Foundations of Bayesian Instrumentalism

By

Olav B. Vassend

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Philosophy)

at the

UNIVERSITY OF WISCONSIN–MADISON

2017

Date of final oral examination: 5/3/2017

The dissertation is approved by the following members of the Final Oral Committee:

Malcolm Forster, Professor, Philosophy

Elliott Sober, Professor, Philosophy

Michael Titelbaum, Associate Professor, Philosophy

Daniel Hausman, Professor, Philosophy

Robert Streiffer, Associate Professor, Bioethics and Philosophy

For Philosophy.

Acknowledgments

I am grateful to many people who have helped me with various parts of my dissertation, several of whom I have credited in footnotes at different points in the dissertation.

I am grateful to Branden Fitelson, Malcolm Forster, Elliott Sober, and Michael Titelbaum for reading and commenting on a previous version of Chapter 2. A version of Chapter 2 was presented at the 2014 Pacific APA, and I'm grateful to several people who provided me with feedback after my presentation, in particular Brad Armendt, Kenny Easwaran, Sam Fletcher, and Greg Gandenberger.

I am grateful to Branden Fitelson and Michael Titelbaum for reading a previous version of Chapter 3, and I am grateful to participants at the 2014 meeting of the PSA – especially Greg Gandenberger – for helpful feedback.

I am grateful to Malcolm Forster and Reuben Stern for reading a previous version of Chapter 4. I am also grateful to the participants at the 2016 meeting of FEW for feedback on Chapter 4, in particular Kenny Easwaran and Jan Sprenger, both of whom also provided me with written feedback.

I am grateful to Malcolm Forster, Elliott Sober, and participants at the 2016 meeting of the PSA for feedback on Chapter 5. I am also very grateful to faculty and students (too many to list individually) at the University of Wisconsin – Madison

and at Nanyang Technological University for providing me with much useful feedback on presentations that were based on Chapter 5.

More generally, I wish to thank Malcolm Forster, Emi Okaysu, Elliott Sober, Reuben Stern, and Michael Titelbaum for many stimulating conversations during my time in Madison. I am likewise grateful to David Buller and Carl Gillett during my time at NIU. I also thank my brother, Nicolai Vassend, and father, Olav Vassend, for many philosophical conversations (often heated) over the past couple of decades.

Thanks are due to the philosophy front office staff: Nina Akli, Lori Grant, Christy Horstmeyer, Cheryl Scutte, Patty Winspur. And thanks are also due to my committee members: Malcolm Forster, Dan Hausman, Elliott Sober, Rob Streiffer, and Mike Titelbaum.

Lastly, I thank my parents, Ellen and Olav Vassend, wife, Emi Okayasu, and parents-in-law, Martha and Ryuichi Okayasu, for all their support.

Contents

Contents iv

Abstract vii

1 Introduction 1

1.1 Goals of the Dissertation and of this Introduction 1

1.2 The Bayesian Framework 2

1.3 Why Bayesianism? 7

1.4 Various Objections to Bayesianism Addressed 11

1.5 Conclusion 15

2 The Ordinal Equivalence Thesis and the Goal-Relativity of Bayesian Norms 17

2.1 Goals of the Chapter 17

2.2 The Ordinal Equivalence Thesis 17

2.3 Formal Characterizations of Various Equivalence Theses 21

2.4 Consequences of Adopting the Ordinal Equivalence Thesis 23

2.5 Applications of Bayesian Confirmation Measures that Rely on a Rejection of OET 33

2.6 On the Goal-Relative Nature of Bayesian Norms 40

2.7 Conclusion 45

3 A New Characterization of Confirmation Measures 46

3.1 Goals of the Chapter 46

3.2 Background Assumptions 47

3.3 The Main Requirement on c 50

3.4 The Main Result 58

3.5 Discussion and Objections 59

3.6 Conclusion 64

4 The Goal-Relativity of Prior Probabilities 65

4.1 Goals of the Chapter 65

4.2 Information Measures and their Importance for Bayesianism 67

4.3 Other Approaches to Measuring Information 69

4.4 Confirmation Measures as Measures of the Informativeness of Data 72

4.5 How to Derive Information Measures From Confirmation Measures 87

4.6 Two Goal-Relative Non-Informative Priors 93

4.7 Goal-Relative Priors Given Objective Background Information 95

4.8 Wider Implications for Bayesianism 98

4.9 Conclusion 102

5 The Interpretive Problem in Bayesian Inference 103

5.1 Goals of the Chapter 103

5.2 The Basics of Bayesian Statistical Inference 104

5.3 The Interpretive Problem in Bayesian Statistical Inference 106

5.4 The Pervasiveness of the Interpretive Problem 109

5.5 Solutions to the Interpretive Problem 112

5.6 The Verisimilitude Interpretation of Probability 114

5.7 The Verisimilitude Interpretation of Probability is Useful 117

5.8 The Counterfactual Interpretation of Probability 123

5.9 Relationship Between the Verisimilitude, Counterfactual, and Standard Interpretations 128

5.10 The Law of Likelihood 130

5.11 What About In-Between Cases? 132

5.12 Conclusion 134

6 Concluding Speculative Thoughts 137

7 Appendix 139

7.1 Derivation of the Main Result of Chapter 3 139

7.2 Derivations of (19) and (20) in Chapter 4 142

7.3 Derivation of (22) in Chapter 4 144

7.4 Derivation of the Rational Requirement on the Prior in Chapter 5, Section 5.7 145

References 147

Abstract

According to Bayesian epistemologists and statisticians, plausibility judgments are representable by real numbers, and these numbers ought, rationally, to obey the probability axioms. For example, if you find P very plausible, then it is irrational for you to also find the negation of P very plausible. More generally, you can use the Bayesian framework to decide which of your plausibility judgments are rational and how your judgments ought to change given evidence. Bayesianism is not the only such framework, but it is arguably the most influential one.

There are good reasons for taking the Bayesian framework seriously. On the one hand, multiple independent arguments purport to show that any numerical plausibility measure ought to be a probability distribution. On the other hand, and perhaps more importantly, the Bayesian framework has been applied very successfully both within philosophy and within statistics.

However, Bayesianism faces multiple objections. For example, Bayesians typically interpret the probability of a hypothesis as the plausibility that the hypothesis is true, which suggests that scientific inference is a search for true hypotheses; however, scientists often use Bayesian methods to explore scientific models they already know to be false. Moreover, the status of Bayesian rational norms is unclear, because

ordinary human beings typically do not have epistemic attitudes as fine-grained as the Bayesian framework seems to assume.

The major goal of my dissertation is to argue that the norms that govern Bayesian quantities, such as an agent's probability function or confirmation measure, are systematically influenced by the goals of the relevant agent. In other words, Bayesianism underwrites a kind of epistemic instrumentalism. In particular, I will argue that the fact that Bayesian norms are goal-relative implies that the Bayesian framework does not make unrealistic assumptions about human psychology. Moreover, I will show that Bayesians are not committed to interpreting probability distributions as representing plausibility judgments, and that in fact some goals require that we interpret the Bayesian probability of a hypothesis as something other than the plausibility that the hypothesis is true.

