Economics and Philosophy, **35** (2019) 321–338
doi:10.1017/S0266267118000305 First published online 14 January 2019
cambridge.org/eap

BAYESIAN OCKHAM'S RAZOR AND NESTED MODELS

Bengt Autzen*

Abstract: While Bayesian methods are widely used in economics and finance, the foundations of this approach remain controversial. In the contemporary statistical literature Bayesian Ockham's razor refers to the observation that the Bayesian approach to scientific inference will automatically assign greater likelihood to a simpler hypothesis if the data are compatible with both a simpler and a more complex hypothesis. In this paper I will discuss a problem that results when Bayesian Ockham's razor is applied to nested economic models. I will argue that previous responses to the problem found in the philosophical literature are unsatisfactory and develop a novel reply to the problem.

KEYWORDS: Ockham's razor, Bayesianism, model, simplicity, prior probability

1. INTRODUCTION

Bayesian methods are widely used in economics and finance. For instance, Bayesian approaches have become popular when it comes to evaluating dynamic stochastic general equilibrium (DSGE) models with empirical data (Smets and Wouters 2007). Similarly, Bayesian methods are invoked to help international corporations that sell their goods abroad to manage the risk of foreign exchange rate exposure that they incur at the time they repatriate the proceeds of their sales (Bos *et al.* 2000). And finally, Bayesian approaches are used to estimate financial risk measures, such as Value-at-Risk, employed by banks and other financial institutions (Hoogerheide and van Dijk 2010).

* Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Ludwigstrasse 31, 80539 München, Germany. Email: b.autzen@lmu.de. URL: www.mcmp.philosophie.uni-muenchen.de/people/faculty/autzen_bengt/index.html

Ockham's razor, sometimes referred to as the principle of parsimony or simplicity, is the idea that simpler hypotheses are more likely to be true. Ockham's razor has made two prominent inroads into Bayesian methodology. Early attempts by Wrinch and Jeffreys (1921) suggested that in Bayesian inference simpler hypotheses should get assigned higher prior probability. One problem with Wrinch and Jeffreys's proposal, which has become known as Jeffreys's simplicity postulate (Jeffreys 1931), is that it does not offer a justification for Ockham's razor. Rather the simplicity postulate only interprets Ockham's razor in a particular Bayesian way. Modern Bayesians therefore prefer a different view on the relationship between Ockham's razor and Bayesian methodology. In the contemporary statistical literature Bayesian Ockham's razor (BOR) refers to the observation that Bayesian inference will automatically assign greater likelihood to a simpler hypothesis if the data are compatible with both a simpler and a more complex hypothesis (Jefferys and Berger 1992).

In this paper I will focus on a particular issue resulting when BOR is applied to nested models.[1] The problem of comparing a set of nested models arises in a number of economic applications. For instance, in studies on asset price forecasting, a baseline model postulating a random walk of asset prices (or a random walk with drift) is compared to a number of more complex models that include additional predictors of asset price movements.[2] In a further example involving nested economic models, Kriwoluzky and Stoltenberg (2016) compare a baseline model of a cashless New Keynesian economy with a number of more complex DSGE models that include additional features, such as the indexation of prices to past inflation, the formation of consumption habits and the inclusion of money in a household's utility function. In particular, they compare the baseline model with indexation, here referred to as the indexation model, with their most complex model, here referred to as the complete model, that allows for habit formation, indexation and money. Importantly, the indexation model is nested in the complete model since the former is a special case of the latter. In line with BOR, the simpler indexation model has both a greater likelihood and a greater posterior probability than the more complex complete model in their Bayesian analysis.

There is, however, something puzzling about this result. The indexation model is assumed to be nested in the complete model, that is, the indexation model only describes a subset of possibilities regarding the data generating mechanism when compared to the complete model. However, if an event A is a subset of event B, then the probability of A,

[1] For a recent philosophical discussion of BOR in the context of non-nested models, see Sober (2015).

[2] Examples include Hong and Lee (2003), who study exchange rates, and Sarno *et al.* (2005), who study interest rates.

$P(A)$ cannot be larger than the probability of B, $P(B)$ for any probability measure P. As such, it seems somewhat miraculous that the simpler, nested indexation model has a greater posterior probability (i.e. 0.82) than the more complex complete model (i.e. 0.18). Phrased differently, BOR seems to lead to an incoherent probability assignment when applied to nested models.

This problem is analogue to an objection raised by Popper (1959) against Jeffreys's simplicity postulate. To recapitulate, the simplicity postulate requires that simpler hypotheses are to be assigned a greater prior probability than more complex hypotheses in a Bayesian analysis. Popper denies the possibility that when a simpler hypothesis is nested in a more complex hypothesis, the simpler hypothesis can have higher probability than the more complex one. While the focus of Popper's critique directed at Jeffreys's simplicity postulate has been on the role of prior probabilities of hypotheses in a Bayesian analysis, the same conceptual problem arises in the discussion of the role of posterior probabilities of nested models in BOR.

