

[7]Including two (infinitely) large classes of information measures, the Rényi measures (Rényi, 1961) and the Tsallis measures (Tsallis, 1988) .

mation measures that also have this feature,[8] so why go for the Shannon entropy rather than one of these other measures?

The standard arguments in favor of Shannon's information measure have nothing in particular to do with Bayesian inference,[9] and it is therefore unclear why Bayesians should care about these arguments.

For example, one of the standard postulates used to derive Shannon's information measure holds that the information content of a probability distribution should decrease as the number of hypotheses increases, all else being equal. This postulate has dubious relevance to Bayesian inference, however, because in Bayesian analyses the hypothesis space is almost always held fixed throughout the analysis. And even if we do demand that our information measure satisfy this requirement, there are many information measures that satisfy it aside from Shannon entropy.

Indeed, in the traditional argument for Shannon's information measure, the only property that distinguishes Shannon's measure from a whole slew of other information measures is that it has a certain additivity property (Rényi, 1961). Although it may make sense to require this additivity property in the original communication theory context in which Shannon information was introduced, it's not clear why an information measure needs to have the property in the context of Bayesian inference.

Some Bayesians have taken a more radical and pluralist approach to information measures. For example, Morris DeGroot (1962) defines "the value of information"

[8] Including all Rényi and Tsallis measures.

[9] A notable exception is Jon Williamson (2010), who uses an argument based on Bayesian scoring rules. However, below I will argue that the scoring rule he relies on is only appropriate in what I call the "learning" situation, where the goal is to identify the true hypothesis in a partition of hypotheses.

as the difference that a piece of evidence makes to the expected utility calculation of an agent. This definition is used by Bernardo (1981) to define "minimally valuable" priors. However, the "minimally valuable" prior is often not the flat distribution and is sometimes even the probability function that assigns all its probability mass to a single hypothesis. Hence, whatever the "minimally valuable" prior is supposed to be, it should not be interpreted as the prior that is maximally uninformative,[10] and DeGroot's measure is therefore not an appropriate measure of the informativeness of probability functions, since the measure clearly fails the previously mentioned sanity checks. The reason DeGroot's measure gives unintuitive results is because the measure depends on the utility function of the agent.

The approach advocated here is intermediate between the preceding two approaches. I do not think information measures should be functionally dependent on agents' utilities, but I also do not think a single measure of information is appropriate in all contexts, nor do I think arguments for information measures should proceed in a complete vacuum from the contexts in which the information measures will play a role. In particular, in a Bayesian context, the prior and the posterior probabilities of a hypothesis are the fundamental quantities that represent how probable the hypothesis is prior to and after the observation of evidence, respectively. Since evidence is the conveyer of information, the starting point of my argument is the following foundational observation about information in a Bayesian context:

> **Observation** Given some hypothesis H and evidence E, the posterior, $p(H|E)$ is *more informed* than the prior, $p(H)$

[10] It is not clear Bernardo (1981) would have endorsed such an interpretation either.

That the posterior probability is more informed than the prior seems to me to be a truism, but the question now arises of *how much* more informed the posterior is when compared to the prior.

5 Confirmation Measures as Measures of the Informativeness of Data

In general, a measure of the distance between the posterior and prior probability of a hypothesis given a piece of evidence is known as a "confirmation measure." Here, I will follow convention and use $c(H, E)$ to refer to the confirmation score of H given E according to some unspecified confirmation measure. Additionally, two specific confirmation measures will play a particularly important role. The *difference measure*, $d(H, E)$, measures the degree of confirmation that E confers on H as $p(H|E) - p(H)$. The *log-ratio measure*, $lr(H, E)$, measures the degree of confirmation as $\log \frac{p(H|E)}{p(H)}$. Note that both the difference measure and the log-ratio measure have the property that 0 signifies that E confers no confirmation on H.

Importantly for our purposes, confirmation measures may naturally be interpreted as quantitative measures of how much information a piece of evidence provides with respect to a hypothesis.[11] For example, if $c(H, E)$ is a large number (either positive or negative), then that means that E provides us with a lot of information about H, since H greatly changes the probability of H; if, on the other hand, $c(H, E)$

[11] That confirmation measures may be interpreted in this way is not to deny that they may also be interpreted in other ways. For example, one prominent strand of confirmation theory (e.g. Crupi and Tentori (2013)) regards confirmation as a generalization of logical entailment. I thank Jan Sprenger for emphasizing this to me.

is 0, then E provides us with no information about H.

It is immediately clear that different confirmation measures will in general disagree on how informative a given datum is, and sometimes the extent of disagreement can be extreme. For example, a change from $Pr(H) = 0.00001$ to $Pr(H|E) = 0.01$ is trivial compared to a change from 0.5 to 0.6 if we use the difference measure; but according to the log-ratio measure, the first change is much greater than the second. How informative E is with respect to H therefore depends on which confirmation measure is used.

The argument put forward here will be that the appropriate way to measure the distance between the posterior and the prior probability of a hypothesis depends on the goals of the agent. Thus, for example, whether the difference between a probability of 0.01 and a probability 0.1 is "big" or "small" depends on pragmatic factors. I will consider two more specific goals that an agent may have in order to demonstrate the point.

5.1 The Learning Situation and the Log-Ratio Measure

In the first situation I consider – let's call it the "learning situation" – the goal is to identify which hypothesis, H, in a partition of hypotheses, $\mathbf{H}$, is true. Translated into the Bayesian framework, the goal is for the posterior probability of the true hypothesis, H_0, to be as large as possible. Ideally, we want $p(H_0|E) = 1$. Given that this is the goal, what is the best way to measure the extent to which E informs us about some H in $\mathbf{H}$?

One way to make the goal more explicit is by creating a "scoring rule" that

more precisely encodes what our epistemic values are in the learning situation. A "scoring rule" is a function of the form $s(p, H_0)$, where H_0 is the true hypothesis in the partition $\mathbf{H}$. The score of p is supposed to represent how well p achieves our goals. The defining feature of the learning situation is that we want to assign as much probability to H_0 as possible. A reasonable way to formalize this goal is to require that a probability function, p, receive a higher score than a different probability function, q, if and only if $p(H_0) > q(H_0)$.

A scoring rule that ranks p as better than q if and only if p assigns the truth a higher probability than does q is sometimes known in the literature as a "local" scoring rule. Such scoring rules are "local" because the probabilities that p and q assign to false hypotheses are irrelevant to which probability function receives a higher score.[12] Sometimes we do care about how inaccurate our probabilities in false hypotheses are, and in those cases locality is a bad requirement to make of our scoring rule. However, locality is a very reasonable requirement to make of a scoring rule in the learning situation, because in the learning situation the objective is precisely and only to identify the truth.

