16
BAYESIANISM AND INFORMATION

Michael Wilde and Jon Williamson

Introduction

In epistemology, Bayesianism is a theory about rational degrees of belief that makes use of the mathematical theory of probability. There is disagreement among Bayesians, however, about which norms govern rational degrees of belief. In this chapter, we first provide an introduction to three varieties of Bayesianism: strictly subjective Bayesianism, empirically based subjective Bayesianism, and objective Bayesianism. Then we discuss how one might appeal to information theory in order to justify the norms of objective Bayesianism.

Bayesianism

Consider the following epistemological question. Given one's body of evidence, should one believe a particular proposition? For example, given that one's body of evidence includes the proposition that the outcome of the roll of the die is three, should one believe that the outcome is also odd? One theory says that one should believe all and only those propositions that follow from one's body of evidence. Call this the classical theory of rational belief. The classical theory maintains that one should believe that the outcome of the roll of the die is odd, if one's body of evidence includes that the roll of the die lands three. This is because it follows from one's body of evidence that the outcome of the roll of the die is odd, if one's body of evidence includes that the roll of the die lands three.

Formally speaking, let ω be an elementary outcome and Ω be a set of mutually exclusive and collectively exhaustive elementary outcomes. In the case of rolling a fair die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then, propositions can be identified with subsets, $F \subseteq \Omega$, of elementary outcomes. For example, the proposition that the die lands odd is identified with the set of elementary outcomes at which the die lands odd, i.e., $odd = \{1, 3, 5\}$. One's body of evidence includes a proposition only if one's evidence rules out possible outcomes at which that proposition is false. For instance, if one's body of evidence includes the proposition that the outcome of the roll of the die is three, then one's evidence eliminates outcomes inconsistent with that proposition, i.e., $\{1, 2, 4, 5, 6\}$. Then, a proposition follows from one's body of evidence if and only if the set of elementary outcomes consistent with one's evidence is a subset of the proposition. To continue the example, if one's body of evidence includes the proposition that the outcome of the roll of the die is *three* = $\{3\}$, then it follows from one's body of evidence that the outcome is *odd*, since $\{3\} \subseteq \{1, 3, 5\}$.

There is another distinct epistemological question. Given one's body of evidence, *how strongly* should one believe a particular proposition? For example, given that one's body of evidence includes that the outcome of the roll of the die is odd, how strongly should one believe that the outcome is three?

The classical theory of rational belief is silent on questions about rational degrees of belief. Bayesianism, however, attempts to address such questions, by making use of the mathematical theory of probability. In particular, it introduces a function on propositions that satisfies the axioms of the probability calculus. (See Chapter 2 for more on the axioms of the probability calculus.) $P_{\mathcal{E}}$ is the probability function which gives the probability of each proposition on one's current body of evidence $\mathcal{E}$. It is called the *prior* probability function on one's evidence, since it gives the relevant probabilities before some novel piece of evidence has been learned. In the case of rolling a fair die, one natural prior probability function gives $P_{\mathcal{E}}(1) = P_{\mathcal{E}}(2) = P_{\mathcal{E}}(3) = P_{\mathcal{E}}(4) = P_{\mathcal{E}}(5) = P_{\mathcal{E}}(6) = 1/6$. Once again, one's body of evidence works to rule out certain basic outcomes, so that if one later learns the proposition that the die lands odd, so that one's new body of evidence is $\mathcal{E}'$ formed by adding *odd* to $\mathcal{E}$, then one would need to update the probability function to $P_{\mathcal{E}'}(1) = P_{\mathcal{E}'}(3) = P_{\mathcal{E}'}(5) = 1/3$. Bayesianism says that one's degree of belief in a proposition should match the probability of the proposition given one's evidence. In this case, given that one's body of evidence includes the proposition that the outcome of the die lands odd, one should believe that the die lands three to degree one-third.