One might be tempted to think that even though this problem has attracted the attention of philosophers of science, it does not have any impact on scientific debates. Turning to a biological context, however, Templeton (2010) critically notices that in some prominent phylogeographic studies nested models have higher posterior probability than their encompassing models. For instance, Fagundes *et al.* (2007) compare three models of human evolution. First, the out-of-Africa replacement model asserts that an expanding African population completely replaced Eurasian populations with no admixture. Second, the assimilation model allows for potential admixture between the African population and the Eurasian populations. The degree of admixture in the assimilation model is measured by the parameter m that can vary between 0 (i.e. no genetic input from archaic Eurasians) and 1 (i.e. no genetic input from the expanding African population into Eurasians). And finally, the gene flow model allows for gene flow between African and Eurasian populations. Since the assimilation model reduces to the replacement model by fixing the parameter m at zero, the replacement model is nested in the assimilation model. Fagundes *et al.* calculate a posterior probability of 0.781 for the replacement model and a posterior probability of 0.001 for the assimilation model. Templeton (2010: 6377) concludes: 'It is mathematically and logically impossible for [the replacement model] to have greater probability than [the assimilation model].'

In this paper I aim to make sense of the apparent Bayesian incoherence found in the economic and statistical literature. The structure of the paper is as follows. Section 2 illustrates the problem of applying BOR to nested models by means of a simple coin tossing example. Section 3 reviews some previous responses to the problem found in the philosophical literature.

Section 4 develops an alternative response based on the notion of a Bayesian model. Section 5 discusses a difficulty associated with the use of Bayesian models in scientific inference. Section 6 concludes.

2. A TOY EXAMPLE

In order to state the problem associated with the application of BOR to nested models as succinctly as possible, I will abstract from the details of the DSGE models discussed by Kriwoluzky and Stoltenberg (2016) and consider a more manageable example that is structurally similar. In particular, I will introduce two statistical models representing different hypotheses about a coin. The first hypothesis, denoted as FAIR, asserts that the coin is fair. More formally, the hypothesis asserts that the coin tosses are independent and identically distributed and that the probability of the coin landing 'Heads' in a single toss is equal to 0.5. According to FAIR, the number of 'Heads' in n tosses is then described by the Binomial distribution $B(n, 0.5)$. In order to simplify the example, the number of tosses is fixed at $n = 10$. FAIR is then given by

$$(\text{FAIR})\{B(10, p_1) : p_1 = 0.5\}.$$

In contrast, the second hypothesis, denoted as HEDGE, asserts that the coin is fair or biased towards heads or biased towards tail. As such the second hypothesis can be represented by the following set of probabilistic hypotheses:[3]

$$(\text{HEDGE})\{B(10, p_2) : 0 \leq p_2 \leq 1\}.$$

In line with standard treatments of model selection theory (e.g. Burnham and Anderson 2002), FAIR is considered as simpler than HEDGE since FAIR has no adjustable parameters (i.e. parameter p_1 is fixed at 0.5) while HEDGE has one.

In order to illustrate the mechanics of BOR, I will examine the posterior probabilities of FAIR and HEDGE given the data D denoting the sequence of 10 coin tosses $(H, T, H, T, H, T, H, T, H, T)$. The posterior probability of, say, FAIR results from applying Bayes's theorem:

$$P(\text{FAIR}|D) = \frac{P(D|\text{FAIR})P(\text{FAIR})}{P(D)}.$$

Since I am mainly interested in the relationship between the posterior probabilities of FAIR and HEDGE, it suffices to assess the ratio of these

[3] Note that HEDGE is not a tautology. It asserts that the coin tosses are independent and identically distributed with parameter p_2 describing the probability of 'Heads' in a single coin toss.

two probabilities:

$$\frac{P(\text{FAIR}|D)}{P(\text{HEDGE}|D)} = \frac{P(D|\text{FAIR})P(\text{FAIR})}{P(D|\text{HEDGE})P(\text{HEDGE})}.$$

In order to evaluate the ratio of posterior probabilities of FAIR and HEDGE, one has then to calculate the likelihood and prior probability of each model. Suppose for the moment that the prior probability of FAIR equals 0.4 and the prior probability of HEDGE equals 0.6.[4] The likelihood of FAIR is straightforwardly given by 0.5^{10} as there is only a single Binomial distribution included in this model. In contrast, the likelihood of HEDGE is an average for Bayesians and therefore requires assigning a prior distribution to the adjustable parameter p_2. More formally, the ('marginal') likelihood $P(D\,|\,\text{HEDGE})$ is given by $\int_0^1 P(D|B(10, p_2)) f(p_2|\text{HEDGE}) dp_2$, where $f(p_2\,|\,\text{HEDGE})$ denotes the conditional prior probability density of parameter p_2. In line with common treatments of BOR suppose that p_2 follows a 'flat' prior such as the uniform distribution on the unit interval $[0, 1]$.[5] Given these assumptions, the ratio of model posterior probabilities is given by

$$\frac{P(\text{FAIR}|D)}{P(\text{HEDGE}|D)} \approx \frac{0.000977 * 0.4}{0.000361 * 0.6} > 1.$$

That is, even though HEDGE has a higher prior probability than FAIR, FAIR has a higher posterior probability than HEDGE.

Underlying BOR is the idea that a more complex model must spread its likelihood more thinly over the data space than a simpler model. This can be illustrated by means of the likelihoods of FAIR and HEDGE, where the likelihood ratio in the previous numerical example is approximately given by $\frac{9.77}{3.61}$ in support of the simpler model FAIR. Depending on the assignment of prior probabilities to the two models, this can result in the simpler model being more probable a posteriori. In the current example, the simpler model FAIR not only has higher likelihood than HEDGE but also has a larger posterior probability given the data.