Out of the well-known and independently plausible scoring rules, the only local scoring rule is the log scoring rule, which assigns a score of $\log p(H_0)$ to p. In fact, the log-scoring rule is the only local scoring rule that is *strictly proper* (Bernardo, 1979a), which is a property that many philosophers have argued any reasonable scoring rule ought to have (see, e.g. Oddie (1997), Gibbard (2007), Joyce (2009), and Horowitz (2014)). The log-scoring rule is therefore a *reasonable* scoring rule in the learning

[12]Clearly, in practice we often do care about which probabilities we assign to false hypothesis, so the learning situation, as characterized here, describes a somewhat idealized epistemic goal.

situation: it appropriately encodes the epistemic goal of learning the truth. Note that this does not mean that the log-scoring rule is the *uniquely* rational scoring rule in the learning situation.

As Steven van Enk (2014) points out in a recent paper, there is a clear connection between scoring rules and confirmation measures. More precisely, the extent to which E confirms (or disconfirms) a hypothesis H can also naturally be understood as the extent to which E changes the *score* of $p(H)$, on the assumption that H is true. The idea is that the scoring rule assigns an epistemic value to the posterior and to the prior, and the difference in score is therefore the difference that the evidence makes to the epistemic value of the hypothesis.

In the learning context, the epistemic value is to learn the truth, so the difference in log-score between $p(H|E)$ and $p(H)$ is therefore the difference that the evidence makes to the goal of learning whether H is true. If we measure this difference by taking the arithmetic difference, we end up with the log-ratio measure of confirmation:

$$\log p(H|E) - \log p(H) = lr(H, E). \tag{5.1}$$

Thus, we get the conclusion that the log-ratio measure is a *reasonable* measure of the informativeness of evidence in the learning situation, where the goal is to learn whether H is true.[13]

[13]Why measure the difference between the log-score of the posterior and the prior using the arithmetic difference rather than, say, the ratio, $\frac{\log p(H|E)}{\log p(H)}$? Of course, we could use the ratio instead of the difference, but the resulting confirmation measure is not independently plausible, in contrast to the familiar log-ratio measure. In any case, I am not claiming that the formal choices I make here and other places in the paper are *uniquely* reasonable, but only that they are reasonable.

The above argument is not intended to be a knock-down argument for the log-ratio measure of confirmation; the argument is only intended to show that the log-ratio measure is *reasonable* in the learning situation, where the goal is to identify the true hypothesis in a partition of hypotheses. Indeed, although the log-ratio measure is reasonable in the learning situation, it is not reasonable in all other situations; in the next subsection, I consider a different situation where the log-ratio measure is not reasonable, while another confirmation measure is.

5.2 The Decision Situation and the Difference Measure

Our goal is not always to find the truth; sometimes the goal is to make a good decision. Thus the second situation I will consider is the "decision situation." In the traditional Bayesian formalization of the decision situation, there is a preference ranking over a partition of various states S_m that the world may be in, and there is also a partition of possible available acts A_n ranked by their expected utility. For example, S_m may represent hypotheses about how much it is going to rain in the next hour, and A_n may represent how far away from home we are willing to venture without an umbrella.

For simplicity, I will assume in this paper that the acts and states are independent according to p.[14] More importantly, I will also assume that the utility function does not depend on p or on possible evidence.[15] The "prior" expected utility of some act

[14]When the acts and states are not independent, there is some controversy over which Bayesian decision theory is the correct one. Some endorse Causal Decision Theory (e.g. Lewis (1981), Pearl (2009), and Joyce (2009)), while others endorse Evidential Decision Theory (e.g. Jeffrey (1983), Eells (1991), and Ahmed (2012).

[15]Hence, the utility function is not a scoring rule in the sense of the previous section. The learning situation as I presented it in the previous section may also be regarded as a kind of decision problem,

A_n is then defined[16] as:

$$EU(A_n) = \sum_m p(S_m)U(S_m\&A_n) \tag{5.2}$$

Here, $U(S_m\&A_n)$ is the utility of performing A_n when S_m obtains. For example, going on a long walk without an umbrella when it rains a lot has a low utility for me, but going on a long walk without an umbrella when it's sunny has a high utility.

Now suppose we also have available a partition of hypotheses, **H**, that can be used to predict whether S_m will obtain. For example, **H** may be hypotheses about what the barometric pressure will be in the next hour. Clearly, if we knew what the barometric pressure H_0 would be in the next hour, then we could use that information to predict how much it would rain using the conditional probability $p(S_m|H_0)$. Unfortunately, we don't know what the barometric pressure is going to be, so we need to put a prior probability over **H**, $p(H_j)$, that represents the probability of each of the possible values the barometric pressure can take in the next hour. Once we have this prior distribution, we can use the H_j's to predict the S_m's by using the law of total probability:

$$p(S_m) = \sum_j p(S_m|H_j) * p(H_j) \tag{5.3}$$

Now, suppose we wanted to use a scoring rule to evaluate the prior probability distribution over **H**. Is the log-scoring rule appropriate in this context? By assump-

but it is important to realize that it is a qualitatively very different decision problem from the kind of decision problem considered in this section, because the utility function (i.e. the scoring rule) in the learning situation depends on the agent's probability function and on the data.

[16]Following Savage (1954).

tion, we do not really care about what the true value of the barometric pressure is; what we care about is how much it will rain in the next hour. The hypotheses about barometric pressure are therefore for us mere *predictive tools*. Clearly, if the goal is to use the H_j's to predict which S_m is going to obtain, then we want to assign high probabilities to predictively accurate hypotheses (irrespective of whether they are true) and low probabilities to predictively inaccurate hypotheses. The true hypothesis only has a special status insofar as it can be expected to have the highest predictive accuracy. But a probability function that assigns a high probability to the truth will not be good for predictive purposes if it also assigns high probabilities to hypotheses that are very predictively inaccurate, and, moreover, it will not be better than a probability function that assigns a low (even 0) probability to the truth, but at the same time only assigns high probabilities to predictively accurate hypotheses. But this means that a local scoring rule, such as the log-scoring rule, is inappropriate, because a local scoring rule scores probability functions only by the probabilities that they assign to the truth.