The dissertation is divided into six chapters. Chapter 1 is an introduction to the Bayesian framework that is intended to clarify how Bayesianism will be understood in the rest of the dissertation. In Chapter 2 I discuss a popular thesis about Bayesian measures of confirmation that I call the "ordinal equivalence thesis" (OET). I argue that OET must be rejected if confirmation measures are to be legitimately used in several ways that philosophers use them, and that a rejection of OET leads to a new conditional norm of rationality. Chapter 3 is a more technical chapter that provides a novel quantitative characterization of the three most common confirmation measures. Chapter 4 argues that the goals an agent has partly determine whether a confirmation measure and probability distribution are rational for the agent. Chapter 5 argues that Bayesianism can accommodate the idea that some false hypotheses are

closer to the truth than others, but only at the expense of giving up the standard Bayesian interpretation of probability. Chapter 6 ends with a few speculative thoughts concerning the generality of the conclusions drawn in the dissertation.

1 Introduction

1.1 Goals of the Dissertation and of this Introduction

Bayesians use probability distributions and quantities derived from probability distributions to model practical and epistemic rationality. My over-arching goal in the dissertation is to argue that the goals you have influence which Bayesian norms it is rational for you to obey, which epistemic state it is rational for you to occupy, and even how you rationally ought to interpret the Bayesian probability distributions you use. Insofar as the Bayesian framework is a serious framework for epistemic and practical rationality, the conclusions of the dissertation have radical upshots for our understanding of epistemic and practical rationality.

In this introduction, I will give a brief overview of the Bayesian framework. The overview is intended to cohere with how I think the framework should best be understood, and will for that reason perhaps be a bit different from the normal way in which Bayesianism is motivated and introduced. After I have introduced the basics

of the Bayesian framework, I will briefly explain why I think we should take the framework seriously, and then I will consider some common objections. In addressing these objections, it will hopefully become clearer to the reader how the Bayesian framework will be understood in the rest of the dissertation.

1.2 The Bayesian Framework

The fundamental objects in the Bayesian framework are probability distributions over Boolean algebras. Given a set **S** of elements of some form – e.g. sentences, propositions, hypotheses, theories, etc. – **A** is a Boolean algebra generated by the elements of **S** just in case **A** consists of all complex expressions that can be formed recursively from elements in **S** using conjunctions, disjunctions, and negations.[1] $\mathsf{Pr}_\mathsf{K}(\mathbf{A}|\mathfrak{x})$ is a probability distribution over **A** conditional on $\mathfrak{x}$ and given background assumptions K if and only if $\mathsf{Pr}_\mathsf{K}(\mathbf{A}|\mathfrak{x})$ satisfies the following axioms:

1. $\mathsf{Pr}_\mathsf{K}(\mathfrak{a}|\mathfrak{x}) = 1$ if $\mathsf{K}\&\mathfrak{x} \models \mathfrak{a}$, for all $\mathfrak{a} \in \mathbf{A}$.

2. $\mathsf{Pr}_\mathsf{K}(\mathfrak{a}|\mathfrak{x}) \geqslant 0$ for all $\mathfrak{a} \in \mathbf{A}$

3. $\mathsf{Pr}_\mathsf{K}(\mathfrak{a} \vee \mathfrak{b}|\mathfrak{x}) = \mathsf{Pr}_\mathsf{K}(\mathfrak{a}|\mathfrak{x}) + \mathsf{Pr}_\mathsf{K}(\mathfrak{b}|\mathfrak{x})$, for all $\mathfrak{a}$ and $\mathfrak{b}$ in **A** such that $\mathsf{K}\&\mathfrak{x} \models \neg(\mathfrak{a}\&\mathfrak{b})$[2]

The *unconditional* probability distribution $\mathsf{Pr}_\mathsf{K}(\mathbf{A})$ is the special case where no $\mathfrak{x}$ is conditioned on. Now suppose we have two algebras **A** and **X**, and suppose we have

[1] Or unions, intersections, and complements.

[2] This is finite additivity. Countable additivity, which will implicitly be assumed at certain times in the dissertation, is a bit stronger.

a conditional distribution $\mathsf{Pr}_{\mathsf{K}}(\mathbf{A}|\mathfrak{x})$ for each $\mathfrak{x} \in \mathbf{X}$ and an unconditional distribution, $\mathsf{Pr}_{\mathsf{K}}(\mathbf{X})$. Then the *joint* distribution over $\mathbf{A}$ and $\mathbf{X}$, $\mathsf{Pr}_{\mathsf{K}}(\mathbf{A}\&\mathbf{X})$ is the set of all probabilities of the following form: $\mathsf{Pr}_{\mathsf{K}}(\mathfrak{a}|\mathfrak{x})\mathsf{Pr}_{\mathsf{K}}(\mathfrak{x})$. Note that if we also have a conditional distribution $\mathsf{Pr}_{\mathsf{K}}(\mathbf{X}|\mathfrak{a})$ for each $\mathfrak{a} \in \mathbf{A}$ and an unconditional distribution $\mathsf{Pr}_{\mathsf{K}}(\mathbf{A})$, then there will be two ways of defining the joint distribution over $\mathbf{A}$ and $\mathbf{X}$, using either $\mathsf{Pr}_{\mathsf{K}}(\mathfrak{a}|\mathfrak{x})\mathsf{Pr}_{\mathsf{K}}(\mathfrak{x})$ or $\mathsf{Pr}_{\mathsf{K}}(\mathfrak{x}|\mathfrak{a})\mathsf{Pr}_{\mathsf{K}}(\mathfrak{a})$. In general, for consistency, the joint distribution should be the same regardless of how we arrive at it. Consequently, we should require that $\mathsf{Pr}_{\mathsf{K}}(\mathfrak{a}|\mathfrak{x})\mathsf{Pr}_{\mathsf{K}}(\mathfrak{x}) = \mathsf{Pr}_{\mathsf{K}}(\mathfrak{x}|\mathfrak{a})\mathsf{Pr}_{\mathsf{K}}(\mathfrak{a})$ for all $\mathfrak{a}$ and $\mathfrak{x}$. Rearranging this equality gives us Bayes's formula: $\mathsf{Pr}_{\mathsf{K}}(\mathfrak{a}|\mathfrak{x}) = \mathsf{Pr}_{\mathsf{K}}(\mathfrak{x}|\mathfrak{a})\mathsf{Pr}_{\mathsf{K}}(\mathfrak{a})/\mathsf{Pr}_{\mathsf{K}}(\mathfrak{x})$. Thus we see that Bayes's formula falls out as a fundamental constraint given that we want the joint distribution over two algebras to be the same regardless of how we construct it.

In Bayesian philosophy of science and statistics, there are two kinds of algebras that are of special interest, namely the algebra generated by a set of competing hypotheses and the algebra generated by a set of possible observations; the joint probability distribution over two (or more) such algebras is the starting point for any Bayesian analysis. More precisely, suppose, on the one hand, that we have a set of competing hypotheses, H_1, H_2, ..., H_n. And suppose, on the other hand, that we have a set of possible observations or evidence E_1, E_2, ..., E_m. Let $\mathbf{H}$ be the algebra constructed out of the set of hypotheses and let $\mathbf{E}$ be the algebra constructed out of the set of possible evidence, then the basic object of study for Bayesians will be some joint distribution over $\mathbf{H}$ and $\mathbf{E}$, $\mathsf{Pr}_{\mathsf{K}}(\mathbf{H} \,\&\, \mathbf{E})$.