Before considering possible replies to the problem, it is worth noting that there is nothing special about using a simpler model that consists only of a point hypothesis regarding the value of the unknown coin tossing parameter. As an alternative to FAIR consider, for instance, the following model FAIR-ε given by

$$(\text{FAIR-}\epsilon)\{B(10, p_1) : 0.5 - \epsilon \leq p_1 \leq 0.5 + \epsilon\}$$

[4] This choice of prior probabilities reflects Popper's remark that simpler, nested models cannot have greater prior probability than more complex models.

[5] Calculating the likelihood $P(D\,|\,\text{HEDGE})$ then involves evaluating the integral $\int_0^1 p_2^5(1 - p_2)^5 dp_2$.

with a small, fixed $\varepsilon > 0$. For a suitable choice of data FAIR-ε will have both larger likelihood and larger posterior probability than HEDGE. Also note that in the modified example a slightly different notion of simplicity is invoked according to which a model is simpler if its adjustable parameter has a more constrained range of possible numerical values than the adjustable parameter of its competitor. While the adjustable parameter p_2 of HEDGE ranges between 0 and 1, the parameter p_1 of FAIR-ε ranges over a much smaller interval that also forms a subset of [0, 1]. FAIR-ε is therefore simpler than HEDGE based on this second reading of simplicity.

MacKay (2003: 349–350) discusses a similar example in which he compares three nested models, denoted as H_1, H_2 and H_3. Each model has one parameter w but assigns a different prior range to that parameter. The model H_3 is the most complex model since it assigns the broadest prior range, while model H_1 is the simplest model since it goes along with the most narrow prior range. It is assumed that all three models have equal prior probability. In the example a data set assigns a higher likelihood to model H_2 than the more complex model H_3. The simplest model H_1 has the smallest likelihood since it can only achieve a poor fit to the observed data. Since all three models have equal prior probability, the model H_2 of intermediate complexity has the largest posterior probability given the data.

The latter examples demonstrate that the canonical philosophical response to Popper's critique of the simplicity postulate cannot accommodate BOR in its full generality. Howson (1988) replies to Popper's argument by restricting the domain of the simplicity postulate to non-nested models. More specifically, Howson essentially asks us to compare FAIR with the model given by

$$\left\{B(10, p_2) : 0 \leq p_2 \leq 1, p_2 \neq \frac{1}{2}\right\}$$

rather than HEDGE. Since HEDGE and the model given by the set $\{B(10, p_2) : 0 \leq p_2 \leq 1, p_2 \neq \frac{1}{2}\}$ have the same marginal likelihood, this re-description seems to solve the conceptual problem while leaving the numerical model probabilities unaffected.[6] This proposal, however, becomes problematic when comparing models such as (FAIR-ε) with HEDGE. Here, HEDGE cannot be re-described in the way suggested by Howson without affecting its likelihood calculation. Since proponents of BOR, such as MacKay (2003), are genuinely interested in comparing models with non-trivially overlapping parameter ranges, restricting the Bayesian analysis to models with non-overlapping parameter ranges amounts to substantively changing the inference problem. In the next

[6] Here, I rely on the fact that taking out a singleton value in the parameter range does not affect the likelihood calculation for a continuous probability distribution.

https://doi.org/10.1017/S0266267118000305 Published online by Cambridge University Press

section, I will turn to some more recent philosophical responses that aim to take the conceptual sting out of BOR.

3. PREVIOUS RESPONSES

3.1. Theories as generators

In their discussion of Bayesian hierarchical modelling Henderson *et al.* (2010) respond to the view that simple, nested models cannot have higher probability than more complex models. Henderson *et al.* phrase their discussion in terms of theories rather than models such that they take issue with the view that logically stronger theories cannot have higher probability than logically weaker theories. Their reply can be best discussed by means of one of their examples.

Curve fitting refers to the problem of finding the curve that best represents the relationship between two variables X and Y in the light of some observed data. Doing so not only involves choosing a particular curve, such as $y = 2x + 1$ or $y = x^2 + x + 2$, but also making a decision about the functional form of the curve (e.g. linear function, quadratic function etc.) that is most appropriate for the task at hand. Henderson *et al.* understand the curve fitting problem as an instance of Bayesian hierarchical modelling. At the lowest level of the hierarchical model one finds particular curves such as $y = 2x + 1$. At the next level of the hierarchy theories are distinguished by means of the maximum degree of the polynomial used to establish particular curves in the lower level hypothesis space. In particular, Henderson *et al.* consider the two higher level theories Poly_1, referring to polynomials with maximum degree 1, and Poly_2, referring to polynomials with maximum degree 2.

Henderson *et al.* contrast their understanding of these higher level theories with the traditional, 'set-based' understanding of higher level theories in the curve fitting problem according to which higher level theories are characterized by sets of curves. For instance, the set of all linear curves of the form $y = a_1x + a_0$ (with $a_0, a_1 \in \mathbb{R}$) constitutes one such higher level theory. A further candidate theory is given by the set of all quadratic functions of the form $y = a_2x^2 + a_1x + a_0$ (with $a_0, a_1, a_2 \in \mathbb{R}$). At the heart of Henderson *et al.*'s account is the idea that higher level theories 'generate' lower level theories meaning that higher level theories provide 'a rule or recipe specifying constraints on the construction of [the lower level theory]' (Henderson *et al.* 2010: 176). In the curve fitting problem Poly_1 and Poly_2 each generate curves of a particular functional form. However, Poly_1 and Poly_2 are not to be identified with the set of curves they generate according to the generative view on higher level theories. Importantly, Poly_1 and Poly_2 are not considered as standing in a subset relationship. Rather, these higher level theories are seen to be

https://doi.org/10.1017/S0266267118000305 Published online by Cambridge University Press

mutually exclusive and thereby to block any concerns about the simpler theory having larger probability than the more complex theory.