In particular, in the prediction of S_m (i.e. formula (5.3)), each H_j is in a sense equally important because each H_j is used in the weighted prediction, so a non-local scoring rule that takes into account the probability assigned to every hypothesis in the partition seems much more appropriate. The most well known non-local scoring rule that does this is the quadratic scoring rule, which assigns a score of $\sum_j (i(H_j) - p(H_j))^2$ to p, where $i(H_j)$ is the indicator function that assigns 1 to H_j if H_j is true and 0 otherwise. The quadratic scoring rule therefore seems more appropriate than the log-scoring rule for the purpose of evaluating our prior over **H**

in the decision situation, where **H** is used as a predictive tool. Moreover, as van Enk (2014) shows, the standard confirmation measure that is associated with the quadratic scoring rule is the difference measure. Hence we get the conclusion that the difference measure, and not the log-ratio measure, is a reasonable measure of the informativeness of evidence in the decision situation.

The above argument is rather sketchy, so here is a more detailed analysis that shows how the difference measure of confirmation naturally arises in the decision situation. First, note that we can plug (5.2) into (5.1) in order to make the dependence of the expected utility of A_n on H_j explicit:

$$EU(A_n) = \sum_m \sum_j p(S_m|H_j)p(H_j)U(S_m\&A_n) \tag{5.4}$$

Next, suppose we receive evidence regarding which hypothesis in **H** is true in the form of data E; for example E may be data about the barometric pressure from two hours ago. What the barometric pressure was two hours ago is clearly relevant to what the barometric pressure will be in the next hour, so if we are good Bayesians, we will update each prior $p(H_j)$ to a posterior $p(H_j|E)$ to take into account this new information. If we do, then the new "posterior" expected utility of A_n is:

$$EU(A_n|E) = \sum_m \sum_j p(S_m|H_j, E)p(H_j|E)U(S_m\&A_n) \tag{5.5}$$

Here, $p(S_m|H_j, E)$ represents the probability that it will rain S_m millimeters in the next hour, given that the barometric pressure in the next hour is H_j and the barometric pressure two hours ago was E. It is natural to assume here and in many

other similar cases that E does not give us any information about S_m except insofar as E provides us with information about H_j.That is, it is natural to assume that $p(S_m|H_j, E) = p(S_m|H_j)$.[17] If we make this assumption, then the posterior expected utility of A_n is simply:

$$EU(A_n|E) = \sum_m \sum_j p(S_m|H_j)p(H_j|E)U(S_m\&A_n) \tag{5.6}$$

Now, if we take the difference between the posterior expected utility of A_n and the prior expected utility of A_n, we arrive at the following expression:

$$\Delta EU(A_n; E) = EU(A_n|E) - EU(A_n) = \sum_i \sum_j p(S_i|H_j)[\mathbf{p(H_j|E)} - \mathbf{p(H_j)}]U(S_i\&A_j) \tag{5.7}$$

Or, in other words,

$$\Delta EU(A_n; E) = \sum_i \sum_j p(S_i|H_j)\mathbf{d(H_j, E)}U(S_i\&A_j) \tag{5.8}$$

Here, $d(H_j, E)$ is the confirmation conferred on H_j by E according to the difference measure $p(H_j|E) - p(H_j)$. Again, the above expressions may look complicated, but the important thing to note is that the difference between the posterior and prior expected utility of A_n depends on the data *only* through $d(H_j, E)$. In the decision situation, we do not care about which H_j is true; we only care about H_j insofar as it can help us predict S_m and thereby influence our preference ranking over A_n. Clearly

[17] As has been pointed out to me by Reuben Stern, this assumption does not always hold, but it holds very widely.

the only way our preference ranking can change given x is if $\Delta EU(A_n; E)$ is non-zero for some A_n. But $\Delta EU(A_n; E)$ depends on the data only through $d(H_j, E)$; hence, in the decision situation, $d(H_j, E)$ arises as a natural measure of the informativeness of E with respect to H_n.

But why use the arithmetic difference between the posterior and prior expected utility to measure the impact that E has on the expected utility of A_n? Isn't that a circular way of arguing in favor of the difference measure? Why not use, say, the ratio instead?

One answer to this objection[18] is that we do not really have a choice, because the ratio between two expected utilities will in general not be a meaningful quantity. This is because utility functions are usually only defined up to arbitrary linear transformations. In other words, if U is the utility function of some agent, then $aU + b$ is usually an equally valid representation of the agent's utilities, for any real number b and positive real number a. For instance, Savage's (1954) famous representation theorem, and its various descendants, only define the utility function up to arbitrary positive linear transformations. As a result of this, utilities and expected utilities exist on an interval scale (Stevens, 1946). But this means that the ratio of two utilities is not meaningful, because the ratio will change if you transform the utility scale with an arbitrary positive linear transformation. As an analogy, celsius and fahrenheit are interval scale measurements of temperature: it is meaningful to say that the difference between 5 and 10 degrees celsius is the same as the difference between 15 and 20 degrees celsius, because these differences remain equal if they are

[18] I will say a bit more about it in the next section.

both transformed to the fahrenheit scale. However, it is not meaningful to say that 10 degrees celsius is "twice as large" as 5 degrees celsius, because the ratio between these temperatures changes if the temperatures are transformed to the fahrenheit scale.

5.3 Numerical Examples Showing Why the Learning and Decision Situations Require Different Measures of Confirmation

Neither of the preceding subsections is intended to offer a knock-down argument; the argument in subsection 5.1 merely shows the log-ratio measure to be an especially reasonable confirmation measure in the learning situation, and the argument in subsection 5.2 just shows the difference measure to be especially reasonable in the decision situation. Furthermore, the arguments may appear rather abstract. Simple numerical examples help illustrate and independently bolster the claim that the decision situation and the learning situation call for different confirmation/information measures.

In particular, suppose you have just two hypotheses, H and $\neg H$ and consider two different scenarios: in the first scenario, the probability of H changes from 0.0001 to 0.1001; in the second scenario, the probability of H instead changes from 0.4 to 0.5. Which of these changes is more informative?

Suppose, first, that you are in the learning situation, so that your goal is to figure out which of H or $\neg H$ is true. According to the odds version of Bayes's formula,

$$\frac{p(H|E)}{p(\neg H|E)} = \frac{p(E|H)}{p(E|\neg H)} \frac{p(H)}{p(\neg H)} \tag{5.9}$$

Thus, if the probability of H changes from 0.0001 to 0.1001, then $\frac{p(E|H)}{p(E|\neg H)} = 1111$. If, on the other hand, the probability of H changes from 0.4 to 0.5, then by the same calculation, $\frac{p(E|H)}{p(E|\neg H)} = 1.5$. Thus, the change from 0.0001 to 0.1001 requires that H predict the evidence much better than $\neg H$, whereas the change from 0.4 to 0.5 does not.

