The Objectivity of Subjective Bayesian Inference

February 29, 2016

Abstract

Subjective Bayesianism is a major school of uncertain reasoning and statistical inference. Yet, it is often criticized for an apparent lack of objectivity. By and large, these criticisms come in three different forms. First, the lack of constraints on prior probabilities, second, the entanglement of statistical evidence and degree of belief, third, the apparent blindness to bias in experimental design. This paper argues that the above criticisms pertain to some specific senses of scientific objectivity. On a more comprehensive understanding of that concept, the criticisms fail to hold water; in fact, subjective Bayesianism may even be more objective than some of its competitors.

1 Introduction

Subjective Bayesianism is a major school of uncertain reasoning and statistical inference that is steadfastly gaining popularity. It is based on the subjective interpretation of probability and describes how prior degree of belief in a scientific hypothesis is updated to posterior degree of belief. This creates a straightforward connection between the mathematical theory of probability and the epistemological question of which hypothesis is confirmed by the evidence. What makes subjective Bayesianism subjective is the honest representation of personal degree of belief in terms of probabilities, rather than the pursuit of uniquely rational degrees of belief (Bernardo and Smith, 1994; Howson and Urbach, 2006).

Subjective Bayesianism should be distinguished from an "anything goes" position where degrees of belief are only constrained by probabilistic coherence: the rational degree of belief to make a certain observation, given a certain parameter value (e.g., the chance of observing both heads and tails in two i.i.d. tosses of a fair coin) may

be uniquely determined. Similarly, we may have theoretical reasons to believe that not all probability distributions over the value of an unknown parameter are equally adequate. Subjective Bayesians may agree in their probabilistic judgments surprisingly often, but it is characteristic of their view that there is no principled way of resolving disagreement.

Being objective is essential for any method of scientific inference, including subjective Bayesianism: objectivity contributes to the reliability of research, conveys an image of epistemic authority and strengthens our trust in science. The 2009 "Climategate" affair and the recent "replication crisis" in psychology (i.e., the widespread failure to replicate experimental results due to various forms of bias, see Makel et al., 2012), illustrate how an apparent lack of objectivity weakens trust in scientific findings. Subjective Bayesian inference is often criticized for an apparent lack of objectivity: "a notion of probability as personalistic degree of belief [...], by its very nature, is not focused on the extraction and presentation of evidence of a public and objective kind" (Cox and Mayo, 2010, 298). This view is echoed in writings of well-known statisticians and philosophers of science such as Fisher (1956), Mayo (1996), Popper (2002) and Senn (2011).

The objectivity-related criticisms of subjective Bayesian inference come, by and large, in three different forms. First, the lack of constraints on prior probabilities, second, the entanglement of statistical evidence and degree of belief, third, the apparent blindness to bias in experimental design. In the light of these objections, one is tempted to conclude that Bayesian inference cannot produce objective knowledge, is not suitable for scientific communication and is therefore inferior to frequentist inference.

This paper addresses the above objections (Section 2–4). The defense concedes that Bayesian inference is—like any method of inference—not fully objective in every possible sense, but that it promotes various important senses of objectivity. Moreover, claims that it is less objective than other inferential frameworks can be rebutted. The final section concludes and embeds our discussion into a broader debate about objectivity in science (Section 5).

2 Objection 1: The Choice of the Prior Distribution

Bayesian inference is based on prior probability distributions. Assume that you are interested in assessing a hypothesis $H_0 \in \mathcal{H}$. You represent your prior belief in H_0 by

means of a probability distribution over the entire space of hypotheses $\mathcal{H}$. Assume further that for any $H \in \mathcal{H}$, your data D follow a definite probability distribution $p(D|H)$. Then, your posterior degree of belief in the null hypothesis H_0 can be calculated by the formula

$$p(H_0|D) = \frac{p(H_0)p(D|H_0)}{p(D)} \tag{1}$$

where $p(D) = \sum_{H \in \mathcal{H}} p(D|H)p(H)$ is the marginal probability of data D. On the basis of the posterior probability $p(H_0|D)$, a Bayesian can form a theoretical judgment about H_0 or make a practical decision. For example, if H_0 is the hypothesis that a new medical drug is not more efficacious than a placebo, and if H_0 is sufficiently probable given the data, then we will not pursue further development of the drug.