How to unpack the idea of higher level theories as generators of lower level theories? And more specifically, how to understand the models Poly_1 and Poly_2? A natural reading identifies Poly_1 with the proposition that the curve describing the relationship between variables X and Y has a linear form. Similarly, Poly_2 is then to be identified with the proposition that the curve describing the relationship between variables X and Y has a quadratic form. These are very general readings of Poly_1 and Poly_2 that provide constraints on the construction of particular curves, such as $y = 2x + 1$ and $y = x^2 + x + 2$. Based on this reading, however, Poly_2 is logically entailed by Poly_1 since every linear curve is a quadratic curve but not vice versa. As such, Poly_1 cannot have a higher probability than Poly_2. Importantly, this line of reasoning does not rely on invoking a set-based approach to state the higher level theories in the curve fitting problem. While probability theory is typically formulated in a set-theoretic framework following Kolmogorov's foundational work (Kolmogorov 1933), probability theory can be equivalently phrased in terms of propositions rather than sets (Jaynes 2003).

Given that the proposed reading of Poly_1 and Poly_2 does not provide a satisfactory reply to the apparent incoherence of BOR, the question remains of how to cash out the idea of higher level theories as generators of lower level theories. The proponent of this account needs to say more about the content of the generative narratives associated with Poly_1 and Poly_2 as well as how these narratives prevent Poly_2 being logically entailed by Poly_1. I will come back to this question in Section 4, where I offer an answer on what additional content might be needed in the generative story of higher level theories. For the moment, I conclude that the generative view on theories in its current form does not offer a satisfactory answer on how BOR applies to nested models.

3.2. Relabelling

Romeijn and van de Schoot (2008) provide an alternative proposal for addressing the challenge resulting from applying BOR to nested models. They propose that nested models, such as FAIR and HEDGE, can be relabelled so that assigning larger probability to the nested model does not violate the probability calculus. Romeijn and van de Schoot (2008: 353) write:

> nothing prevents us from using two distinct sets of hypotheses ... which are different from a set-theoretical view by virtue of being labelled differently, even while they have exactly the same likelihood functions over the data.

https://doi.org/10.1017/S0266267118000305 Published online by Cambridge University Press

To illustrate, consider again the model FAIR. FAIR consists of the single Binomial distribution denoted as $B(10, 0.5)$. Now, relabel the hypothesis in the set. For instance, let us denote the same probability distribution as $B*(10, 0.5)$. Romeijn and van de Schoot suggest that in virtue of the different label, the model consisting only of the single Binomial distribution $B*(10, 0.5)$, here denoted as FAIR$*$, and the model FAIR form two disjunct sets of hypotheses. As a result, FAIR$*$ can be assigned a different probability than FAIR. In particular, FAIR$*$ can be assigned a larger probability than HEDGE without violating the probability calculus, or so they claim.

Romeijn and van de Schoot propose that the models FAIR and FAIR$*$ are non-nested in virtue of the fact that the hypotheses in each set are labelled differently. While I agree that, technically speaking, the models FAIR and FAIR$*$ are non-nested, the question arises of whether FAIR and FAIR$*$ constitute an adequate mathematical representation of the inference problem at hand. Relabelling does not alter the fact that both FAIR and FAIR$*$ are empirically equivalent. The two statistical models consist of pairwise identical probabilistic hypotheses in the sense that these hypotheses assign the same probabilities to the events in the underlying probability space. As a result it is not clear why one would reasonably assign different probabilities to these two models. Put differently, simple relabelling seems to amount to a case of mislabelling. Assigning different labels to sets of pairwise identical probability distributions, gives a misleading picture regarding the possible hypotheses about the data generating mechanism.

In a more recent article developing a Bayesian account of abductive inference Romeijn (2013) elaborates how models, again understood as sets of probability distributions, relate to empirical data from a Bayesian perspective. Romeijn (2013: 430) writes

> A central point in this paper is that models whose distinction is theoretical [i.e. the models consist of probability distributions assigning identical probabilities to the data] may still differ in empirical content, because of the priors we define over them. We will look at models ... that differ theoretically in the sense specified above, but that are associated with different stories concerning the data generating system. Such stories motivate different priors over the models in question, and these priors again lead to a different empirical content for the two models. The data may be used to choose between the models in virtue of their association with different priors.

Romeijn uses the example of a normal and a magical coin to illustrate his view. A normal coin is most probably fair, that is, has a probability of 'Heads' that is close to 0.5. In contrast, the magical coin is most probably biased, that is, has a probability of landing 'Heads' that is close to 0 or 1.