Let's make the example more vivid by providing some concrete numbers. Suppose $\neg H$ assigns E a probability of 0.0009 and that H assigns E a probability of 0.9999, and suppose E is observed. Intuitively, the observation of E very strongly suggests that H is true and that $\neg H$ is false because $\neg H$'s prediction was that E was basically impossible whereas H predicted that E was almost sure to happen. Under these conditions, if H's prior probability is 0.0001, then H's posterior will be 0.1001. Thus, the difference between 0.0001 and 0.1001 is actually extremely large in this context.

Suppose, on the other hand, that H assigns E a probability of 0.9 and that $\neg H$ assigns E a probability of 0.6, and suppose again that E is observed. In this scenario, the observation of E only weakly suggests that H rather than $\neg H$ is true. H and $\neg H$ both predicted E as more likely than not, and both also assigned $\neg E$ a fairly high probability. Under these conditions, if H's prior probability probability is 0.4, then H's posterior probability will be 0.5. Hence the difference between 0.4 and 0.5 is not very large in this context.

This numerical example, which has nothing to do with scoring rules and therefore provides an argument independent of the one provided in the subsection (5.2),

strongly suggests that the change from 0.0001 to 0.1001 is much more informative regarding H's truth value than is the change from 0.4 to 0.5. This is exactly the verdict delivered by the log-ratio measure.[19]. The difference measure, on the other hand, says that these changes in probability are equally informative, which does not seem reasonable.

But now suppose you are instead in the decision situation, and suppose you are calculating the expected utility of some action A. Then, as the following calculation shows, the change in the expected utility of A is the same whether the probability of H changes from 0.0001 to 0.1001 or from 0.4 to 0.5. For suppose first that the probability of H changes from 0.0001 to 0.1001. Then:

$$\Delta EU(A;E) = \sum_m p(S_m|H)U(S_m\&A)\big[p(H|E) - p(H)\big] + \sum_m p(S_m|\neg H)U(S_m\&A)\big[p(\neg H|E) - p(\neg H)\big] \tag{5.10}$$

$$= \sum_m p(S_m|H)U(S_m\&A)\big[0.1001 - 0.0001\big] + \sum_m p(S_m|\neg H)U(S_m\&a)\big[0.8999 - 0.999\big] \tag{5.11}$$

$$= \sum_m p(S_m|H)U(S_m\&A) * 0.1 - \sum_m p(S_m|\neg H)U(S_m\&A) * 0.1 \tag{5.12}$$

Suppose, on the other hand, that the probability of H changes from 0.4 to 0.5; then the change in the expected utility of A is,

[19]Of course, other confirmation measures also deliver this verdict, such as the log-likelihood ratio, for example. The argument presented here therefore does not single out – and is not intended to single out – the log-ratio confirmation measure as better than *all* other confirmation measures; the argument only establishes the log-ratio measure of confirmation as a reasonable measure of confirmation.

$$\Delta EU(A;E) = \sum_m p(S_m|H)U(S_m\&A)\big[0.5 - 0.4\big] + \sum_m p(S_m|\neg H)U(S_m\&a)\big[0.5 - 0.6\big] \tag{5.13}$$

$$= \sum_m p(S_m|H)U(S_m\&A) * 0.1 - \sum_m p(S_m|\neg H)U(S_m\&A) * 0.1 \tag{5.14}$$

The fact that (5.12) to (5.14) are identical implies that the change in the expected utility of A is the same whether the probability of H changes from 0.0001 to 0.1001 or from 0.4 to 0.5. Thus, in this context, a change in probability from 0.4 to 0.5 is exactly as informative as a change in probability from 0.0001 to 0.1001. And this is the verdict delivered by the difference measure.

On the other hand, the log-ratio measure is *not* a reasonable measure of informativeness in the decision situation. In fact, for every $\epsilon > 0$, no matter how small, and every $B > 0$, no matter how large, it is easy to come up with examples[20] such that the degree of confirmation (or disconfirmation) conferred by E according the log-ratio measure is greater than B, while at the same time, for every n, $|E(A_n;E) - E(A_n)| < \epsilon$ and $|E(A_n;E)/E(A_n) - 1| < \epsilon$, i.e. the difference that E makes to the expected utility of every action under consideration is arbitrarily small, regardless of whether you measure the impact that E has on the expected utility ranking of actions as a difference or as a ratio.

Arguably, in the decision situation, where what you care about is the expected utility ranking of the actions under consideration, it does not make sense to say that

[20]For reasons of space, I will omit the details here.

E provides you with a lot of information if E has essentially no influence on the expected utility of any action. But that is what you have to say if you measure informational impact with the log-ratio measure. The log-ratio measure is therefore not a reasonable measure of informativeness in the decision situation.

6 How to Derive Information Measures From Confirmation Measures

So far, I've argued that how informative a piece of evidence is depends on the goal. In the learning situation, the informativeness of E with respect to H is reasonably quantified by the log-ratio measure, whereas the difference measure is not reasonable. However, in the decision situation, the informativeness of E with respect to H is reasonably quantified by the difference measure, whereas the log-ratio measure is not reasonable.

However, the ultimate goal of the paper is to show that how much information there is in a *probability distribution* depends on how the probability distribution will be used. The next goal of the paper is therefore to show how information measures may reasonably be derived from confirmation measures. As before, I do not claim that the formal choices made are uniquely rational: I only claim that they are reasonable. There may be other reasonable ways of deriving information measures from confirmation measures, but the point will still stand that the decision situation and the learning situation call for different information measures because they call for different confirmation measures.

6.1 How to Extend a Confirmation Measure to a Partition of Hypotheses, or: How to Measure the Information Distance Between the Prior and Posterior Distributions

A confirmation measure tells us how informative E is with respect to some particular H_j in $\mathbf{H}$. Of course, E will have an impact on each H_j in the partition. How may we quantify the effect that E has on the partition overall? Or, to put the same question in somewhat different terms, how do we measure the "information distance" between the whole posterior distribution and the whole prior distribution? One way to do so is to simply add up all the individual confirmation scores, $\sum_j c(H_j, E)$, for each H_j in the partition. This implicitly weighs each confirmation score as equally significant. An alternative approach that is more appealing from a Bayesian perspective is to weigh each term in the sum using either the prior or the posterior. Since the posterior is more well-informed than the prior, it makes more sense to use the posterior than the prior. Following this idea leads us to quantify the impact of E on $\mathbf{H}$ as $\sum_j c(H_j, E)p(H_j|E)$.