The posterior probability depends on the prior probability, and often, there is not sufficient background knowledge to establish consensus on prior probabilities. Subjective Bayesians such as Ramsey (1926) and de Finetti (1972) have stressed that in principle, *any* coherent prior probability distribution can be defended as rational. This attitude seems to jeopardize any claims to objectivity that subjective Bayesians could possibly make. What kind of epistemic warrant does a Bayesian inference still provide? After all, the choice of the prior can hide all kind of pernicious values, e.g., financial interests of the experiment sponsor. This is particularly worrying in sensitive areas such as medicine, where the need for impartial inference methods is particularly high, due to the manifest financial interests in clinical trials and the ethical consequences of wrong decisions. As the medical methodologist Lemuel Moyé writes:

> Without specific safeguards, use of Bayesian procedures will set the stage for the entry of non-fact-based information that, unable to make it through the "evidence-based" front door, will sneak in through the back door of "prior distributions". There, it will wield its influence, perhaps wreaking havoc on the research's interpretation. (Moyé, 2008, 476)

The objection claims that Bayesians can bias the final result in their preferred direction by choosing an appropriate prior. This objection is thus based on the value-free ideal that the core business of scientific reasoning, namely evaluating evidence, assessing and accepting theories, should be free of non-cognitive values and individual biases—a requirement that Bayesian inference seems to violate blatantly. Adherence to the value-free ideal has, however, in one form or another, been upheld as a trademark of scientific objectivity (e.g., Weber, 1904; Lacey, 1999; Reiss and Sprenger, 2014), and for

practitioners, it plays an even greater role due to regulatory constraints and conflicts of interests. Even if one doubts that the value-free ideal can be attained in practice—e.g., because any inference involves an implicit tradeoff of false negatives and false positives (Rudner, 1953)—, values should not be allowed to *replace* scientific evidence (Douglas, 2008, 2009). How can Bayesian inference be safeguarded against this danger?

The first defense notes that subjective opinion need not be the same as individual bias. Two medical doctors may, on the basis of their experience, give a different judgment about what might be a good therapy for a patient with a given set of symptoms. The fact that they disagree does not mean that one of them or both are biased: they may have enjoyed a different training, come from different disciplines or have different experience in dealing with those symptoms. Prior probability distributions provide a way to make explicit a judgment that is fed by individual expertise and track record. This is also a reason why many models of expert judgment and decision-making use subjective Bayesian inference—even when "objective risk assessments" or the like are required (Cooke, 1991).

The second defense notes that prior probabilities are open to rational criticism. Whenever a prior distribution is used, be its shape conventional or peculiar, the researcher should justify her particular choice and explain which considerations (theoretical and empirical ones) led her to this choice. We cannot justify an extreme posterior simply by choosing a suitably extreme prior because it is part of the Bayesian model of reasoning that also the prior needs to be justified. This is also explicit in regulations for medical trials, such as the guidelines for the use of Bayesian statistics, issued by the Food and Drug Administration of the United States:

> We recommend you be prepared to clinically and statistically justify your choices of prior information. In addition, we recommend that you perform sensitivity analysis to check the robustness of your models to different choices of prior distributions. (US Food and Drug Administration, 2010)

The above quote hints to a second requirement in Bayesian reasoning: to perform a sensitivity analysis on the choice of the prior and to check whether the main result of the research remains intact under different prior assumptions. Such an analysis also contributes to scientific objectivity in terms of "convergent objectivity" (Douglas, 2004, 2009, 2011), according to which a scientific result can claim to be objective when it is validated from different assumptions and perspectives. Checking how a variation in the prior affects variation in the results therefore contributes to drawing conclusions

which satisfy this sense of objectivity.