Bayesianism is an attractive theory of rational belief. First, it is a simple and natural generalization of the classical theory of rational full belief. To begin to see this, note that Bayesianism preserves the classical theory: $P_{\mathcal{E}}(F) = 1$, if F follows from one's body of evidence; $P_{\mathcal{E}}(F) = 0$, if F is inconsistent with one's body of evidence. In addition, Bayesianism extends the classical theory by giving an account of the cases in between these two extremes. Having specified a probability function, a measure is provided of how close one's body of evidence comes to entailing or ruling out some proposition. Second, Bayesianism can help itself to all the many results in the mathematical theory of probability. Thus it has powerful resources to call upon in providing a theory of rational belief. Third, Bayesianism accommodates many of our intuitive judgements about rational degrees of belief. For instance, it accommodates the judgement that hypotheses are confirmed by their successful predictions, and that hypotheses are better confirmed if they make more surprising successful predictions. Bayesianism can account for all these intuitions and more (Howson and Urbach, 2006, pp. 91–130).

But why should one's degrees of belief be probabilities? The usual answer is to provide a Dutch book argument. Loosely speaking, if one is willing to bet the farm in exchange for a penny if some proposition is true, then it would seem that one has a high degree of belief in that proposition. On the other hand, if one is only willing to bet a penny in exchange for a farm if the proposition is true, then it seems that one has a low degree of belief in the proposition. This leads naturally to a betting interpretation of belief, according to which one's degree of belief in any given proposition is identified with one's willingness to bet on that proposition, i.e., with one's betting quotient for that proposition. The Dutch book argument then aims to show that betting quotients that are not probabilities are irrational in a certain sense. In particular, it is assumed that betting quotients susceptible to the possibility of sure loss are irrational. It is then shown that betting quotients are probabilities if and only if they are not susceptible to the possibility of sure loss. Thus one's betting quotients are rational only if they are probabilities. Given the betting interpretation of belief, one's degrees of belief are rational only if they are probabilities.

Thus the proponents of Bayesianism all tend to agree that degrees of belief should be representable by a probability function. That is, in order to be rational, one's degrees of belief must be probabilities. This is sometimes called the probability norm for degrees of belief. But is meeting the probability norm enough for one's degree of belief to be rational?

Strictly subjective Bayesians maintain that the probability norm is enough. They allow that any choice of prior degrees of belief is rational, as long as these degrees of belief are probabilities. One advocate of strict subjectivism is Bruno de Finetti (1937). However, other Bayesians have argued that certain probabilistic degrees of belief are more appropriate than others, given one's body of evidence. For instance, in the case of rolling a die, if one's evidence is only that the die is fair, then arguably one's degrees of belief are best represented by a probability function that gives $P_{\mathcal{E}}(1) = P_{\mathcal{E}}(2) = P_{\mathcal{E}}(3) = P_{\mathcal{E}}(4) = P_{\mathcal{E}}(5) = P_{\mathcal{E}}(6) = 1/6$. It looks like strictly subjective Bayesianism cannot account for the more or less objective nature of rational belief.

This has led some proponents of Bayesianism to advocate *empirically based subjective Bayesianism*. Richard Jeffrey (2004) and Colin Howson (2000) may be considered advocates of empirically based subjectivism.

Empirically based subjective Bayesians argue that meeting the probability norm is not sufficient for one's degrees of belief to be rational. Instead of allowing an arbitrary selection of a prior probability function, they argue that one's prior probability function should also be calibrated to evidence of physical probabilities. This is sometimes called the calibration norm. For instance, let $P^{\star}$ be the physical probability function. Then, if one's evidence includes that $P^{\star}(1) = P^{\star}(2) = P^{\star}(3) = P^{\star}(4) = P^{\star}(5) = P^{\star}(6) = 1/6$, one should choose as one's prior the probability function that gives $P_{\mathcal{E}}(1) = P_{\mathcal{E}}(2) = P_{\mathcal{E}}(3) = P_{\mathcal{E}}(4) = P_{\mathcal{E}}(5) = P_{\mathcal{E}}(6) = 1/6$. More generally, where $\mathbb{E}$ is the set of probability functions constrained in this way by the empirical evidence, one should choose $P_{\mathcal{E}} \in \mathbb{E}$. Empirical constraints such as this can also be justified by betting arguments (Williamson, 2010, pp. 39–42). But what if there is absolutely no empirical evidence? Then, the selection of a prior probability function is once again unconstrained, and which degrees of belief are rational again looks like a matter of personal choice.