Plugging in various confirmation measures for c yields different measures of the information distance between the posterior distribution and the prior distribution. For example, plugging in the log-ratio confirmation measure for c yields the well known Kullback-Leibler divergence (Kullback and Leibler, 1951), which lends further credence to the idea that $\sum_j c(H_j, E)p(H_j|E)$ is a reasonable general measure of the information distance between the posterior and the prior. Quantifying the impact of E on $\mathbf{H}$ in this way is also endorsed by Crupi and Tentori (2014).

Thus, I contend, the following is a reasonable (though not necessarily uniquely reasonable) quantification of the information distance between the prior distribution and the posterior distribution, given some piece of evidence E:

The information distance between the posterior and the prior distribution. *Given a confirmation measure c, a piece of evidence E, and a probability function p, the information distance between the prior distribution $p(\mathbf{H})$ and the posterior distribution $p(\mathbf{H}|E)$ is defined as follows:*

$$\text{InfDis}(p(\mathbf{H}|E), p(\mathbf{H})) = \sum_j c(H_j, E) * p(H_j|E) \tag{6.1}$$

(6.1) tells us the information distance between p and the posterior given some particular E_k in $\mathbf{E}$. Different E_k's will, of course, result in different posteriors. Before we receive the evidence, how much evidence can we *expect* to receive from $\mathbf{E}$? Or, put differently, how much information – on average – will $\mathbf{E}$ provide us about $\mathbf{H}$? A reasonable way to quantify the answer to this question is to simply average $\text{InfDis}(p(\mathbf{H}|E), p(\mathbf{H}))$ over the partition $\mathbf{E}$ (again, this is also suggested by Crupi and Tentori (2014)):

$$\text{InfDis}(p(\mathbf{H}|\mathbf{E}), p(\mathbf{H})) = \sum_i \text{InfDis}(p(\mathbf{H}|E), p(\mathbf{H})) * p(E_i) \tag{6.2}$$

(6.2) tells us how much information, on average, the partition of evidence $\mathbf{E}$ can be expected to provide us about the partition of hypotheses $\mathbf{H}$. A trick due to Jose Bernardo (1979) is now all we need in order to derive information measures.[21]

[21] I emphasize that my interpretation of Bernardo's trick differs significantly from Bernardo's own.

6.2 How to Define Information Measures From Measures of the Information Distance Between the Posterior and the Prior Distribution

More precisely, a prior is intuitively non-informative to the extent that it is distant from most posteriors that are *heavily influenced* by data. To formalize this idea, imagine that we are going to receive a large amount of evidence $\mathbf{E}^1, \mathbf{E}^2, \ldots\ldots \mathbf{E}^n$. As the amount of evidence increases ($n \to \infty$), the posterior distribution will gradually become increasingly informed by the evidence, and in the limit of infinite evidence, the posterior distribution will be maximally informed and maximally opinionated; that is, some hypothesis (we do not know which one) will have a probability of 1.[22] A prior distribution is then non-informative in proportion to how informationally distant, on average, it will be from the maximally informative posterior distribution, whatever the maximally posterior distribution turns out to be. Using the definition of InfDis (6.2), we can formally quantify the preceding ideas, and define the information content of the prior distribution, $p(\mathbf{H})$, as follows:

$$\mathrm{Inf}(p) = \lim_{n\to\infty} \mathrm{InfDis}(p(\mathbf{H}|\mathbf{E}^n), p(\mathbf{H})) \tag{6.3}$$

It is very important to note that we do not need an actual sequence of evidence in order to make sense of (6.3). The imagined sequence of evidence, $\mathbf{E}^1, \mathbf{E}^2, \ldots\ldots \mathbf{E}^n$, merely functions as a way of formalizing the idea that the posterior gets increasingly

For more faithful presentations of Bernardo's views, see Bernardo (1979b), Berger et al. (2009), or Sprenger (2012).

[22] Well known convergence results guarantee that the probability distribution will converge under widely applicable conditions (see, e.g. Hawthorne (manuscript)).

informed as more evidence comes in. The derivation in Appendix A shows that, when the hypothesis space is finite, properties of the sequence of imagined evidence (e.g. the distribution of the evidence) do not make a difference to the information content of $Pr(\mathbf{H})$.[23]

If we plug (6.1) and (6.2) into (6.3), we get the following alternative expression for Inf(p), which makes the dependence on the choice of confirmation measure explicit:

$$\mathrm{Inf}(p) = \lim_{n\to\infty} \sum_i \sum_j c(H_j, E_i^n) * p(H_j, E_i^n) \tag{6.4}$$

Now we can plug different confirmation measures into (6.4) and get different information measures. In the case of a finite hypothesis space, it is actually possible to explicitly calculate (6.4) for several well known confirmation measures, and in particular for the difference measure and the log-ratio measure. More precisely, if we plug in the difference measure and the log-ratio measure, respectively, and perform the relevant calculations, we arrive at the following two alternative information measures (see Appendix A for the derivations):

The *lr* information measure. *Given p defined on a finite hypothesis space,* $\{H_i\}$,

[23]When the hypothesis space is continuous, the situation is a bit more subtle—in this case, the information content depends on the *statistical model* in which the hypotheses are situated. But this is reasonable because, in the continuous case, the hypotheses are generally indexed by continuous parameters, and it is those parameters that must be assigned probabilities. Moreover, the meaning of a parameter generally depends on the statistical model of which it is a part. For example, the parameter B in the linear model $Bx + C$ picks out the slope of a line; but in the quadratic model $Ax^2 + Bx + C$, B does not pick out the slope of a line. Thus, it is not strange that the information content of Pr(B should depend on which statistical model B is embedded in.

The information content of p according to the log-ratio measure is defined as,

$$\mathrm{Inf}_{lr}(p) = -\sum p(H_i) \log p(H_i) \tag{6.5}$$

The d information measure. *Given p defined on a finite hypothesis space, $\{H_i\}$, The information content of p according to the difference measure is defined as,*

$$\mathrm{Inf}_{d}(p) = 1 - \sum p(H_i)^2 \tag{6.6}$$

Both of the above information measures have a long and rich history, and it is both surprising and interesting in its own right that these measures have such a close connection with Bayesian measures of confirmation. $-\sum p(H_i) \log p(H_i)$ is the Shannon information of p (Shannon, 1948), which has been defended as a measure of the non-informativeness of probability functions by, among others, Edwin Jaynes (2003) and Jon Williamson (2010). $1 - \sum p(H_i)^2$ is known to ecologists as the Simpson index of diversity (Simpson, 1949) and to machine learning theorists as the Gini index. Jaynes discusses $1 - \sum p(H_i)^2$ as a possible alternative information measure (Jaynes, 2003, p. 345), but rejects it for reasons I will explain later. The diagnosis of the present paper is that both $-\sum p(H_i) \log p(H_i)$ and $1 - \sum p(H_i)^2$ are appropriate information measures, but that the two measures should be used in different contexts: $-\sum p(H_i) \log p(H_i)$ is appropriate in a learning situation, but in a decision situation $1 - \sum p(H_i)^2$ is more appropriate.