Finally, the third defense notes that the explicit choice of a prior distribution exposes modeling assumptions more clearly than competing paradigms. In frequentist inference, for example, such assumptions are more implicit and harder to identify. This makes it easier for the Bayesian to criticize a particular choice, contributing to scientific objectivity in the sense of a reasoning process that is transparently conducted and open to rational criticism (Longino, 1990). We will get back to this point in the final section.

The bottom line is that the choice of the prior is just like any other modeling assumption in science open to rational criticism. Indeed, if the prior were not varied and judged critically, there would be no corrective mechanism for gauging to what extent personal bias has influenced the results through the choice of the prior. But the same is true of scientific inference in general: if assumptions are not examined critically, biased results are the necessary consequence. Screwing up a subjective Bayesian analysis with a biased prior is as easy or difficult as screwing up a non-Bayesian analysis with biased modeling assumptions. Therefore, this objection is not more fearsome for Bayesians than for any other framework of inductive inference. We now move to the next objection: that Bayesians mix up belief and evidence.

3 Objection 2: Belief vs. Evidence

The second objection contends that scientific reasoning, and statistical analysis in particular, is not about assessing the subjective probability of hypotheses, but about finding out whether a certain effect is real or due to chance. The task of science is to state the objective *evidence* for the truth of the hypothesis. In this view, the Bayesian statistician commits a category mistake: she tries to answer a question that scientists are not (and should not be) interested in. Statistical reasoning is about the truth of hypotheses and should be independent of subjective plausibility judgments. Ronald A. Fisher, one of the fathers of modern statistics, forcefully articulated this view:

> Advocates of inverse probabilities [ascribing probabilities to scientific hypotheses given some data, J.S.] are forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes (Fisher, 1935, 6–7)

Royall (1997, 4) makes a similar distinction between three major questions in statistical analysis: "What should we believe?", "What should we do?" and "What is the evidence?". A good answer to one of them need not be a good answer to another question. The Bayesian answers the belief questions by providing prior and posterior probabilities, but what does a satisfactory response to the evidence question look like?

Underlying this objection is the idea of "detached objectivity" (Douglas, 2009, 459): claims to scientific knowledge should be detached from personal belief and wishful thinking. Bayesians also struggle to achieve "concordant objectivity" (Douglas, 2004, 462–463), that is, intersubjectively agreed assessments of evidence. As Quine (1992, 5) stated it: "The requirement of intersubjectivity is what makes science objective." However, the "psychological tendencies" that correspond to personal degrees of belief do not fulfil this requirement.

Many philosophers and scientists share the view that subjective Bayesian inference falls short of achieving concordant objectivity. Williamson (2007) notes that "full objectivity—i.e. a single probability function that fits available evidence" cannot be achieved in the subjective Bayesian framework. Bem et al. (2011, 718) quote a *Psychological Science* referee as saying

> I have great sympathy for the Bayesian position [...] The problem in implementing Bayesian statistics for scientific publications, however, is that such analyses are inherently subjective, by definition [...] with no objectively right answer as to what priors are appropriate. I do not see that as useful scientifically. It is unclear to me how we can have agreed upon priors for a collective such as the body of psychological researchers.

In other words, even if the priors are not contaminated by extra-scientific values (see Section 2), they still mirror individual perspectives. It is not clear how objective evidence can be extracted from them, without tying each interpretation to a particular (subjective) model.

To address this point, we study the most popular Bayesian measures of evidential support in some detail. The Bayes factor (Kass and Raftery, 1995; Lee and Wagenmakers, 2013) expresses the support for H_0 over the alternative H_1 in terms of the ratio of posterior and prior odds. Equivalently, the Bayes factor can be expressed as the ratio of (integrated) likelihood of H_0 and H_1:

$$B_{01}(D) \quad := \quad \frac{p(H_0|D)}{p(H_1|D)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{\int_{\theta\in\Theta_0} p(D|\theta)p(\theta)d\theta}{\int_{\theta\in\Theta_1} p(D|\theta)p(\theta)d\theta} \tag{2}$$

It is important to note that the Bayes factor is not affected by $p(H_0)$ and $p(H_1)$ simpliciter. For two point hypotheses H_0 and H_1, it is even fully independent of the prior probability distribution: it is just the likelihood ratio $p(D|H_0)/p(D|H_1)$, indicating how much D favors H_0 over H_1. Nothing depends on personal belief.