In light of this, some proponents of Bayesianism advocate *objective Bayesianism*. One proponent of objective Bayesianism is Jon Williamson (2010). Objective Bayesians argue that one's prior degrees of belief are rational if and only if they are probabilities calibrated with the empirical evidence, that are otherwise sufficiently non-committal with regard to elementary outcomes. What does it mean for one's degrees of belief to be non-committal? The standard answer is that one's degrees of belief commit oneself to a particular elementary outcome over another to the extent that one believes the former to a greater degree than one believes the latter. This means that one's degrees of belief are *fully* non-committal between elementary outcomes when one believes all such outcomes to the same degree. Then one's degrees of belief are *sufficiently* non-committal only if they are as close to fully non-committal as meeting the probability and calibration norms permits. In the case of rolling a die when one has no evidence either way that the die is fair, the selection of a prior probability function is not a matter of personal choice according to objective Bayesianism. Instead, the sufficiently non-committal prior probability function gives $P_{\mathcal{E}}(1) = P_{\mathcal{E}}(2) = P_{\mathcal{E}}(3) = P_{\mathcal{E}}(4) = P_{\mathcal{E}}(5) = P_{\mathcal{E}}(6) = 1/6$.

To sum up, there is no consensus among proponents of Bayesianism about which norms govern rational degrees of belief. In particular, there is disagreement regarding the following core norms:

> **Probability**: Degrees of belief should be probabilities;
> **Calibration**: Degrees of belief should be calibrated to evidence of physical probabilities;
> **Equivocation**: Degrees of belief should be sufficiently non-committal.

There are three main lines of thought. Strict subjectivists advocate only the probability norm. They hold that one's prior degrees of belief are rational if and only if they satisfy the axioms of the probability calculus. For the strict subjectivist, then, one's prior degrees of belief are a matter of personal choice, so long as they are probabilities. Empirically based subjectivists go further by holding that one's prior degrees of belief are rational if and only if they are probabilities that are appropriately constrained by the empirical evidence; in particular, they hold that these degrees of belief should be calibrated to physical probabilities, insofar as one has evidence of them. That is, they advocate both the probability and the calibration norm, but not the equivocation norm. Objective Bayesians go further still, holding that one's prior degrees of belief are rational only if they are also sufficiently non-committal.

How is this disagreement to be settled? Some have looked to information theory to provide an answer.

Information theory and Bayesianism

The core norms of Bayesianism have been justified by appealing to information theory. In this section we provide an introduction to this line of research.

Information theory grew out of the pioneering work of Claude Shannon. Given finitely many elementary outcomes $\omega \in \Omega$, Shannon (1948, §6) argued that the uncertainty as to which outcome occurs is best measured by the entropy of the probabilities of the outcomes. The entropy of a probability function P is defined by:

$$H(P) = -\sum_{\omega \in \Omega} P(\omega) \log P(\omega)$$

Entropy increases the more evenly the probability is spread out over the possible outcomes; it is minimal when probability 1 is concentrated on one of the outcomes and maximal when each outcome has the same probability. Shannon argued that entropy is the only measure of uncertainty that (i) is continuous (small changes in probability lead to small changes in entropy), (ii) increases as the number of possible outcomes increases, and (iii) sums up in the right way when a problem is decomposed into two sub-problems.