7 Two Goal-Relative Non-Informative Priors

The general definition provided in (6.4) gives us a way of selecting maximally non-informative priors. More precisely, given some confirmation measure, a probability function that *maximizes* (6.4) is a natural candidate for a prior that is maximally non-opinionated. Both (6.5) and (6.6) are maximized by a single prior – namely the flat prior – so if the hypothesis space is finite, whether you use the log-ratio or the difference measure as the confirmation measure in (6.4) will not make a difference to the non-informative prior you select. In the next section, I consider what happens when (6.5) and (6.6) are maximized relative to constraints; it turns out they can then yield different priors, and so the confirmation measure you use makes a difference when you have background information.

However, if the hypothesis space is continuous, the confirmation measure you use makes a difference even if the maximization is not relative to any constrains. For concreteness, let us consider the problem of estimating the bias θ of a coin given n coin flips. In other words, the problem is to estimate the parameter θ in the binomial distribution. Then we have:[24]

The *lr* non-informative prior. *Given the problem of estimating the parameter θ of a binomial distribution, the maximally non-informative prior density function that corresponds to the log-ratio measure is*

$$\mathrm{NonInf}_{\mathrm{lr}}(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} \tag{7.1}$$

[24]For reasons of space, I'm omitting the proof. However, a proof can be found in Bernardo (1979b).

The above prior is known as "the Jeffreys prior" after its discoverer, Harold Jeffreys (1946). We also have:[25]

The d non-informative prior. *Given the problem of estimating the parameter θ of a binomial distribution, the maximally non-informative prior density function that corresponds to the difference measure is*

$$\mathrm{NonInf}_{\mathrm{d}}(\theta) = 1 \tag{7.2}$$

The main take-away message here is that the goals you have influence which non-informative prior it is rational for you to have. Or to put the point differently: whether a probability function is "completely non-informative" or opinionated depends on the context. The Jeffreys prior can justifiably be regarded as maximally non-informative in a learning situation, but in a decision situation it is opinionated. The reverse is true for the flat prior, which is maximally non-informative in a decision situation, but opinionated in a learning situation.

8 Goal-Relative Priors Given Objective Background Information

As noted earlier, the information measures $-\sum p(H_i) \log p(H_i)$ and $1 - \sum p(H_i)^2$ are both uniquely maximized by the flat prior. However, if we have background information available, it is reasonable to maximize the two information measures

[25]For reasons of space, I again omit the proof. Unfortunately, I'm not aware of any reference where a proof may be found. However, the proof is straightforward.

relative to that background information. This is the procedure recommended by the objective Bayesians Jaynes (2003) and Williamson (2010), for example.[26]

If (6.5) and (6.6) are maximized relative to background information, they will in general not be maximized by the same priors. As a simple illustration, consider again the example provided in the introduction.[27] The example was as follows: suppose you are about to roll a six-sided die and you want a probability distribution $p(X)$ over the possible outcomes $X = 1, 2, 3, 4, 5, 6$. I have rolled the die many times already, and I tell you that – on average – the die has landed on 5. Let's first formalize the information that I give you. The natural way for you to formalize that the die has landed on 5 on average is to demand that the expected value of the die roll according to your prior should be 5. In other words, you should require that $\sum_{i=1}^{6} X_i p(X_i) = 5$. The additional constraints are, of course, that $\sum_i p(X_i) = 1$ and that $p(X_i) \geq 0$ for each X_i since probabilities must be non-negative and add up to 1.

If you maximize $-\sum p(H_i) \log p(H_i)$ relative to all of the above three constraints, you end up with the probability distribution summarized in the following table:[28]

[26]How are we to understand the learning of "background information"? This is a deep question that I do not have the space to discuss here. But, very briefly, the learning of background information cannot be the result of conditionalizing because conditionalizing requires that there already be a prior, but background information is supposed to be a constraint that is used in the construction of the prior and must therefore be "prior to the prior." For a discussion of these issues, see Williamson (2010).

[27]Again, this admittedly artificial example is structurally similar to many real examples.

[28]I have omitted the very tedious calculation.

Die	Probability
1	0.02053
2	0.03853
3	0.07232
4	0.13574
5	0.25475
6	0.47812

The distribution that maximizes $1 - \sum p(H_i)^2$, on the other hand, is as follows:[29]

Die	Probability
1	0
2	0
3	0.1
4	0.2
5	0.3
6	0.4

Perhaps the most striking difference between the two tables is the fact that the second table has zeros in it whereas the first table does not.[30] This is not incidental to this example: whereas the prior that maximizes $-\sum p(H_i) \log p(H_i)$ will never assign a probability of 0 to any hypothesis unless background information logically excludes the hypothesis, the prior that maximizes $1 - \sum p(H_i)^2$ sometimes does assign 0 to

[29] I have again omitted the tedious calculation.

[30] Another thing that may strike the reader is how nice the numbers look in the second table; however, that is incidental to this specific example and will not happen in general.

hypotheses. Indeed, it is precisely for this reason that Jaynes rejects $1 - \sum p(H_i)^2$ as a measure of non-informativeness, because he does not think that any hypothesis should ever be assigned a probability of 0 unless the hypothesis is logically excluded (Jaynes, 2003, p. 346).

The requirement that a prior never assign 0 to any outcome or hypothesis is reasonable in the learning situation. After all, the goal in the learning situation is to learn the truth, and if you assign probabilities of 0 to hypotheses, you run the risk of assigning a probability of 0 to the truth, which would ruin your chances of learning what the truth is. However, in the decision situation, the requirement that every hypothesis receive a non-zero probability is unmotivated. After all, the goal in the decision situation is not to learn the truth; therefore, accidentally assigning a probability of 0 to the truth is not necessarily a bad thing. Thus, the learning situation is inherently a more "risk-averse" setting than the decision situation, and this is reflected in the fact that $1 - \sum p(H_i)^2$ is maximized by "riskier" priors than the priors that maximize $-\sum p(H_i) \log p(H_i)$.