Almost invariably, the joint distribution over $\mathbf{H}$ and $\mathbf{E}$ is constructed in the following way: first, a so-called "prior" probability distribution $\mathfrak{p}_{\mathsf{K}}$ is assigned over $\mathbf{H}$.

Next, a conditional probability is assigned to each E_i given each H_j, thus creating a set of conditional probability distributions $q_K(\mathbf{E}|H_j)$, one for each H_j . Finally, the joint distribution, $Pr_K(\mathbf{H}\&\mathbf{E})$ is the set of probabilities $q_K(E_i|H_j)p_K(H_j)$.

More generally, if there are multiple sets of hypotheses and multiple sets of possible observations, then the Bayesian approach is to create a joint distribution over all the sets of hypotheses and all of the sets of possible observations. The fundamental tenet of Bayesianism is thus the following: any time you consider a set of possibilities – multiple competing hypotheses, multiple possible assumptions, multiple possible observations, etc. – you ought to assign a joint probability distribution over all the possibilities. There are other rational frameworks that also rely on the use of probability distributions, but what distinguishes the Bayesian framework is the *unrestricted* use of probabilities. For example, frequentism and likelihoodism are two alternative probabilistic frameworks,[3] but in contrast to Bayesianism both these frameworks eschew assigning probabilities to hypotheses.

Given a particular piece of evidence, x and given that $\mathbf{H}$ is the set of competing hypotheses of interest, Bayesians do inference by following a three-step procedure. In the first step, the prior probability of each H_i in $\mathbf{H}$ is identified. In the second step, the "likelihood" of each hypothesis is identified. The likelihood of H_i is the probability that H_i assigns to x, $p_K(x|H_i)$ Finally, in the third step, the prior and likelihood of each hypothesis is combined using Bayes's formula in order to arrive at the "posterior" probability of each hypothesis given the evidence, $p_K(H_i|x)$.

[3]See Sober (2008) for a discussion.

The above presentation of the Bayesian framework may seem rather abstract and unfamiliar. This is intentional: the presentation I have given is completely abstract and mathematical in order to make it clear that Bayesians do not need to commit to any particular interpretation of either the probability function or the elements over which the probability function ranges. How the probabilities that Bayesians make use of ought to be interpreted will be discussed in more depth later in the dissertation. But as a first approximation – and for most of the dissertation – $\mathsf{Pr}_{\mathsf{K}}(\mathsf{H})$ may be interpreted as a judgment of how plausible H is and $\mathsf{Pr}_{\mathsf{K}}(\mathsf{E}|\mathsf{H})$ may be interpreted as a judgment of how plausible E is in light of H. In general, there is no such thing as an unqualified plausibility judgment. Implicitly, any plausibility judgment you make will rest on a large number of assumptions, some mundane and some more risky. In the above mathematical characterization of the Bayesian framework, I have made this explicit by adding the K subscript. From now on I will for the most part leave K implicit; however, it is important to remember that there is always some K lurking in the background.

Once we have a joint probability distribution over a set of competing hypotheses and a set of possible observations, we can explicate important epistemic concepts quantitatively, and we can also define new important quantities that help us explicate yet other epistemic concepts. For example, according to the standard Bayesian explication of (incremental) confirmation, evidence E confirms hypothesis H if and only if $\mathsf{Pr}(\mathsf{H}|\mathsf{E}) > \mathsf{Pr}(\mathsf{H})$.[4] This criterion suffices to answer the qualitative question of whether or not E confirms H, but it does not answer the quantitative question of

[4]Disconformation and absence of confirmation (neutrality) can be defined analogously.

how much E confirms H, nor does it answer the comparative question of which of two confirmed hypotheses is confirmed more by E. To answer the quantitative and comparative questions, we must adopt a confirmation measure that quantifies the degree to which E confirms H. The following are some of the confirmation measures that have been suggested in the literature:

The plain ratio measure, $\mathsf{r}(\mathsf{H},\mathsf{E}) = \frac{\mathsf{Pr}(\mathsf{H}|\mathsf{E})}{\mathsf{Pr}(\mathsf{H})}$

The log-ratio measure, $\mathsf{lr}(\mathsf{H},\mathsf{E}) = \log \mathsf{r}(\mathsf{H},\mathsf{E})$[5]

The difference measure, $\mathsf{d}(\mathsf{H},\mathsf{E}) = \mathsf{Pr}(\mathsf{H}|\mathsf{E}) - \mathsf{Pr}(\mathsf{H})$

The log-likelihood measure, $\mathsf{l}(\mathsf{H},\mathsf{E}) = \log \frac{\mathsf{Pr}(\mathsf{E}|\mathsf{H})}{\mathsf{Pr}(\mathsf{E}|\neg\mathsf{H})}$

The alternative difference measure, $\mathsf{s}(\mathsf{H},\mathsf{E}) = \mathsf{Pr}(\mathsf{H}|\mathsf{E}) - \mathsf{Pr}(\mathsf{H}|\neg\mathsf{E})$[6]

The Kemeny-Oppenheim measure, $\mathsf{k}(\mathsf{H},\mathsf{E}) = \frac{\mathsf{Pr}(\mathsf{E}|\mathsf{H})-\mathsf{Pr}(\mathsf{E}|\neg\mathsf{H})}{\mathsf{Pr}(\mathsf{E}|\mathsf{H})+\mathsf{Pr}(\mathsf{E}|\neg\mathsf{H})}$

Once we have a measure of confirmation, we can explicate yet other concepts. For example, the overall impact that a piece of evidence has on our epistemic state (assuming we interpret the probability distribution as somehow representing our epistemic state) may be quantified as the sum, or weighted sum, of the confirmation scores of all the hypotheses under consideration given the evidence. As we will see in Chapter 4, each such *divergence measure* can, in turn, be used to define a measure of how opinionated an epistemic state is. Thus we see that the Bayesian framework is

[5]It is not customary to specify the base of the logarithm.

[6]This measure is also sometimes called the "Joyce-Christensen measure," after Joyce (1999) and Christensen (1999)

rich enough to explicate in a formal fashion many concepts of epistemic significance. Later in the dissertation we will see that pragmatic factors sometimes influence which formal explication is appropriate.

1.3 Why Bayesianism?

The goal of this dissertation is to explore the foundations of the Bayesian framework, and in particular to explore various ways in which the quantitative structure and the norms of the framework interact with epistemic and practical goals. This project is only valuable insofar as Bayesianism is a serious framework for epistemic and practical rationality. Why should we think that it is? I cannot give a full answer to that question in an introduction; nonetheless, I feel compelled to give an "apologetic" of sorts for the Bayesian framework. Furthermore, since "Bayesianism" means different things to different people, it seems necessary to clarify – at least somewhat – what I take the framework to be.

The main – but far from only – reason why I think Bayesianism should be taken seriously as a framework for rationality is because of its successful applications in science and in philosophy. The successful applications of the Bayesian framework in statistical inference are obvious; here, I will focus my attention on Bayesianism as a framework for rationality in philosophy of science.