Edwin Jaynes applied the information-theoretic notion of entropy to the problem of choosing a prior probability function (Jaynes, 1957). Jaynes suggested that one should choose a prior function that is maximally non-committal with respect to missing information, i.e., a function that is compatible with what information is available, but which is maximally uncertain with regard to questions about which no information is available. Applying Shannon's notion of entropy, this means that one should choose as one's prior, a probability function, from all those functions compatible with available evidence, that has maximum entropy. If $\mathbb{E}$ is the set of probability functions that are compatible with available evidence, Jaynes' maximum entropy principle says that one should choose $P_{\mathcal{E}} \in \text{maxent}\, \mathbb{E}$, where

$$\text{maxent}\,\mathbb{E} \overset{\text{df}}{=} \{P \in \mathbb{E} : H(P) \text{ is maximized}\}.$$

The maximum entropy principle can be understood as an explication of the equivocation norm advocated by objective Bayesians.

The question remains as to why one's prior should be maximally non-committal. What advantage is there to adopting a non-committal prior?

Topsøe (1979) provided an interesting line of argument. Suppose the loss incurred by believing ω to degree $P_{\mathcal{E}}(\omega)$ when ω turns out to be the true outcome is logarithmic:

$$L(\omega, P_{\mathcal{E}}) = -\log P_{\mathcal{E}}(\omega).$$

Thus the loss is zero when ω is fully believed, but increases exponentially as degree of belief $P(\omega)$ decreases towards 0. Suppose $P^\star$ is the true chance function, so that one's expected loss is

$$\sum_{\omega \in \Omega} P^\star(\omega) L(\omega, P_{\mathcal{E}}) = -\sum_{\omega \in \Omega} P^\star(\omega) \log P_{\mathcal{E}}(\omega).$$

All one knows is that $P^\star \in \mathbb{E}$. Thus one's worst-case expected loss is

$$\sup_{P^\star \in \mathbb{E}} - \sum_{\omega \in \Omega} P^\star(\omega) \log P_{\mathcal{E}}(\omega).$$

It turns out that, as long as $\mathbb{E}$ is non-pathological (e.g., if $\mathbb{E}$ is closed and convex), the prior probability function which minimizes worst-case expected loss is just the probability function in $\mathbb{E}$ that maximizes entropy. Thus the maximum entropy principle is justified on the grounds that the resulting prior minimizes worst-case expected loss. Note that the maximum entropy principle thus construed explicates the calibration norm as well as the equivocation norm, because it says that, when evidence determines just that the chance function $P^\star \in \mathbb{E}$, one should take $P_{\mathcal{E}}$ to be *a function in* $\mathbb{E}$ – i.e., a calibrated probability function – that has maximum entropy.

One question immediately arises: why should loss be logarithmic? Topsøe, appealing to Shannon's work on communication and coding, suggested that loss is logarithmic if it is the cost incurred by transmitting the results of an observation. Grünwald and Dawid (2004) recognized that this limits the scope of the maximum entropy principle to certain communication problems. They generalized Topsøe's justification to cope with other loss functions, leading to a generalized notion of entropy which depends on the loss function in operation, and to a generalized maximum entropy principle which says that one should choose a prior probability function in $\mathbb{E}$ that maximizes generalized entropy.

This approach remains rather limited to the extent that one needs to know the true loss function in order to choose one's prior probability function, because one needs to know which generalized entropy function is to be maximized. Often, however, one wants to choose a prior in advance of knowing the uses to which one's beliefs will be put and the losses (or gains) which might result. Thus one needs to identify a *default* loss function – a loss function that encapsulates what one might presume about one's losses, in the absence of information about the true loss function. Williamson (2010, pp. 64–65) put forward four principles that constrain this default loss function:

L1: Fully believing the true outcome may be presumed to lead to zero loss.
L2: One can presume that loss strictly increases as $P_{\mathcal{E}}(\omega)$ decreases from 1 towards 0.
L3: The presumed loss $L(\omega, P_{\mathcal{E}})$ depends on $P_{\mathcal{E}}(\omega)$ but not on $P_{\mathcal{E}}(\omega')$, for other outcomes ω'.