The model of the normal coin, here denoted as NORMAL, is given by

$$(\text{NORMAL})\{H_\theta : \theta \in [0, 1]\},$$

while the model of the magical coin, here denoted as MAGIC, is given by

$$(\text{MAGIC})\{H^*_\theta : \theta \in [0, 1]\},$$

where H_θ (H^*_θ) asserts that the probability of the coin landing 'Heads' is equal to θ. (Again, it is assumed that the coin tosses are independent and identically distributed.) The fact that the normal coin is most probably fair while the magical coin is most probably biased is then captured by means of distinct prior probability distributions assigned to the parameter θ in the two models.

If the notion of a model as a set of probabilistic hypotheses about a data-generating mechanism is maintained, Romeijn's position leads to an unclear and rather awkward terminology. It is unclear how the two sets of hypotheses NORMAL and MAGIC can represent two different data-generating mechanisms, that is, the normal and the magical coins. They only do so in the light of the different priors assigned to the adjustable parameters of the models. Furthermore, Romeijn (2013: 430) seems to suggest that two models consisting of hypotheses with identical (classical) likelihood functions simultaneously have the *same* and *different* empirical content. That is, two models have the same empirical content because the hypotheses in the models have pairwise identical likelihood functions. At the same time these models can have different empirical content in virtue of the different priors assigned to the adjustable parameters of the two models even though these priors are not part of the models.

While I disagree with the particulars of Romeijn's (and van de Schoot's) account, I still think there is an important lesson to be learned. I concur with Romeijn (2013: 430) that 'data may be used to choose between the models in virtue of their association with different priors'. This observation suggests to include the prior of adjustable parameters *into* the notion of a model. By including the prior of the adjustable parameter into the model, it becomes clear how models that contain pairwise identical probabilistic hypotheses about the data-generating mechanism can have different empirical content. The next section will follow up this idea.

4. BAYESIAN MODELS

The notion of a statistical model typically invoked in the philosophical literature understands models as sets of probabilistic hypotheses about the data generating process (e.g. Forster and Sober 1994). More formally, a statistical model describes the generating process of data y in terms of probability distributions or, more conveniently, probability densities $p(y \mid \theta)$ characterized by reference to the parameter $\theta \in \Theta$. A statistical model is then given by $\{p(y\,|\,\theta) : \theta \in \Theta\}$. There is, however, a different reading of a statistical model found in the contemporary Bayesian literature. For Bayesian statisticians, such as Box (1980), Spiegelhalter *et al.* (2002) and Gelman *et al.* (2004), a (Bayesian) model consists of both a set of probabilistic hypotheses about the data-generating process and a prior probability distribution on its adjustable parameters. Schematically, a Bayesian model is given by $(\{p(y\,|\,\theta) : \theta \in \Theta\}, p(\theta))$, with $p(\theta)$ denoting the prior probability density of θ.

By adopting the notion of a Bayesian model, the fair coin hypothesis can be construed as follows:

$$(\text{FAIR}^{**})(\{B(10, p_1) : p_1 = 0.5\}, \mu(p_1)),$$

with $\mu(p_1)$ denoting the (trivial) prior probability measure on parameter p_1 that assigns all the probability mass to the singleton value $p_1 = 0.5$. Formally, probability measure $\mu(p_1)$ is defined on the event space containing only the element $p_1 = 0.5$ (i.e. $\Omega = \{p_1 = 0.5\}$) and the σ-algebra consisting of the empty set and the event $\{p_1 = 0.5\}$. More interestingly, the hypothesis that the coin is either fair or biased is to be formalized as follows:

$$(\text{HEDGE}^{**})(\{B(10, p_2) : 0 \leq p_2 \leq 1\}, \nu(p_2)),$$

with $\nu(p_2)$ denoting the prior probability distribution on parameter p_2 in the HEDGE model. In the numerical example discussed earlier a uniform prior over the unit interval was assumed.

Adopting the Bayesian notion of a model has important consequences for the present discussion. Since FAIR$**$ is not a subset of HEDGE$**$, there are no formal reasons to think that FAIR$**$ cannot have a larger probability than HEDGE$**$. As such, the aforementioned concerns raised against BOR disappear by modifying the notion of a statistical model used in Bayesian inference. Similarly, the present proposal can make sense of Kriwoluzky and Stoltenberg's Bayesian evaluation of DSGE models. By including the prior probability distributions of their adjustable parameters into the indexation and the complete model, these DSGE models turn out to be mutually exclusive. As a result, there are no mathematical reasons speaking against the simpler indexation model having a greater posterior probability than the more complex complete model in the light of the data.

Having introduced and subsequently criticized previous philosophical responses to the problem of how to apply BOR to nested models, it is natural to ask how my proposal fares in the light of the aforementioned criticism. To begin with, consider the generative view on theories proposed by Henderson *et al.* (2010). I argued that the generative view needs a more detailed account of the generative story associated with higher level theories. Based on the notion of a Bayesian model outlined in this section, I suggest that the additional content of higher level theories (or models) is to be found in the prior probability distribution over the adjustable model parameters. Generating a lower level theory then amounts to specifying a particular prior over the model parameters.

This can be illustrated in the curve fitting context. Suppose the higher level theory asserting that the curve describing the relationship between variables X and Y is of the form $y = a_1x + a_0$ (with $a_0, a_1 \in \mathbb{R}$) is conjoined with information about the prior probability distribution over the parameters a_0 and a_1. Doing so amounts to specifying a Bayesian model. Now, one can choose a particular prior that assigns all probability mass to the numerical parameter values $a_0 = 1$ and $a_1 = 2$. As a result, the higher level theory gives rise to or 'generates' the particular curve $y = 2x + 1$. Hence, the notion of a Bayesian model promises to provide a more substantive account of how higher level theories generate lower level theories.