I agree with Clark Glymour (1980) that it is, in general, not useful to give "Bayesian reconstructions" of famous historical episodes that try to explain, say, why one theory was rationally replaced by another one. Nor do I think the traditional Bayesian view

of induction – positing, as it does, an ideal agent who has probabilities over every conceivable theory and hypothesis – provides a realistic model of how induction in the sciences actually works. In my view, Bayesian philosophy of science is useful because it offers us a systematic way of exploring when and why various inference rules that scientists actually use are defensible and when they are indefensible.[7]

But why use the Bayesian framework for this purpose? To take the verdicts provided by Bayesian analyses seriously, you have to have some initial trust in the framework; where is this initial trust supposed to come from? Here I can only give a brief argument for why you should have initial trust in the Bayesian framework.

Given a set of hypotheses, H_1, H_2 ..., H_n, we often want to "score" the hypotheses relative to each other – we want to say that some of them are more plausible, or more plausibly closer to the truth, or more plausibly explanatory, etc., in light of the evidence. That's not to say we want to be able to score the set of *all* hypotheses for *all* good-making features on a single scalar scale. Rather, given a *specific* set of hypotheses, and given a *specific* "scientific virtue" (e.g. truth, closeness to the truth, predictive accuracy, etc.), we want to be able to give a provisional evaluation of which of the hypotheses we think is doing best with respect to the scientific virtue under consideration. Formally, it's natural to use a numerical scale for this purpose, and once you agree to using a numerical scale to score the various hypotheses, the typical arguments (e.g. accuracy-arguments (Pettigrew, 2016) and axiomatic arguments

[7]Of course, the fact that an inference rule doesn't have a Bayesian justification does not necessarily show that the inference rule is indefensible, as it may have a justification in some other framework, e.g. the "Akaike framework" (see, e.g. Forster and Sober (1994) and Sober (2015)). However, if Bayesianism is understood broadly enough, it may be that these other frameworks themselves ultimately have Bayesian justifications. I will not explore that possibility here.

(Jaynes, 2003)) for the Bayesian framework do not require many extra assumptions in order to kick in. These arguments are by no means decisive, but they certainly give you an initial reason to place some trust in the Bayesian framework. Once you have placed your trust in the framework, you can use it to analyze various rules of scientific inference. The fact that these analyses – in my experience – invariably seem reasonable, if not perfect, lends further credence to the hypothesis that Bayesianism is a good framework. To illustrate the kind of analysis I have in mind, I will give three recent examples.

First, consider the widespread use of robustness analysis in many sciences, and especially in the social and life sciences. The intuitive idea that motivates robustness analysis is the thought that a hypothesis is more plausible if the hypothesis is supported by multiple "diverse and independent" analyses. In a probabilistic framework, it's possible to make precise various candidate kinds of independence and diversity and investigate which are the important and relevant ones. In particular, Orzack and Sober (1993) argue convincingly that probabilistic independence is not a relevant sort of independence. In a recent paper, Jonah Schupbach (forthcoming) argues that partial independence and confirmational independence are also not the correct kind of independence, but rather that it's a kind of explanatory independence that is important; the relevant explanatory independence can be made precise in the Bayesian framework.

Another method used in many areas of science is inference to the best explanation. The basic idea behind inference to the best explanation is that the most confirmed hypothesis given a set of evidence is the hypothesis that strikes the best balance

between a number of "explanatory virtues," such as simplicity, unifying power, precision, and so on. By doing a Bayesian analysis, we can better understand which of these virtues are confirmational, and in what sense they are confirmational, as well as which explanatory virtues are not confirmational (see Cabrera (2017)). For example, one virtue may be "confirmational" in the sense that having the virtue can be expected to increase the posterior probability of the hypothesis; another virtue may in fact decrease the probability of the hypothesis (compared to not having the virtue), but can be expected to increase the incremental confirmation of the hypothesis using some confirmation measure. All of this can be investigated rigorously in the Bayesian framework.

The Bayesian framework can also be used to investigate the methodology of specific scientific fields. For example, the Comparative Method in Historical Linguistics relies on the use of various "rules of thumb," such as parsimony, directionality, majority rule, etc. In a Bayesian framework, it is possible to investigate necessary and/or sufficient conditions under which these inference rules are justified (see Okayasu (2017)). Often the relevant conditions are subtler than one would have initially expected.

The above Bayesian analyses of scientific rules of inference are in my view good applications of the Bayesian framework in the philosophy of science – in each case, the Bayesian framework is used to investigate the justificatory grounds of some inferential method. It may be the case that explicit Bayesian calculation will not be the best way to perform scientific inference in many scientific disciplines – this remains an open question. Analogously, it is probably the case that Utilitarianism is not a good decision procedure for making ethical decisions—even for a Utilitarian, commonsense

moral reasoning may be more reliable and have a higher utility. Still, a Utilitarian will insist that the ultimate ethical justification for any decision must be grounded in a consequentialist calculation. Similarly, it may be that Bayesianism is not the best way to do induction—even for Bayesians. However, it may still be the case that the rational justification of any scientific inference ultimately bottoms out in some sort of probabilistic calculation. I am not convinced that this is the case, but I think it's an intriguing possibility.[8]

In any case, all I have hoped to make a case for here is that Bayesianism has a role to play in whatever the complete framework for rationality turns out to be. Bayesianism is *a* correct framework for rationality, not necessarily *the* correct framework. The weaker claim is all I need in my dissertation.

1.4 Various Objections to Bayesianism Addressed

In the previous section, I gave my "positive" case for the Bayesian framework. However, misconceptions and misguided objections to the Bayesian framework abound, so in this section I have collected some of the most common objections with brief replies. My replies to these objections will also hopefully further clarify how Bayesianism will be understood in this dissertation.

Objection 1: Bayesianism conflates different kinds of uncertainty. Some uncertainty – "Knightian risk" (Knight, 1921) – is such that we can sensibly assign probabilities,

[8] I am grateful to Frank Cabrera for bringing the analogy with Utilitarianism to my attention.

e.g. when we are rolling a die and we are wondering how probable it is that the die will land on six. But there is also "Knightian uncertainty" where we have no rational basis for assigning any probability, for example when we are trying to assess how plausible it is that some new technology will render the Internet obsolete.

Reply: Every probability assessment, including every Bayesian probability assessment, is implicitly premised on a range of assumptions. Sometimes we know that the assumptions are true, in which case the probability is well-founded and we are in a Knightian risk case: this is the case with a die roll, for example. Other times, we are not sure whether the assumptions are accurate. Other times, again, we know the assumptions are false but we have reason to believe that they are close enough to the truth to not render the probability assessments seriously misleading.[9] Finally, sometimes we know the assumptions are false and we have no idea whether they are close to the truth – the last case is an extreme case of Knightian uncertainty. Once we recognize that all Bayesian probability assignments are premised on assumptions, we see that Bayesianism in fact incorporates the distinction between Knightian risk and uncertainty – and all the cases in between.

Objection 2: Bayesianism assumes that the plausibility of every propositions can be evaluated and ranked on a single numerical scale, but that is absurd.