For composite hypothesis (e.g., $H_1 : \theta \in [a, b], \theta \neq \theta_0$), things are more complicated. The value of the Bayes factor depends on how likely the observed evidence is under the various components of H_0 and H_1, weighted with their relative prior probability. It is important to realize that this dependency is benign and not pernicious in the context of null hypothesis testing. Imagine the frequent case that we are testing the null hypothesis that a certain intervention, e.g., taking vitamin C tablets as a cure for the common flu, has no effect at all: $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$, where θ is the variable denoting the effect size. Of course, it is implausible that the effect of the vitamin C intervention is *exactly* zero: the tablets will cause a biochemical reaction in the human body even if it is negligibly small. The test aims at finding out whether we can use the null hypothesis as a simple and precise, but strictly speaking wrong idealization of a complex reality (Gallistel, 2009). In order to assess whether a finding is evidence for or against H_0, we need to know which effect sizes are plausible and clinically relevant. Only if this is clarified, we can state meaningfully that the observed results speak in favor of or against the null hypothesis.

Frequentist inference proceeds similarly. When deciding on the sample size N of an experiment, the choice of N reveals the power of an experiment for relevant alternative hypotheses (=effect sizes). An experiment that would have low power for plausible effect sizes would be misdesigned and futile. Therefore, the relative plausibility of the alternative hypotheses affects the level of evidential support. The view that degrees of belief must not play any role in assessing evidential support is taking the value-free ideal and the idea of detached objectivity one step too far. Indeed, also Douglas (2004, 460) stresses that objectivity in scientific reasoning should not imply the elimination of personal perspective; this would actually be a gross misrepresentation of how science works. Values must not play a detrimental role in inference, but this is arguably not the case in Bayesian hypothesis testing.

Regarding intersubjectivity/concordant objectivity, we have seen that the raw observations often underdetermine levels of evidence, and in such a situation, the goal of intersubjective agreement may be elusive. The Bayesian is able to differentiate such a situation from those where the evidence is agreed upon, but inconclusive.

I conclude this section with an example of how considerations pertaining to scientific objectivity may lead one to a quasi-Bayesian position on measuring evidential support. Birnbaum's Likelihood Principle (all experimental evidence about an unknown parameter θ is contained in the likelihood function $L(\theta) = p(D|\theta)$ for observed data D) is one of the cornerstones of Bayesian inference. I do not want to contribute to the debate about the Likelihood Principle, but just notice that it can be derived from the following two principles (Birnbaum, 1962):

Sufficiency Principle If $T(X)$ is a sufficient statistic for a parameter θ, that is,

$$P(X = x|T(X) = t, \theta) = P(X = x|T(X) = t) \tag{3}$$

and we observe $T(x_1) = T(x_2)$ in two separate experiments, then both experiment generate the same evidence about θ.

Conditionality Principle If a chancy trial (e.g., the toss of a coin) decides which of two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ about parameter θ shall be performed, then only the outcome of the actually performed experiment is evidentially relevant.

Both principles can be read as restricting the inferential role of information that does not directly serve an epistemic goal: the Sufficiency Principle claims that evidential support only depends on observations of a particular kind, and the Conditionality Principle marks certain procedural aspects of an experiment (e.g., which alternative experiments could have performed) as evidentially irrelevant. We thus see that Bayesian and quasi-Bayesian concepts of evidence can even be foundationally justified from requirements that express the need for a certain type of objectivity in inference. This is remarkable, given the large history of objections to Bayesian inference that are based on an apparent lack of objectivity, as exemplified in the quotes by Fisher, Cox and Mayo.