The reader may object that assigning a probability of 0 to a hypothesis implies that you would be willing to accept absurd bets. For example, assigning a probability of 0 to H apparently implies that you would be willing to pay USD 1,000,000,000 for a bet that pays 1 cent if H is false. That does not seem rational. However, this objection implicitly assumes that every probability is a *betting* probability. But this assumption begs the question against the arguments made in this paper. In fact, as soon as I offer you a bet over a partition of hypotheses, your goal becomes to identify which hypothesis in the partition is true. In other words, you enter

the learning situation with respect to that partition. However, according to the arguments presented here, you should only ever assign 0 to a hypothesis if you are in the decision situation, i.e. if you do not care about which of the hypotheses is true, but rather aim to use the hypotheses as a predictive tool in order to predict something else. We may call the probabilities you assign to hypotheses in the decision situation "predictive probabilities"; thus, the predictive probability you assign to H_i reflects how much trust you put in H_i's prediction. Crucially, you can have trust in the predictions of a hypothesis, even if you know that the hypothesis is false. On the other hand, your betting probability in H_i reflects the bets you would be willing to accept on the truth of H_i. Clearly you would not be willing to accept any bets on a hypothesis you know to be false; your betting probability in a hypothesis you *know* to be false is 0 or close to 0. Hence, predictive probabilities and betting probabilities are very different. In general, you should not use predictive probabilities as your betting probabilities.[31]

9 Wider Implications for Bayesianism

The arguments in the preceding sections have important upshots for both objective and subjective Bayesians, as I hope to make clear in the following two subsections.

[31] I thank a referee for pressing me to be clearer in this paragraph.

9.1 Upshots for Objective Bayesianism

According to most versions of objective Bayesianism, a probability function is rational for an agent if and only if the probability distribution is maximally uninformative while still being consistent with the agent's background information. Because most objective Bayesians have assumed that there is only one correct way of measuring the informativeness of a probability function, most objective Bayesians have accepted the Uniqueness Thesis (see Feldman (2007) and White (2005)). According to the Uniqueness Thesis (applied to the case of prior probability functions), given any body of background information, there is a unique rational prior probability function. However, if the arguments presented in this paper are sound, the Uniqueness Thesis, as stated, is clearly false and can only be salvaged if it is relativized to goals. Thus, a version of the Uniqueness Thesis consistent with the arguments presented in this paper is as follows: given any body of background information, and given a fixed goal, there is a uniquely rational prior probability function.

However, modifying the Uniqueness Thesis in this way makes apparent the second consequence for objective Bayesians: if the arguments that have been presented are sound, then objective Bayesians must apparently admit that pragmatic factors systematically influence which prior it is rational to use.

9.2 Upshots for Subjective Bayesianism

Whereas the upshots for objective Bayesians are, I think, relatively clear, the upshots for subjective Bayesians are likely to be more controversial. In contrast to objective Bayesians, subjective Bayesians do not think there are substantial rational

requirements that agents' probability distributions need to satisfy. Rather, an agent's probability distribution is supposed to accurately reflect the agent's epistemic state. Hence, for subjective Bayesians, the construction of a prior is not a search for the rationally ideal prior probability function; instead, it is the search for a probability distribution that will faithfully capture the agent's actual opinions. Since agents do not literally have probability functions in their heads, the epistemic state of the agent must somehow be *translated* into a probability function, either by the agent herself or by others. But how this translation exercise is to be solved will in general depend on the goals of the agent.

This is perhaps most easily seen in cases where you want to represent probabilistically a lack of opinion. Suppose, for example, that you are trying to determine which probability distribution most faithfully represents your opinions regarding the bias of some coin, and suppose, moreover, that you consider yourself completely uninformed and unopinionated, so that you would like your probability distribution over the possible biases of the coin to reflect your lack of an opinion. According to the calculation in Section 7, the probability distribution that is maximally unopinionated and that therefore most accurately reflects your epistemic state is relative to whether you are in the learning situation or the decision situation. If you are in the learning situation, the Jeffreys prior is the most faithful probabilistic representation of your lack of an opinion, but if you are in the decision situation, the flat prior more faithfully represents your epistemic state.

Of course, it is possible that you are in both the learning situation and in the decision situation simultaneously with respect to a single partition of hypotheses.

In that case, both probability distributions will be accurate representations of your epistemic state, but the two probability distributions should be used for different purposes. The predictive probability distribution – appropriate in the decision situation – should be used and updated (given evidence) whenever your goal is to use the partition of hypotheses to predict the future. But the learning probability distribution should be used and updated (given evidence) when you are interested in identifying the true hypothesis in the partition. If you have both goals at the same time, both probability distributions should be used. Note that your epistemic state is the same in both situations – you are completely unopinionated. But how you should best represent your lack of an opinion over the set of hypotheses probabilistically depends on why you care about the set of hypotheses.

More generally, suppose you consider yourself both epistemically risk-averse and empirically-minded and that you therefore want your epistemic state to be as unopinionated as possible given objective background information, such as, e.g., publicly available frequency data. Naturally, you will want your probability distribution to accurately reflect your epistemic risk-averseness. According to the results in Section 6, you will need to take into account your goals when you are deciding how to translate your epistemic state into a probability distribution, because whether a probability distribution counts as unopinionated given background information can only be determined once a goal has been specified. Thus the upshots we saw for objective Bayesians also carry over to at least some agents, namely those agents who see themselves as epistemically risk-averse.

10 Conclusion

I will end by briefly summarizing what I take to be the main novel contributions and conclusions of the paper. First, I have argued that the decision situation and the learning situation require different confirmation measures in order to accurately quantify the informational impact that a piece of evidence has on the probability of a hypothesis. Thus, I have argued for a version of "confirmation measure pluralism." Second, I have shown how various information measures may reasonably be derived from confirmation measures, and I have shown that how opinionated a probability distribution is for an agent therefore depends on whether the agent is in the decision situation or in the learning situation. Thus, I have also argued for a kind of "information measure pluralism." Finally, I have argued that the goal-relative nature of information has important upshots for both objective and subjective Bayesians. Most importantly, objective Bayesians must concede that whether a probability distribution is rational is partly determined by pragmatic factors, and subjective Bayesians must similarly concede that pragmatic factors sometimes partly determine which probability distribution most accurately represents an agent's epistemic state.

A Derivations of (6.5) and (6.6)

The first goal is to show that $\mathrm{Inf}_{lr}(p) = -\sum p(H_i) \log p(H_i)$ under the condition that the posterior mass converges on some H_i as $n \to \infty$, for any imagined sequence E^n of evidence. In other words, for any E^n, we require that there exists an H_i such that $\lim_{n\to\infty} P(H_i|E^n) = 1$. To avoid clutter, I will suppress n in the notation henceforth.