L4: If one decomposes a problem into two sub-problems which are presumed to be unrelated, then the total loss can be presumed to be the sum of the losses incurred on each of the two sub-problems.

It turns out that the default loss function must be logarithmic if it is to satisfy these four principles. Thus one can apply Topsøe's original justification of the maximum entropy principle in the (rather typical) case in which one does not know the true loss function, and one can apply Grünwald and Dawid's generalization if one does happen to know the true loss function.

A second concern arises for this kind of justification of the maximum entropy principle. Recall that the probability norm is usually justified by means of the Dutch book argument: degrees of belief must be probabilities if one is to avoid exposing oneself to the possibility of sure loss, i.e., $L(\omega, P_{\mathcal{E}}) > 0$ for all ω. There are two respects in which this argument does not cohere well with the above argument for the maximum entropy principle. First, in the Dutch book argument the objective is to avoid sure loss, rather than minimize worst-case expected loss. Second, the notion of loss invoked by the Dutch book argument is not logarithmic loss. Instead,

$$L(\omega, P_{\mathcal{E}}) = (P_{\mathcal{E}}(\omega) - 1)S(\omega) + \sum_{\omega' \neq \omega} P_{\mathcal{E}}(\omega')S(\omega')$$

where the $S(\omega)$, $S(\omega')$ are stakes chosen by an adversary, which may be positive or negative and which may depend on one's belief function $P_{\mathcal{E}}$.

It is clearly less than satisfactory if the justification of one tenet of objective Bayesianism – the probability norm – is incompatible with the justification of the others, namely the calibration and equivocation norms, cashed out in terms of the maximum entropy principle. In view of this, Landes and Williamson (2013) attempted to reconcile the Bayesian norms, by extending the justification of the maximum entropy principle so as to justify the probability norm at the same time. The justification of the maximum entropy principle outlined above presumes the probability norm, since it shows that the *probability function* that minimizes worst-case expected loss is the probability function in $\mathbb{E}$ which maximizes entropy. What is needed is to show that the *belief function* that minimizes worst-case expected loss is the function in $\mathbb{E}$ with maximum entropy; that it is in $\mathbb{E}$ implies that the prior belief function is a probability function, i.e., it implies the probability norm.

Thus Landes and Williamson (2013) extend the concepts of loss and expected loss to handle losses incurred by an arbitrary belief function B, which is not necessarily a probability function, in order to show that belief functions which are not probability functions expose one to sub-optimal worst-case expected loss. The main issue is that in the original notion of expected loss,

$$\sum_{\omega \in \Omega} P^{\star}(\omega) L(\omega, P_{\mathcal{E}}) = -\sum_{\omega \in \Omega} P^{\star}(\omega) \log P_{\mathcal{E}}(\omega),$$

one considers a single partition of outcomes, namely the partition of elementary outcomes $\omega \in \Omega$. This is appropriate if one assumes the probability norm from the outset, as the probability of any proposition $F \subseteq \Omega$ is determined by the probability of the elementary outcomes,

$$P_{\mathcal{E}}(F) = \sum_{\omega \in F} P_{\mathcal{E}}(\omega),$$

i.e., the probabilities of the elementary outcomes tell us everything about the probability function. For example, if the elementary outcomes correspond to outcomes of a roll of a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$, then $P(odd) = P_{\mathcal{E}}(\{1,3,5\}) = P_{\mathcal{E}}(1) + P_{\mathcal{E}}(3) + P_{\mathcal{E}}(5)$. However, it is no longer appropriate to consider only the partition of elementary outcomes when we do not assume the probability norm from the outset, because the degree of belief in F may be unrelated to the degrees of belief in the elementary outcomes that make up F. Thus we need to consider all partitions π of Ω when defining expected loss:

$$\sum_{\pi} g(\pi)\sum_{F\in\pi} P^{\star}(F)L(F,B) = -\sum_{\pi} g(\pi)\sum_{F\in\pi} P^{\star}(F)\log B(F).$$