4 Objection 3: Experimental Design and Error Control

The third objection to subjective Bayesianism concerns the problem of bias in trials with interim looks at the data. The problem can best be motivated with an example from medicine. Randomized Controlled Trials (RCTs) are currently the gold standard within evidence-based medicine. They are usually conducted as *sequential trials* allowing for monitoring for early signs of effectiveness or harm. In sequential trials, data

are typically *monitored* as they accumulate. That is, we have interim looks at the data and we may decide to stop the trial before the planned sample size is reached. By terminating a trial when overwhelming evidence for the effectiveness or harmfulness of a new drug is available, the prohibitive costs of a medical trial can be limited and in-trial patients are protected against receiving inferior treatment.

However, such truncated trials are often seen as problematic. In a review of 134 trials stopped early for benefit, Montori et al. (2005) point to an inverse correlation between sample size and treatment effect: the smaller the sample size achieved by the trial at the moment of stopping, the larger the estimate it provided for the effect. These findings are supported by a more recent study by Bassler et al. (2010) where truncated trials report significantly higher effects than trials that were not stopped early. While the authors of these studies do not object to monitoring and truncating trials in general, they advocate that results (e.g., effect size estimates) from such trials be treated with caution. Truncating a trial seems to introduce a bias toward overestimating effect sizes. A good measure of evidential support should take this into account.

Bayesian measures of evidence such as the Bayes factor do not depend on the sampling protocol or experimental design and evaluate truncated trials like fixed-sample trials. Indeed, critics of Bayesian inference such as Deborah Mayo (1996) complain that decoupling statistical inference from the sampling protocol "can lead to a high probability of error, and [...] this high error probability is not reflected in the interpretation of data" (Mayo and Kruse, 2001). In the context of medical research, the Bayesian seems to provide *carte blanche* for implementing any design that favors the pursuit of certain non-cognitive values, such as the financial interests of the trial sponsor. For instance, we could sample on until a significant result is reached and then decide not to report that the results were reached in a biased way. After all, the sampling protocol is evidentially irrelevant for the Bayesian. Again, the perceived threat to the objectivity of Bayesian inference comes from the hidden intrusion of bias and non-cognitive values into statistical reasoning.

Three responses can be made to this criticism. First, the phenomenon on which the criticism is based can also be described differently. Higher effect sizes in truncated trials are not surprising, but *predictable* (Goodman et al., 2010). Of all treatments, highly efficacious ones will be most prone to early termination for benefit. That is, when the actual effect size is large, it is more probable that we also observe a large effect in our sample and decide to terminate the trial. Hence, the observed difference between truncated and completed trials is precisely what we should expect. Comparing trun-

cated to completed trials amounts, as highlighted by Berry et al. (2010), to selecting the trials to be compared on the basis of their outcome. In that light, it is questionable whether the observed effect size difference is really problematic.

Second, prior knowledge or empirically-based prior expectations are highly relevant for dealing with overestimated effects. Imagine that we are interested in the relative risk reduction which a medical drug provides. A Bayesian represents her uncertainty by means of a prior probability distribution over that quantity. By means of Bayes' Theorem, this distribution is updated to a posterior probability distribution that synthesizes the observed evidence with the background knowledge. Then, the Bayesian framework naturally accounts for the intuition that truncated trials should be treated with caution: for the same observed effect size, small sample sizes change the prior distribution less than large sample sizes. The posterior distribution visualizes these differences in an intuitive way that can be directly used for decision-making (Goodman, 2007; Nardini and Sprenger, 2013). In other words, the subjective Bayesian has an automatic safeguard against rash conclusions which other inference schools do not possess.

Third, that Bayes factors do not depend on the sampling protocol does not imply that Bayesians should ignore matters of experimental design. Procedural objectivity in the form of following certain regulatory constraints and standard procedures can be helpful to eliminate certain forms of institutional bias (Douglas, 2004, 2009). In fact, guidelines for the use of Bayesian statistics (such as the ones issued by the Food and Drug Administration) stress that Bayesians should be as conscious and diligent in matters of experimental design as frequentists. For instance, also from a Bayesian perspective, a test with high type I and type II errors is evidently a bad test. The point of disagreement is different: while the frequentist bases her post-experimental evaluation of the evidence on the pre-experimental design and the properties of the entire experiment, the Bayesian considers these properties as essential for obtaining valid data, but as orthogonal to the question of how to interpret them once they are in.