Reply: The Bayesian framework does not assume that the plausibility of every proposition can be compared on the same scale. It's true that some Bayesians like to imagine an "ideal agent" who assigns probabilities to every conceivable proposition;

[9]What is a "misleading probability"? Suppose we base our probability assessment on assumptions $\mathtt{a}$. Generally, $\mathtt{a}$ will be false to some extent. Let $\mathtt{t}$ be the set of assumptions that are in fact true. Then $\mathtt{Pr}_{\mathtt{a}}(\mathtt{H})$ is a misleading probability if $\mathtt{Pr}_{\mathtt{a}}(\mathtt{H})$ is very different from $\mathtt{Pr}_{\mathtt{t}}(\mathtt{H})$.

but the Bayesian framework itself is not committed to the existence of such an agent. If you have a probability function $\mathfrak{p}$ over hypotheses H_1, H_2, ... H_n, and another probability function $\mathfrak{q}$ over h_1, h_2, ..., h_m, then the Bayesian framework does not commit you to directly comparing, say, H_1 to h_2 in any substantively important way. Suppose $\mathfrak{p}(H_1) > \mathfrak{q}(h_2)$, for example; then, on a standard Bayesian interpretation of probability, it follows that you think H_1 is more plausible given the assumptions incorporated in $\mathfrak{p}$ than is h_2 given the assumptions incorporated in $\mathfrak{q}$; but it does not follow that you judge H_1 to be more plausible than h_2 tout court. Nor does it follow that you would, say, prefer a bet on H_1 to a bet on h_2. If the assumptions that ground the probability assignment given to h_2 are much more secure than the assumptions that ground the probability assignment given to H_1, then you would be foolish to prefer a bet on H_1 to a bet on h_2.

Objection 3: The quantitative framework assumed by Bayesianism is too strong. If we assign a probability of 0.4 to H_1 and a probability of 0.8 to H_2, we are implicitly saying that H_2 is twice as plausible as H_1. But that is ridiculous; people's plausibility judgments are at best merely ordinal, so a weaker and purely ordinal framework for epistemology is more appropriate.

Reply: It's not correct that our plausibility judgments are always merely ordinal. It is correct that we do not always have strong quantitative intuitions, in the same way that we do not always have strong ordinal intuitions. Our quantitative plausibility judgments, like our ordinal ones, are gappy. We should want our formal framework to be able to capture both types of judgments. I discuss this issue again in Chapter 2.

Objection 4: Bayesianism assumes that truth is always the goal of inquiry. But

often we know that all of our hypotheses are false and we are just trying to figure out which hypothesis is the best one.

Reply: The Bayesian framework is not committed to the pursuit of truth being the goal of inquiry. According to the standard Bayesian interpretation of probability, truth is implicitly the goal; but the Bayesian framework is not committed to this interpretation. I discuss this issue in Chapter 5.

Objection 5: Bayesianism faces the "Problem of Old Evidence." If we know that E is true, then $\mathsf{Pr}(\mathsf{E}) = 1$, and so E cannot confirm any hypothesis, because we will have $\mathsf{Pr}(\mathsf{H}|\mathsf{E}) = \mathsf{Pr}(\mathsf{H})$. But this mischaracterizes how confirmation works in science, since new theories are often confirmed by evidence we know to be true.

Reply: The problem of old evidence is a problem for Bayesians with a certain interpretation of probability, namely those who interpret the probability function as representing the actual degrees of belief of some agent. That interpretation will not be assumed in this dissertation. Let's look more closely at how the problem of old evidence arises. According to the first probability axiom, $\mathsf{Pr}_{\mathsf{K}}(\mathsf{E}) = 1$ if $\mathsf{K} \models \mathsf{E}$. Note that if this is the case, we will also have $\mathsf{Pr}_{\mathsf{K}}(\mathsf{E}|\mathsf{H}) = 1$ for every H, i.e. the likelihood of every hypothesis will be 1. Frequentists and likelihoodists therefore face the problem of old evidence just as much as Bayesians do, because frequentists and likelihoodists base their inferences on the likelihood function just like Bayesians do—the problem of old evidence is usually presented as a problem for Bayesians, but that isn't so. In any case, the problem of old evidence clearly only arises if $\mathsf{K} \models \mathsf{E}$. But of course it would be foolish to construct Pr in such a way that $\mathsf{K} \models \mathsf{E}$, so the problem of old evidence will not arise for any probability distribution that is actually

of interest.

1.5 Conclusion

No doubt not everyone will be convinced by either the positive case for Bayesianism given in Section 1.3 or the replies to the objections given in Section 1.4. However, hopefully the reader now has a better idea of how the Bayesian framework will be construed in this dissertation.

In the remainder of the dissertation, my main goal will be to show how pragmatic factors interact with the Bayesian framework in order to determine what's practically and epistemically rational. In the end we will see that Bayesianism in fact underwrites a rather radical form of epistemic instrumentalism. In recent years, several philosophers have suggested that there is "pragmatic encroachment" on epistemic norms (e.g. Fantl and McGrath (2002), Stanley (2005), and Armendt (2009)). The chapters of this dissertation may be regarded as establishing increasingly strong theses about how the pragmatic encroaches on the epistemic in a Bayesian framework. In particular, in Chapter 2, I will argue that pragmatic factors influence which norms are applicable to you. In Chapter 3, I will argue that the general goal of having a confirmation measure that is interpretable as an interval measure puts rational constraints on which measure you ought to use. In Chapter 4, I will argue that pragmatic factors influence what probability distribution it is rational for you to use. And finally, in Chapter 5, I will argue that pragmatic factors influence how you should interpret the probability distributions that you use. Along the way, several debates of

interest internal to Bayesian epistemology and Bayesian philosophy of science will be addressed, including – among others – the debates concerning confirmation measure pluralism (Hajek and Joyce, 2008), the Uniqueness Thesis (Kopec and Titelbaum, 2016), and the status of the Law of Likelihood (Royall, 1997).

2 The Ordinal Equivalence Thesis and the Goal-Relativity of Bayesian Norms

2.1 Goals of the Chapter

The narrow goal of this chapter is to argue against a commonly accepted thesis in Bayesian confirmation theory. The broader goal is to argue that it is fruitful to regard Bayesian norms as goal-relative. The chapter also serves as a foundation for the rest of the dissertation, because I argue that several applications of Bayesian confirmation measures – including several applications in this dissertation – require that the measures be interpreted as providing us with more than a merely ordinal ranking of hypotheses.

2.2 The Ordinal Equivalence Thesis

Recall that a Bayesian measure of confirmation is a measure of the distance between the posterior and prior probability of a hypothesis given evidence that represents the confirmational impact the evidence has on the probability of the hypothesis. For

example, the ratio measure, r quantifies the confirmational impact E has on H as $\mathsf{Pr}(\mathsf{H}|\mathsf{E})/\mathsf{Pr}(\mathsf{H})$ and the difference measure d quantifies the impact as $\mathsf{Pr}(\mathsf{H}|\mathsf{E})-\mathsf{Pr}(\mathsf{H})$.