Now, definition (6.4) with $c = lr$ yields,

$$\mathrm{Inf}_{lr}(p) = \lim_{n\to\infty} \sum_E \sum_i \log \frac{p(H_i|E)}{p(H_i)} p(H_i, E) \tag{A.1}$$

$$= \lim_{n\to\infty} \sum_E \sum_i \log p(H_i|E)p(H|E)p(E) - \lim_{n\to\infty} \sum_E \sum_i \log p(H_i)p(H_i, E) \tag{A.2}$$

$$= \sum_i \lim_{n\to\infty} \sum_E p(E)\log p(H_i|E)p(H|E) - \sum_i \log p(H_i)p(H_i) \tag{A.3}$$

For each term of the form $p(E)\log p(H_i|E)p(H|E)$, by assumption, either $p(H_i|E) \to 1$ as $n \to \infty$, in which case $p(E)\log p(H_i|E)p(H_i|E) \to 0$; or else $p(H_i|E) \to 0$, in which case $p(E)\log p(H_i|E)p(H_i|E) \to 0$ again.[32] Thus,

$$\mathrm{Inf}_{lr}(p) = -\sum_i \log p(H_i)p(H_i) \tag{A.4}$$

Which was the first thing to be proven. Note that no assumptions were made about the sequence of evidence in the above derivation. This shows that the derivation does not depend on any such assumptions.

Now suppose that we instead plug $c = d$ into definition (6.4). Then the calculation becomes:

[32]This latter limit can be shown by an application of l'Hopital's rule.

$$\mathrm{Inf}_d(p) = \lim_{n\to\infty} \sum_E \sum_i \big[p(H_i|E) - p(H_i)\big] p(H_i, E) \tag{A.5}$$

$$= \lim_{n\to\infty} \sum_E \sum_i p(H_i|E) p(H_i, E) - \lim_{n\to\infty} \sum_E \sum_i p(H_i) p(H_i, E) \tag{A.6}$$

$$= \lim_{n\to\infty} \sum_E \sum_i p(H_i|E)^2 p(E) - \sum_i p(H_i)^2 \tag{A.7}$$

$$\tag{A.8}$$

By assumption, there is a unique term of the sum $\sum_i p(H_i|E)^2$ such that $p(H_i|E) \to 1$ as $n \to \infty$; moreover, for all of the other members of the sum, $p(H_i|E) \to 0$. Therefore, as $n \to \infty$, the entire sum $\sum_i p(H_i|E)^2$ converges to 1. Consequently,

$$\mathrm{Inf}_d(p) = \lim_{n\to\infty} \sum_E p(E) - \sum_i p(H_i)^2 \tag{A.9}$$

$$= 1 - \sum_i p(H_i)^2 \tag{A.10}$$

$$\tag{A.11}$$

References

Ahmed, Arif (2012), "Push the Button." *Philosophy of Science*, 79, 386–95.

Berger, James O., José M. Bernardo, and Dongchu Sun (2009), "The Formal Definition of Reference Priors." *The Annals of Statistics*, 37, 905–938.

Bernardo, José M. (1979a), "Expected Information as Expected Utility." *The Annals of Statistics*, 7, 686–690.

Bernardo, José M. (1979b), "Reference Posterior Distributions for Bayesian Inference." *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 113–147.

Bernardo, José M. (1981), "Reference Decisions." *Symposia Mathematica*, XXV, 85–94.

Crupi, Vincenzo and Katya Tentori (2013), "Confirmation as Partial Entailment: A Representation Theorem in Inductive Logic." *Journal of Applied Logic*, 11, 364–372.

Crupi, Vincenzo and Katya Tentori (2014), "Measuring Information and Confirmation." *Studies in the History and Philosophy of Science*, 47, 81–90.

DeGroot, Morris H. (1962), "Uncertainty, Information, and Sequential Experiments." *The Annals of Mathematical Statistics*, 33, 404–419.

Eells, Ellery (1991), *Probabilistic Causality*. Cambridge: Cambridge University Press.

Feldman, Richard (2007), "Reasonable Religious Disagreements." In *Philosophers Without Gods* (L. Anthony, ed.), 194–214, Oxford: Oxford University Press.

Gibbard, Allan (2007), "Rational Credence and the Value of Truth." In *Oxford Studies in Epistemology* (Tamar Szabo Gendler and James Hawthorne, eds.), OUP Oxford.

Hawthorne, James (manuscript), "A Better Bayesian Convergence Theorem." Manuscript.

Horowitz, Sophie (2014), "Immoderately Rational." *Philosophical Studies*, 167, 41–56.

Jaynes, E. T. (1989), "Where Do We Stand on Maximum Entropy?" In *Papers on Probability, Statistics and Statistical Physics* (Roger Rosenkrantz, ed.), volume 158 of *Synthese Library*, 210–314, Springer Netherlands.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffrey, Richard (1983), *The Logic of Decision*, second edition. Cambridge University Press, Cambridge.

Jeffreys, Harold (1946), "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186, 453–461.

Joyce, James (2009), "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief." In *Degrees of Belief* (Franz Huber and Christoph Schmidt-Petri, eds.), Synthese.

Kullback, Solomon and Richard Leibler (1951), "On Information and Sufficiency." *Annals of Mathematical Statistics*, 22, 79–86.

Lewis, David (1981), "Causal Decision Theory." *Australasian Journal of Philosophy*, 59, 5–30.

Oddie, Graham (1997), "Conditionalization, Cogency, and Cognitive Value." *British Journal for the Philosophy of Science*, 48, 533–541.

Pearl, Judea (2009), *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge: Cambridge University Press.

Rényi, Alfréd (1961), "On Measures of Entropy and Information." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 547–561.

Savage, Leonard J. (1954), *The Foundations of Statistics*. New York: Dover Publications.

Seidenfeld, Teddy (1986), "Entropy and Uncertainty." *Philosophy of Science*, 53, 467–491.

Shannon, Claude E. (1948), "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27, 379–423.

Simpson, Edward H. (1949), "Measurement of Diversity." *Nature*, 163, 688–688.

Sprenger, Jan (2012), "The Renegade Subjectivist: Jose Bernardo's Objective Bayesianism." *Rationality, Markets and Morals*, 3, 1–13.

Stevens, Stanley S. (1946), "On the Theory of Scales of Measurement." *Science*, 103, 577–80.

Tsallis, Constantino (1988), "Possible Generalization of Boltzmann-Gibbs Statistics." *Journal of Statistical Physics*, 52, 479–487.

van Enk, Steven J. (2014), "Bayesian measures of confirmation from scoring rules." *Philosophy of Science*, 81, 101–113.

White, Roger (2005), "Epistemic Permissiveness." *Philosophical Perspectives*, 19, 445–459.

Williamson, Jon (2010), *In Defence of Objective Bayesianism*. Oxford University Press.