In total, the claim that Bayesian inference in sequential trials contains an implicit bias can be soundly rebutted. The particular problem of sequential analysis and monitoring ongoing trials poses no challenges to Bayesian inference that it does not also pose to competing paradigms.

5 Conclusion: A Digression on Scientific Objectivity

The concept of scientific objectivity is a notoriously difficult one, with various aspects and interpretations. It is a commonly shared view, though, that objective conclusions support the epistemic authority of science, distinguishing it from religion or political ideology. No wonder that statistical approaches are also valued according to their ability to provide an image of objectivity. Objective reasoning can manifest itself in different ways, e.g., leading to intersubjective agreement on evidence, priority of evidence over values, freedom of idiosyncratic bias, standardization of inference procedures, responsiveness to criticism, and so on. The standard criticisms of Bayesian inference relate to selected aspects of the complex notion of scientific objectivity. We recap the main ideas below.

First, there is the idea that subjective Bayesian inference is particularly vulnerable to the intrusion of bias and non-cognitive values since there is apparently no restriction on choosing prior probabilities. However, prior degrees of belief can incorporate valuable expertise and background information and they can (and should!) be criticized like any statistical model assumption. Once these points are recognized, the objection loses its bite. It can also be demonstrated that sensitivity analysis in Bayesian inference contributes to convergent objectivity in Douglas's sense: validation of a result from different independent perspectives (Section 2).

Second, there is the fear that on a Bayesian approach, scientific evidence is always entangled with (possibly idiosyncratic and biased) subjective judgments of belief. Similarly, one may argue that intersubjective agreement on levels of evidence—the concordant dimension of objectivity—is hard to achieve on a Bayesian approach. We have shown that for simple hypothesis testing, these fears are not substantiated. And for composite hypotheses, the Bayes factor (=the Bayesian's standard measure of evidence) only depends on the relative weight of the individual hypotheses—a dependency which we have argued to be benign and necessary for meaningful scientific inference.

Finally, I would like to gloss on aspects of objectivity that relate to interaction and mutual criticism in a research community. Here, Helen Longino (1990) has forcefully argued that scientific objectivity is not only about scientific reasoning itself, but also about the structure of scientific discourse: the possibility of openly criticizing each other's assumptions, providing a floor for the exchange of rational arguments, etc. In this respect, Bayesian inference has several important assets: it is honest and trans-

parent about the assumptions it makes and clearly distinguishes between prior belief, evidence, and conclusions (=posterior belief). This points out clear avenues for model criticism and allows for a straightforward detection of inappropriate bias, such as prior assumptions that heavily favor a particular hypothesis. Such bias is usually more hidden in frequentist inference, e.g., in experimental designs that deliver impressively low p-values and barely discernible, scientifically irrelevant effect sizes at the same time. Moreover, subjective Bayesianism provides a rigorous description of what happens when the prior assumptions on a parameter value are varied. The transparency of the role of individual degrees of beliefs, hidden and implicit in other schools of statistical inference, can be seen as a plus of subjective Bayesianism from the vantage point of scientific objectivity.

In the light of these arguments, claims that subjective Bayesians cannot quantify evidence in an objective way must be rejected as unjustified. They rely on a too narrow and one-sided view of scientific objectivity, on a too simplified picture of Bayesian inference and on a blind eye regarding the shortcomings of classical, frequentist inference. Even more, it has been shown that the diversity of prior distributions that characterizes subjective Bayesianism can also be a strength from the point of view of scientific objectivity.

Finally, a caveat. The purpose of this article is not to promote subjective Bayesianism as a one-size-fits-all solution for problems of statistical and scientific inference. It is well known that in many modeling problems, it is difficult to come up with meaningful subjective degrees of belief. Rather, I have argued that when an inference problem is such that subjective Bayesianism can be used to solve it, the apparent lack of objectivity should not prevent us from applying Bayesian methods. The objectivity problem is no more and less pressing than for any other scientific method.