It is well known that confirmation measures do not always order hypothesis-evidence pairs in the same way: the measures are sometimes ordinally non-equivalent. For instance, r and d are ordinally non-equivalent since they differ in how they rank certain hypothesis-evidence pairs.[1,2] It is obvious that two confirmation measures that are ordinally non-equivalent ought not be considered the same confirmation measure. In other words, the following is uncontroversial:

> *The Ordinal Non-Equivalence Thesis*: If two confirmation measures are ordinally non-equivalent, then the two confirmation measures are not the same confirmation measure.

I have called the ordinal non-equivalence thesis a "thesis," but perhaps it is more appropriate to call it a truism. The task of this chapter will be to investigate the converse of the non-equivalence thesis. Namely,

> *The Ordinal Equivalence Thesis*: If two confirmation measures are ordinally equivalent, then they are the same confirmation measure.

According to the ordinal equivalence thesis (OET), d and d^3 are the same confirmation measure since, even though they have differing functional forms, they rank all hypothesis-evidence pairs in the same order. The ordinal equivalence thesis has

[1]Example: $\mathsf{Pr}(\mathsf{H}) = 0.1$, $\mathsf{Pr}(\mathsf{H}|\mathsf{E}) = 0.9$, $\mathsf{Pr}(\mathsf{H}') = 0.01$, $\mathsf{Pr}(\mathsf{H}'|\mathsf{E}) = 0.5$. Here H is better confirmed than H' according to d, but H' is better confirmed than H according to r.

[2]Interestingly, the standard measures do correlate fairly well (Tentori et al., 2007).

arguably become a widespread tacit – and sometimes explicit – commitment among philosophers who work on Bayesian confirmation theory. For example, Branden Fitelson writes:

> "If two relevance measures are ordinally equivalent, then, as far as we are concerned, they are identical. So, when we say 'according to c', we really mean 'according to any measure ordinally equivalent to c"' (Fitelson, 2007, p. 7n7).

Other philosophers reveal their commitment to OET by treating ordinally equivalent measures as interchangeable, which is only legitimate given OET.[3] For example, David Glass and Mark McCartney write:

> "ł satisfies (C4) provided division by zero is equated with infinity. To avoid this, the ordinally equivalent measure proposed by Kemeny and Oppenheim (1952) can be used instead." (Glass and McCartney, 2015, p. 62n4)

Still other philosophers do not unconditionally commit to the ordinal equivalence thesis, but hold that ordinally equivalent measures are often interchangeable. For example, Tomoji Shogenji writes that, "For many purposes, ordinally equivalent measures are essentially the same measure" (Shogenji, 2012, p. 5n4). Shogenji may be right that ordinal equivalence suffices for many purposes, but a major goal of

[3]Numerous conversations I have had with philosophers who work on Bayesian confirmation theory have convinced me that it is standard for philosophers to regard ordinally equivalent measures as interchangeable.

this chapter is to show that there are also several purposes for which the ordinal equivalence thesis is insufficiently strong.

Here's the layout of the chapter. In Section 2.3, I describe various competing theses that we may choose to adopt; each of these theses corresponds to an alternative level of analysis that we may choose to prioritize. In Section 2.4, I show the shortcomings of the ordinal equivalence thesis by contrasting it with alternative theses, and in particular I show that adopting the ordinal equivalence thesis renders one unable to set various thresholds that can be used to interpret a set of confirmation scores, and that a thesis stronger than OET must be adopted if merely ordinal conclusions are to be drawn from the expected values of a confirmation measure. In Section 2.5, I show that several arguments given by philosophers already rely on a rejection of OET. In Section 2.6, I discuss possible reasons why OET has been accepted. A major reason why philosophers have focused on the ordinal level is probably because a rejection of OET seems to imply that human beings have epistemic states that have a very fine-grained quantitative structure. However, I suggest that the normative upshot of the chapter has a conditional form, so that it is only applicable when the antecedent of the conditional applies. I furthermore suggest that other Bayesian norms can fruitfully be understood as having a similar conditional form.

2.3 Formal Characterizations of Various Equivalence Theses

As is well known in the literature, ordinal equivalence can be formally characterized. More precisely, two confirmation measures are ordinally equivalent if and only if there is a strictly increasing function from each measure to the other. We can state the preceding characterization of ordinal equivalence more formally as follows:

> *Ordinal equivalence characterization*: Confirmation measures c and c' are ordinally equivalent if and only if there is a strictly increasing function, f, such that, for all H and E, $\mathsf{c}(\mathsf{H}, \mathsf{E}) = \mathsf{f}(\mathsf{c}'(\mathsf{H}, \mathsf{E}))$.

To better understand the ordinal equivalence thesis, it is useful to contrast it with alternative theses that we may instead choose to adopt. Inspired by the above characterization of ordinal equivalence, we can use the following abstract schema to derive alternative equivalence theses:

> *Confirmation Equivalence Schema*: Confirmation measures c and c' are equivalent if and only if there is an invertible function, f, such that $\mathsf{c} = \mathsf{f}(\mathsf{c}')$.

Different confirmation equivalence theses can then be characterized by what requirements they put on f. In theory, we could produce infinitely many theses from the above schema since there are potentially infinitely many requirements we could choose to put on f. Certain theses are of more theoretical interest than others,

however. Following Stevens (1946), I will call the theses I consider "ordinal," "interval," "ratio," and "absolute," where these theses are distinguished by the increasingly strong demands they place on f.

> *Ordinal Equivalence Thesis* (OET): The requirement on f is that it be strictly increasing.
>
> *Interval Equivalence Thesis* (IET): The requirement on f is that it be strictly increasing and linear.
>
> *Ratio Equivalence Thesis* (RET): The requirement on f is that it be strictly increasing and linear with constant term 0.
>
> *Absolute Equivalence Thesis* (AET): The requirement on f is that it be the identity function.

Adopting OET amounts to carving the set of all possible confirmation measures into classes of ordinally equivalent measures and treating the measures in each class as interchangeable. Similarly, IET and RET carve the space of confirmation measures into classes of measures that are what we might respectively call "interval" and "ratio" equivalent. The fourth thesis, AET, is the strongest possible thesis: its equivalence classes contain only a single confirmation measure each.

My choice of singling out the above four theses is not arbitrary. The first three theses correspond to three of the four "levels of measurement" outlined by Stevens (1946) in the context of scientific measurement. As Stevens points out, the strength of the conclusions one is licensed to draw from data depends on the strength of the

measurement scale used. What is true in the case of measurement scales is also true in the case of confirmation measures, as I show in the next section when I discuss the consequences of adopting OET by contrasting it with the consequences of instead adopting IET.

2.4 Consequences of Adopting the Ordinal Equivalence Thesis

In the following two subsections, I discuss general consequences associated with adopting OET and treating confirmation measures as mere ordinal measures. In Section 2.5, I show why these consequences matter for several of the arguments and applications of confirmation measures that have been discussed in the literature.