Finally, let us revisit Romeijn's writings on how to apply BOR to nested models. Setting aside Romeijn and van de Schoot's relabelling approach, I objected that Romeijn's treatment of the prior of adjustable parameters leads to an unclear and rather awkward terminology. In particular, I took issue with the fact that two models consisting of identical classical likelihood functions $p(y \mid \theta)$ can simultaneously have the same and different empirical content based on Romeijn's account. The difference in empirical content results from the influence of the prior of the adjustable parameters in a Bayesian analysis. In particular, the prior of the adjustable parameters will influence the numerical value of the marginal likelihood calculated in a Bayesian analysis. By including the prior of the adjustable parameters into the model, there is nothing mysterious about why two models can have different empirical content even though they also contain identical classical likelihood functions. As such, I consider the notion of a Bayesian model as a natural further development of Romeijn's ideas.

5. REPRESENTATION

While adopting the notion of a Bayesian model blocks the concerns regarding the coherence of BOR, methodological questions remain. Here, I will focus on what I consider the most pressing issue, that is, the question of whether Bayesian models adequately represent scientific hypotheses. Consider again the coin tossing example discussed earlier. While the prior probability measure μ in the model FAIR$**$ follows directly from the meaning of the fair coin hypothesis, matters are less clear in the case of the hypothesis that the coin is either fair or biased. The latter hypothesis does not naturally induce a prior probability measure on parameter p_2 needed to specify the model HEDGE$**$; the choice of the uniform prior over the unit interval in the previous numerical example was nothing more than a convenient choice. Matters would be different if, say, HEDGE$**$ were to represent the hypothesis that the coin had been selected by a random mechanism that samples the parameter describing the probability of 'Heads' from a uniform distribution. In that case the prior probability distribution of parameter p_2 would be prescribed by the hypothesis under consideration.

A reader familiar with the subjective (or 'Bayesian') interpretation of probability might point out that there is nothing problematic with choosing a 'flat' prior (or any other coherent prior) for an adjustable model parameter since prior probabilities merely reflect the subjective degrees of belief of an agent. Matters are more complicated, however, when it comes to the assignment of prior probabilities to the adjustable parameter(s) in a Bayesian model. Since every Bayesian model has a distinct prior probability distribution for its parameter(s), it is generally unclear whose degrees of belief these priors are to be identified with. For instance, comparing the models FAIR$**$ and HEDGE$**$ involves the assignment of different prior probability distributions to the probability of a coin landing 'Heads'. While FAIR$**$ assigns a Dirac measure centred at 0.5 to the probability that the coin will land heads, HEDGE$**$ assigns the uniform distribution to this probability. As a result it is difficult to understand these two different prior probability distributions as the degrees of a belief that the coin will land 'Heads' of a *single* agent.

To consider a further example, Romeijn's normal and magical coins find natural representations by means of a Bayesian model. According to the definition, a normal coin is most probably fair. As such, it can be represented by means of a Bayesian model that includes a prior probability distribution of the model parameter assigning most of its weight to a small interval including the numerical value 0.5. In contrast, the magical coin is most probably biased, that is, has a probability of landing 'Heads' that is close to 0 or 1. As such, the magical coin can be represented by means of a Bayesian model that includes a prior probability distribution of the model parameter assigning most of its weight to the ends of the interval [0,1]. Even though the normal and the magical coin do not go along with unique prior probability distributions of their model parameter, their definitions impose clear constraints on the choice of prior probability distributions of these model parameters.

Having considered a number of different coin tossing models, it is time to take stock. At the one end of the spectrum we find the fair-coin hypothesis that can be directly represented by means of the Bayesian model FAIR$**$. At the other hand of the spectrum, we find the coin tossing model HEDGE$**$. Here, it is not clear how to assign a prior to the adjustable parameter of the Bayesian model HEDGE$**$ supposed to represent the hypothesis that the coin is either fair or biased. In between

https://doi.org/10.1017/S0266267118000305 Published online by Cambridge University Press

these two extremes we find the normal and magical coins discussed by Romeijn. Both the definitions of the normal and the magical coin refer to a prior on the probability of 'Heads' in a single coin toss. As such, the normal and the magical coin can be naturally represented by means of a Bayesian model. In contrast to the prior in FAIR$**$, however, the priors of the adjustable parameters in these models are not uniquely specified but allow for some choice on behalf of the investigator.

What are the implications for the role of Bayesian model comparison in scientific inference? Based on this assessment a constrained role for Bayesian model comparison suggests itself. While some scientific hypotheses may be represented by means of a Bayesian model in a natural way, other hypotheses may not. In the former case the Bayesian approach offers adequate inferential tools, in the latter case its usefulness for scientific inquiry is less clear. Phrased differently, if a scientific hypothesis is not sufficiently detailed to specify a Bayesian model, then it is unclear whether the assessment by means of Bayes's theorem is epistemologically relevant.