Here g is a weighting function that provides each partition π with a weight that determines the extent to which that partition contributes to the expectation. Entropy may be defined similarly:

$$H_{g}(B)\overset{\mathrm{df}}{=}-\sum_{\pi} g(\pi)\sum_{F\in\pi} B(F)\log B(F).$$

This gives a generalized notion of entropy that depends on the weighting function. (Note that this generalization is different to the generalized entropies of Grünwald and Dawid (2004).) The case of standard entropy corresponds to the weighting g_{Ω} which gives weight 1 to the partition $\{\{\omega\}:\omega\in\Omega\}$ of elementary outcomes and weight 0 to every other partition. It turns out that, as long as the weighting function g is inclusive in the sense that for each proposition F, g gives positive weight to some partition that contains F, the belief function that minimizes worst-case expected loss is the probability function in $\mathbb{E}$ that maximizes entropy. This gives an integrated justification of the probability norm and the maximum entropy principle, albeit with respect to a generalized notion of entropy that is defined in terms of g. It is suggested in Landes and Williamson (2013) that the standard notion of entropy stands out as uniquely appropriate among the generalized entropies if we impose language invariance as a further desideratum: i.e., that one's prior belief function should not depend on the language in which the elementary outcomes are expressed.

Conclusion

In the first half of this chapter we introduced Bayesianism as a theory about rational degrees of belief. On the way, we noted some of the arguments in favor of Bayesianism, but we also noted a difficulty. If probabilities are given an interpretation in terms of rational degrees of belief, and rational degrees of belief are largely a matter of personal choice, it begins to look as if rational belief is a matter of personal opinion. However, this fails to do justice to the more or less objective nature of rational belief. To resolve this difficulty, the Bayesian usually attempts to reduce the element of personal choice by advocating further constraints on rational degrees of belief, namely the calibration and equivocation norms. The issue then becomes how to justify those norms. In the second half of this chapter we argued that one can appeal to information theory in order justify the Bayesian norms. The standard information-theoretic justification of the equivocation norm is incompatible with the standard Dutch book justification of the probability norm. However, recent results show that the norms of objective Bayesianism can receive a unified information-theoretic justification.

Acknowledgements

This research was supported by UK Arts and Humanities Research Council grants which fund the projects *From objective Bayesian epistemology to inductive logic* and *Evaluating evidence in medicine*.

Further reading

For more on the mathematical theory of probability see Chapter 2 of this volume. For an introduction to the philosophy of probability, see Gillies (2000). Gillies gives also a clear exposition of Dutch book arguments (2000, pp. 53–65). Bayesianism is named after the Reverend Thomas Bayes, who lived and preached in Kent (Bayes, 1764). One popular introduction and defense of Bayesianism is Howson and Urbach (2006). For a critical evaluation of Bayesianism see Earman (1992). Edwin Jaynes' magnum opus is Jaynes (2003). One recent defense of objective Bayesianism is Williamson (2010).

References

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In Kyburg, H. and Smokler, H. (eds), *Studies in Subjective Probability*, pages 53–118.

Earman, J. (1992). *Bayes or Bust*. Cambridge, MA: MIT Press.

Gillies, D. (2000). *Philosophical Theories of Probability*. Abingdon: Routledge.

Grünwald, P. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4): 1367– 1433.

Howson, C. (2000). *Hume's Problem: Induction and the Justification of Belief*. Oxford: Oxford University Press.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court, 3rd edition.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, 106(4): 620–630.

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

Jeffrey, R. (2004). *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press.

Landes, J. and Williamson, J. (2013). Objective Bayesianism and the maximum entropy principle. *Entropy*, 15(9): 3528–3591.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27: 379–423 and 623–656.

Topsøe, F. (1979). Information theoretical optimization techniques. *Kybernetika*, 15: 1– 27.

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