Interpreting a Set of Confirmation Scores

Suppose I give you the results of a 100m race with three runners by listing the order in which the runners finished. Then you are not entitled to say that the difference in performance between the winner and the runner-up is roughly the *same* as the difference in performance between the runner-up and the third-place finisher; nor may you conclude that the winner performed substantially better than the other two runners. The ordinal data with which you have been provided simply does not contain this information. Suppose you learn, however, that the winner is Usain Bolt and that the other two runners are recreational runners. Then you have reason to believe that the winner's performance was in fact much better than the performance

of the other two runners. If, on the other hand, you learn that all three runners are recreational runners, you no longer have any reason to think that the winner's performance was substantially better than the performance of the other two runners. Thus, if all you learn about the three competitors are their ordinal ranks, you cannot draw conclusions about their performance relative to each other. You can only infer such conclusions on the basis of further information.

But now suppose that you instead learn the *times* of the three runners. Suppose, for instance, that you learn that the winner's time was 10 seconds while the other two runners finished in 14 and 14.5 seconds. Then you really can say that the winner was substantially better than the other two runners, and moreover you can say that the two losers performed about equally well. To be sure, your conclusions still depend on background knowledge about running and about the time scale used, but the conclusions you are entitled to draw are much more robust than in the case where you are just given ordinal ranks in the sense that the conclusions do not depend sensitively on knowledge about the particular runners.

The above example illustrates the differences between ordinal and interval/ratio scales. If we adopt OET, then the proper way to interpret the numerical outputs of confirmation measures is as ranks. Although the outputs of e.g. the log-ratio measure $\mathfrak{lr}$ can be any real number, only the ordinal properties of the real numbers are being used. Suppose, for instance, that our favored confirmation measure — call it "$\mathfrak{m}$" — outputs the three numbers 0.91, 0.9, and 0.1 for evidence-hypothesis pairs (H, E), $(\mathsf{H}', \mathsf{E})$, and $(\mathsf{H}'', \mathsf{E})$, respectively. In that case we are entitled to say that E confirms H more than H', and that E confirms H' more than H'', but we cannot say that H'

and H'' are confirmed to approximately the same degree by E, or that each is much more highly confirmed than H''. To make any of these claims is to go beyond the merely ordinal properties of 0.91, 0.9, and 0.1.

Indeed, if we adopt OET, then any conclusion we draw from $\mathfrak{m}$'s output is valid only if it still holds when we choose to use a different ordinally equivalent measure. This is because by adopting OET we agree to treat ordinally equivalent measures as interchangeable. But it is easy to transform our $\mathfrak{m}$ into an ordinally equivalent measure that instead outputs, say, the numbers 3, 2, and 1 for the above three evidence-hypothesis pairs. All one needs to do is device a suitable strictly increasing function. For example, $\mathfrak{g}(\mathfrak{x}) = 1.25\mathfrak{x} + 0.875$ for $\mathfrak{x} \leqslant 0.9$, and $\mathfrak{g}(\mathfrak{x}) = 100\mathfrak{x} - 88$ for $\mathfrak{x} > 0.9$. Of course, $\mathfrak{g}$ is not a very "natural" function, but that is beside the point. The point is that $\mathfrak{g}$ is a strictly increasing (even continuous) function that transforms $\mathfrak{m}$ into $\mathfrak{m}'$; therefore, by OET, $\mathfrak{m}$ and $\mathfrak{m}'$ are equivalent confirmation measures. Performing the preceding transformation makes it clear that the only conclusion we are justified in drawing from the data is the ordinal ranking itself, $\mathfrak{m}(\mathsf{H}, \mathsf{E}) > \mathfrak{m}(\mathsf{H}', \mathsf{E}) > \mathfrak{m}(\mathsf{H}'', \mathsf{E})$.

The situation is different if we adopt one of the other equivalence theses. Suppose we adopt IET instead. Then any other confirmation measure in the same equivalence class as $\mathfrak{m}$ must be of the form $\mathfrak{m}' = \mathfrak{a}\mathfrak{m} + \mathfrak{b}$, with positive $\mathfrak{a}$. Thus, it must be the case that:

$$\frac{\mathfrak{m}(\mathsf{H}, \mathsf{E}) - \mathfrak{m}(\mathsf{H}', \mathsf{E})}{\mathfrak{m}(\mathsf{H}', \mathsf{E}) - \mathfrak{m}(\mathsf{H}'', \mathsf{E})} = \frac{\mathfrak{m}'(\mathsf{H}, \mathsf{E}) - \mathfrak{m}'(\mathsf{H}', \mathsf{E})}{\mathfrak{m}'(\mathsf{H}', \mathsf{E}) - \mathfrak{m}'(\mathsf{H}'', \mathsf{E})} \tag{1}$$

In other words, any functional transformation of $\mathfrak{m}$ allowed by IET preserves

relative interval sizes. The consequence of this is that while a measure outputting 0.91, 0.90, and 0.1 can be transformed into an interval equivalent measure that instead outputs the values 91, 90, and 10, respectively, it is not possible to transform it into a measure that outputs 3, 2, and 1. Thus, if we have narrowed down the range of confirmation measures to a class of interval equivalent measures and we adopt IET, then we are entitled to draw robust conclusions from the distances between the numbers outputted by our measure. If we adopt IET, then we are no longer merely using the ordinal properties of the real numbers — the difference between 0.91 and 0.9 really *is* smaller than the difference between 0.9 and 0.1.

Setting Thresholds with IET

But the fact that interval equivalent confirmation measures preserve relative interval sizes does not yet mean that we are able to conclude that, e.g., H and H' are confirmed to roughly the same degree by E. In order to draw a conclusion of this kind, we need specific knowledge about $\mathfrak{m}$'s behavior that allows us to determine that H and H' are confirmed to roughly the same degree (by E and E' respectively; of course E and E' may be identical) if and only if $|\mathfrak{m}(\mathsf{H}, \mathsf{E}) - \mathfrak{m}(\mathsf{H}', \mathsf{E}')| < \delta$, for some (small) δ. In the same way, we can establish a threshold that says that H is substantially better confirmed by E than is H' by E' if and only if $|\mathfrak{m}(\mathsf{H}, \mathsf{E}) - \mathfrak{m}(\mathsf{H}', \mathsf{E}')| > \epsilon$ for some suitably chosen ϵ.

Royall (1997, p. 11) does the preceding for the likelihood ratio, $\mathsf{Pr}(\mathsf{E}|\mathsf{H}_1)/\mathsf{Pr}(\mathsf{E}|\mathsf{H}_2)$.[4] He considers the following "canonical experiment": suppose an urn contains either all white balls or else an equal number of white and black balls. Suppose you then draw three balls with replacement and all the balls turn out to be white. Intuitively, this seems to be "pretty strong" evidence that all the balls are white rather than that half of them are black. The likelihood ratio favoring all white balls is in this case 8. Thus, Royall concludes, 8 is the threshold that signifies "pretty strong" evidence (in an everyday context, let us add) favoring one hypothesis over another. Of course, the choice of this particular canonical experiment is somewhat arbitrary, but note that the particular choice of canonical experiment does not matter much if we accept IET. A different canonical experiment may have instead yielded 7 or 9, say, as the threshold that signifies "pretty strong" or maybe just "strong" evidence. But fortunately the real numbers 7 and 9 are relatively *close* to each other, and when we adopt IET we make use of these facts about the real numbers. Therefore, nothing significant hinges on choosing either 7 or 8 or 9 as the threshold.