Returning to macroeconomics, one might wonder whether DSGE models can be naturally represented by means of Bayesian models. At first sight some sceptical thoughts come to mind. Consider, for instance, the DSGE models analysed by Kriwoluzky and Stoltenberg (2016). Their models incorporate a staggered price-setting mechanism originally proposed by Calvo (1983). Suppose that firms cannot change their prices freely each time period. More specifically, assume that when a firm sets a (nominal) price there is a constant probability $1 - \alpha$ that the firm is able to adjust its price in any period. The Calvo parameter α will govern the average duration between price changes. If the Calvo parameter is small, then the firms can update their prices frequently. If the Calvo parameter is large, then it is very probable that a firm will be stuck for a long time with whatever price it chooses today. When assigning a prior probability distribution to the Calvo parameter, Kriwoluzky and Stoltenberg (2016: 336) opt for a 'relatively uninformative' beta distribution with mean 0.5 and standard deviation of 0.2. Does this choice indicate that the prior for the Calvo parameter merely reflects the ignorance of the investigators rather than any substantive claim about how price setting works? Following this line of thought, one might be tempted to conclude that DSGE models are more similar to the model HEDGE$**$ than the fair coin hypothesis (or Romeijn's normal and magical coins) and, hence, that Bayesian model comparison is not suitable for analysing DSGE models.

Such a conclusion, however, would be premature. Taking a closer look at the motivation for assigning a prior to the Calvo parameter in DSGE models, reveals that there are typically good empirical reasons for its choice. Since the Calvo parameter describes the average period between price changes, one can calibrate the parameter by looking at data on the average duration between price adjustments. Bils and Klenow (2004) find that this number is between 6 months and one year. Taking into account Bils and Klenow's empirical study, Smets and Wouters (2007) assume the Calvo parameter to be typically around 0.5, thereby suggesting an average length of price contracts of half a year. As a mathematical description of the empirical distribution of the lengths of price contracts, they choose the beta distribution with mean 0.5 and standard deviation 0.15. This choice of prior distribution is mirrored by Kriwoluzky and Stoltenberg, who closely follow Smets and Wouters in their assignment of prior probabilities to model parameters.[7]

More generally, the prior distributions of the adjustable parameters of DSGE models are largely chosen to reflect empirical observations. Del Negro and Schorfheide (2008) identify three main categories of empirical regularities that are used to motivate the choice of these prior distributions. First, economists use pre-sample data to design prior distributions. For instance, investigators use pre-1982 observations to select a prior when the Bayesian analysis is restricted to post-1982 data. Second, data from other countries are used for prior specification. For instance, a prior for a DSGE model for the Euro zone is specified based on data from the US economy. And third, economists employ observations that are concurrent to the analysed data sample but excluded from the DSGE model. As an example, Del Negro and Schorfheide refer to micro-level data that are informative about features such as labour supply behaviour or price rigidities described in the DSGE model.

The general picture that emerges from the discussion of DSGE models is conducive to the approach taken in this paper. Since there are regularly empirical arguments given to motivate the design of the prior distributions of DSGE parameters, there is good reason to believe that DSGE models can be captured naturally by means of Bayesian models. Returning to the taxonomy of coin tossing models, DSGE models are reminiscent of Romeijn's normal and magical coins. Even though the priors of adjustable parameters are typically well motivated by the hypotheses at hand, there is generally no unique prior probability distribution for DSGE model parameters such as the Calvo parameter. Since the empirical information does not uniquely specify the assignment of a prior distribution, there is some room for choice on behalf of the investigator. This does not mean, however, that these prior distributions are completely subjective in nature. To the contrary, they

[7] To be precise, Kriwoluzky and Stoltenberg choose a beta distribution with a slightly larger standard deviation compared to Smets and Wouters.

capture what economists currently know about a feature or process of the economy.

It is informative to relate this view on Bayesian model comparison to other philosophical positions arguing for a limited role of Bayesian methods in scientific methodology due to the role of the priors in Bayesian inference. Sober (2008), for instance, argues that Bayesianism does not provide an adequate account of scientific inference in cases where there is no objective basis for the prior probabilities of scientific hypotheses. Sober (2008: 26–27) contrasts the case of medical hypotheses such as a patient having tuberculosis where objective prior probabilities can be assigned by reference to frequency data with the case of scientific theories such as Darwin's theory of natural selection where no such objective assignment is possible. While in the former case the Bayesian approach to scientific inference provides the right answers from an epistemological perspective, in the latter case it does not or so it is argued. The difference to the position on Bayesian model comparison outlined in this paper, however, is that I am concerned with the question of whether scientific hypotheses can be adequately represented by means of a Bayesian model. The question of how to assign prior probabilities to these models is a further question that I have not commented on but that needs to be addressed in any Bayesian analysis.

To state this difference more clearly, consider a scenario in which a scientific hypothesis can be adequately represented by mean of a Bayesian model but does not have an objective prior along the lines demanded by Sober. DSGE models with an empirically motivated prior of the Calvo parameter, for instance, can be adequately represented by means of a Bayesian model.[8] It is much less clear, however, whether an objective prior can be assigned to these models. Phrased differently, while authors such as Sober would, in all likelihood, agree with the demand to employ an empirically grounded prior for the Calvo parameter, they would go one step further in demanding the assignment of an empirically grounded prior to the DSGE model. The latter step, however, is not required in my view on Bayesian model comparison. As a result, the constrained form of Bayesian inference proposed in this paper is, in an important sense, less restrictive than the limited Bayesianism with its emphasis on objective priors of scientific hypotheses put forward by Sober and others, which is typically discussed in the philosophical literature.