References

Bassler, D., Briel, M., Montori, V. M., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D., Walter, S. D., Guyatt, G. H., N Flynn, D., and Others (2010). Stopping randomized trials early for benefit and estimation of treatment effects. *JAMA*, 303(12):1180–1187.

Bem, D. J., Utts, J., and Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of personality and social psychology*, 101(4):716–719.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

Berry, S. M., Carlin, B. P., and Connor, J. (2010). Bias and Trials Stopped Early for Benefit. *JAMA*, 304(2):156.

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, 57(298):269–306.

Cooke, R. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, Oxford.

Cox, D. and Mayo, D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In Mayo, D. G. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, pages 276–304. Cambridge University Press, Cambridge.

de Finetti, B. (1972). *Probability, Induction and Statistics: the Art of Guessing*. John Wiley & Sons, New York.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, 138(3):453–473.

Douglas, H. (2008). The Role of Values in Expert Reasoning. *Public Affairs Quarterly*, 22:1–18.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. Pittsburgh University Press, Pittsburgh.

Douglas, H. (2011). Facts, Values and Objectivity. In Jarvie, I. and Zamora Bonilla, J., editors, *The SAGE Handbook of the Philosophy of Social Sciences*, pages 513–529. SAGE Publications, London.

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd, Edinburgh.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Hafner, New York.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological review*, 116(2):439–453.

Goodman, S. J., Berry, D., and Wittes, J. (2010). Bias and Trials Stopped Early for Benefit. *JAMA*, 304(2):157.

Goodman, S. N. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine*, 146(12):882.

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

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90:773–795.

Lacey, H. (1999). *Is Science Value Free? Values and Scientific Understanding*. Routledge, London.

Lee, M. D. and Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, Cambridge.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, NJ.

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, 7(6):537–542.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press, Chicago.

Mayo, D. G. and Kruse, M. (2001). Principles of inference and their consequences. In Corfield, D. and Williamson, J., editors, *Foundations of Bayesianism*. Springer, New York.

Montori, V. M., Devereaux, P. J., Adhikari, N. K. J., Burns, K. E. A., Eggert, C. H., Briel, M., Lacchetti, C., Leung, T. W., Darling, E., Bryant, D. M., Bucher, H. C.,

Schünemann, H. J., Meade, M. O., Cook, D. J., Erwin, P. J., Sood, A., Sood, R., Lo, B., Thompson, C. A., Zhou, Q., Mills, E., and Guyatt, G. H. (2005). Randomized trials stopped early for benefit: a systematic review. *JAMA : the journal of the American Medical Association*, 294:2203–2209.

Moyé, L. A. (2008). Bayesians in clinical trials: Asleep at the switch. *Statistics in Medicine*, 27:469–482.

Nardini, C. and Sprenger, J. (2013). Bias and Conditioning in Sequential Medical Trials. *Philosophy of Science*, 80:1053–1064.

Popper, K. R. (2002). *Logic of Scientific Discovery*. Routledge, London.

Quine, W. (1992). *Pursuit of Truth*. Harvard University Press, Cambridge MA.

Ramsey, F. P. (1926). Truth and Probability. In Mellor, D. H., editor, *Philosophical Papers*, pages 52–94. Cambridge University Press, Cambridge.

Reiss, J. and Sprenger, J. (2014). Scientific Objectivity. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 10:1–6.

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals*, 2:48–66.

US Food and Drug Administration (2010). Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials.

Weber, M. (1904). Die Objektivität sozialwissenschaftlicher und sozialpolitischer Erkenntnis. In *Gesammelte Aufsätze zur Wissenschaftstheorie*, pages 146–214. Mohr, Tübingen, 1988, 7th edition.

Williamson, J. (2007). Motivating Objective Bayesianism: From Empirical Constraints to Objective Probabilities. In Harper, W. and Wheeler, G., editors, *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr.*, pages 151–179. Manuscript under review., London.