[8] In order to simplify the discussion, I set aside the variety of different parameters used in DSGE models and focus only on the Calvo parameter. I do not think that this restriction threatens the generality of the treatment since, as Del Negro and Schorfheide (2008) point out, economists typically aim to provide empirically grounded priors for the parameters of DSGE models.

6. CONCLUSION

In this paper I discussed a problem that results when BOR is applied to nested models in the economic literature. I pointed out that based on the orthodox reading of a statistical model, a simpler, nested model cannot coherently have a higher posterior probability than a more complex model. I argued that previous responses to the problem found in the philosophical literature are unsatisfactory before developing a novel reply to the problem invoking the notion of a Bayesian model. The approach outlined in this paper considers the prior of adjustable parameters as part of a (Bayesian) model. From this perspective choosing an empirically (and perhaps theoretically) informed prior is an integral part of specifying a model in Bayesian inference. While I offered a partial vindication for the application of BOR to nested models as well as contemporary work on Bayesian inference in economics, such as Kriwoluzky and Stoltenberg's Bayesian analysis of DSGE models, there is room for a more systematic inquiry into the process of model building in economics. As such, the present paper should be seen as opening a methodological discussion rather than providing the final word on the subject matter.

ACKNOWLEDGEMENTS

I would like to thank Ken Binmore, Jason Konek, Richard Pettigrew, Samir Okasha and Jan-Willem Romeijn as well as the anonymous referees of the journal for helpful comments on earlier versions of the manuscript. Awards from the British Academy Postdoctoral Fellowship Scheme and the Alexander von Humboldt Foundation are gratefully acknowledged.

REFERENCES

Bils, M. and P.J. Klenow. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112: 947–985.

Bos, C.S., R.J. Mahieu and H.K. van Dijk. 2000. Daily exchange rate behaviour and hedging of currency risk. *Journal of Applied Econometrics* 15: 671–696.

Box, G.E.P. 1980. Sampling and Bayesian inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A* 143: 383–430.

Burnham, K. and D. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer.

Calvo, G.A. 1983. Staggered price setting in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Del Negro, M. and F. Schorfheide. 2008. Forming priors for DSGE models and how it matters for nominal rigidities. *Journal of Monetary Economics* 55: 1191–1208.

Fagundes, N.J.R., N. Ray, M. Beaumont, S. Neuenschwander, F.M. Salzano, S.L. Bonatto and L. Excoffier. 2007. Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences USA* 104: 17614–17619.

Forster, M. and E. Sober. 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science* 45: 1–37.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin. 2004. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall.
Henderson, L., N.D. Goodman, J.B. Tenenbaum and J.F. Woodward. 2010. The structure and dynamics of scientific theories: a hierarchical Bayesian perspective. *Philosophy of Science* 77: 172–200.
Hong, Y. and T.H. Lee. 2003. Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* 85: 1048–1062.
Hoogerheide, L. and H.K. van Dijk. 2010. Bayesian forecasting of Value at Risk and Expected Shortfall using adaptive importance sampling. *International Journal of Forecasting* 26: 231–247.
Howson, C. 1988. On the consistency of Jeffreys's simplicity postulate. *Philosophical Quarterly* 38: 68–83.
Jaynes, E.T. 2003. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.
Jefferys, W.H. and J.O. Berger. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80: 64–72.
Jeffreys, H. 1931. *Scientific Inference*. Cambridge: Cambridge University Press.
Kolmogorov, A.N. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.
Kriwoluzky, A. and C.A. Stoltenberg. 2016. Nested models and model uncertainty. *Scandinavian Journal of Economics* 118: 324–353.
MacKay, D. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.
Popper, K.R. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.
Romeijn, J.W. 2013. Abducted by Bayesians? *Journal of Applied Logic* 11: 430–439.
Romeijn, J.W. and R. van de Schoot. 2008. A philosopher's view on Bayesian evaluation of informative hypotheses. In *Bayesian Evaluation of Informative Hypotheses*, ed. H. Hoijtink, I. Klugkist and P.A. Boelen, 329–357. New York, NY: Springer.
Sarno, L., D.L. Thornton and G. Valente. 2005. Federal funds rate prediction. *Journal of Money, Credit and Banking* 37: 449–472.
Smets, F. and R. Wouters. 2007. Shocks and frictions in US business cycles: a Bayesian DSGE approach. *American Economic Review* 97: 586–606.
Sober, E. 2008. *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.
Sober, E. 2015. *Ockham's Razor: A User's Manual*. Cambridge: Cambridge University Press.
Spiegelhalter, D.J., N.G. Best, B.P. Carlin and A. van der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* 64: 583–639.
Templeton, A.R. 2010. Coherent and incoherent inference in phylogeography and human evolution. *Proceedings of the National Academy of Sciences USA* 107: 6376–6381.
Wrinch, D. and H. Jeffreys. 1921. On certain fundamental principles of scientific inquiry. *Philosophical Magazine* 42: 369–390.

BIOGRAPHICAL INFORMATION

Bengt Autzen is a postdoctoral fellow at the Munich Center for Mathematical Philosophy at Ludwig-Maximilians-Universität München. His research interests lie in the philosophy of biology, the philosophy of statistics and the philosophy of social sciences.