The precise values of the thresholds are therefore not important — in fact, the thresholds ought not be treated too precisely; what a set of thresholds allows us to do is to better interpret a set of confirmation scores. Importantly, IET allows us to use the same threshold throughout the whole confirmation scale. That is because IET implies that the difference $\mathfrak{m}(\mathsf{H},\mathsf{E}) - \mathfrak{m}(\mathsf{H}',\mathsf{E}')$ has the same meaning

[4]Note: the likelihood ratio is not a Bayesian measure of confirmation. Rather, it is a direct measure of the evidential support that one hypothesis enjoys compared with another one. As Fitelson (2007) points out, the standard Bayesian confirmation measure that agrees with using the likelihood ratio to compare the relative support of two hypotheses is the ratio measure. Thus, implicitly, Royall is setting thresholds for interpreting quantities of the form $\frac{\mathfrak{r}(\mathsf{H},\mathsf{E})}{\mathfrak{r}(\mathsf{H}',\mathsf{E})}$.

(i.e. describes the same difference in confirmation) regardless of where on the scale $\mathfrak{m}(\mathsf{H}, \mathsf{E})$ and $\mathfrak{m}(\mathsf{H}', \mathsf{E}')$ happen to be. This is exactly what (1) guarantees will be the case. And the fact that $\mathfrak{m}(\mathsf{H}, \mathsf{E}) - \mathfrak{m}(\mathsf{H}', \mathsf{E}') = \mathfrak{a}$ describes the same difference in confirmation regardless of the values of $\mathfrak{m}(\mathsf{H}, \mathsf{E})$ and $\mathfrak{m}(\mathsf{H}', \mathsf{E}')$ allows us to say, given the confirmation scores of two hypotheses, whether the two hypotheses are confirmed to essentially the same degree, or whether one of the hypotheses is better confirmed, or much better confirmed, than the other one.

It is important to appreciate the importance of being able to make these kinds of comparisons between $\mathfrak{m}(\mathsf{H}, \mathsf{E})$ and $\mathfrak{m}(\mathsf{H}', \mathsf{E})$. Indeed, the question of *whether* a piece of evidence confirms one hypothesis more than it confirms another hypothesis is essentially uninformative unless we can also at the very least determine whether the difference in confirmation is substantial or trivial. Indeed, even if we are ultimately mostly interested in the ordinal ranking provided by the confirmation measure, having confirmation scores that are at least on an interval scale prevents us from *over-interpreting* a difference in confirmation score between two hypotheses. If $\mathfrak{m}(\mathsf{H}, \mathsf{E}) > \mathfrak{m}(\mathsf{H}', \mathsf{E})$, then E confirms H more than it confirms H', but if the difference between the confirmation scores is small, the inequality may be practically insignificant, especially when measurement error is taken into account: that is, the inherent accuracy of our measurement procedure may be such that, had we repeated our measurement, the new E' could easily be such that $\mathfrak{m}(\mathsf{H}, \mathsf{E}') < \mathfrak{m}(\mathsf{H}', \mathsf{E}')$.

Setting Thresholds Without IET?

IET allows us to set thresholds that determine, e.g. whether H and H' are confirmed to roughly the same degree by some piece of evidence. Are there equivalence theses weaker than IET that allow us to do the same thing?

In general, in order to make an assessment of the "distance" between two confirmation scores, we need a function that takes as its input two confirmation scores and outputs a (non-negative) number that represents the distance between the two scores. Suppose we have available some such function, D. In order for us to be able to set up a threshold δ according to which $\mathfrak{x}$ and $\mathfrak{y}$ are "approximately equal" if and only if $\mathsf{D}(\mathfrak{x}, \mathfrak{y}) < \delta$, it needs to be the case that $\mathsf{D}(\mathfrak{x}, \mathfrak{y}) = \mathfrak{a}$ means the same thing regardless of what $\mathfrak{x}$ and $\mathfrak{y}$ happen to be. Thus, in particular, if $\mathsf{D}(\mathfrak{x}, \mathfrak{y}) = \mathfrak{a}$ and $\mathsf{D}(\mathfrak{z}, \mathfrak{w}) = \mathfrak{a}$, then it should be the case that the distance $\mathsf{D}(\mathfrak{x}, \mathfrak{y})$ means the same thing as the distance $\mathsf{D}(\mathfrak{z}, \mathfrak{w})$, so that we can say that $\mathsf{D}(\mathfrak{x}, \mathfrak{y}) = \mathsf{D}(\mathfrak{z}, \mathfrak{w})$. In order for this to be the case, the class of admissible transformations must obey something very analogous to (1). More precisely, in order for it to be legitimate to conclude that $\mathsf{D}(\mathfrak{x}, \mathfrak{y}) = \mathsf{D}(\mathfrak{z}, \mathfrak{w})$ from the fact that $\mathsf{D}(\mathfrak{x}, \mathfrak{y}) = \mathfrak{a}$ and $\mathsf{D}(\mathfrak{z}, \mathfrak{w}) = \mathfrak{a}$, it needs to be the case that $\mathsf{D}(\mathsf{f}(\mathfrak{x}), \mathsf{f}(\mathfrak{y})) = \mathsf{D}(\mathsf{f}(\mathfrak{z}), \mathsf{f}(\mathfrak{w}))$ whenever $\mathfrak{x}$, $\mathfrak{y}$, $\mathfrak{z}$, and $\mathfrak{w}$ are transformed using any admissible transformations f. Hence, the class of all admissible transformations must satisfy the following equation:

$$\frac{\mathsf{D}(\mathfrak{x}, \mathfrak{y})}{\mathsf{D}(\mathfrak{z}, \mathfrak{w})} = \frac{\mathsf{D}(\mathsf{f}(\mathfrak{x}), \mathsf{f}(\mathfrak{y}))}{\mathsf{D}(\mathsf{f}(\mathfrak{z}), \mathsf{f}(\mathfrak{w}))} \tag{2}$$

Thus, given a distance measure D, we can say that two confirmation scores are

approximately equal, or that one confirmation score is substantially greater than another confirmation score *only if* we adopt an equivalence thesis according to which only transformations that obey (2) are admissible. Now, given very weak conditions on D, the class of transformations that obey (2) will be a proper subset of the class of all strictly increasing functions.[5] It follows that OET will in general will be too weak to set thresholds. In order for us to be able say anything more specific about how strong of an equivalence thesis is required, more specific assumptions must be made about the distance measure, D.

The most natural and simplest distance measure on the real numbers is arguably the absolute distance metric, $\mathsf{D}(\mathsf{x}, \mathsf{y}) = |\mathsf{x} - \mathsf{y}|$. If we plug the absolute distance metric into (2) we recover (1). Furthermore, the linear functions are the only functions that obey (1); therefore, all admissible transformations must be linear.[6] It follows that IET is the *weakest* thesis that allows us to set thresholds of the sort discussed above, *provided* the distance measure is the absolute value metric. If some other distance measure is used, then some other thesis than IET may instead (indeed, probably must) be adopted. But in any case, OET is too weak, because any comparison of confirmation scores requires a distance measure, and the distance measure will impose the requirement that the admissible transformations obey (2).

[5]There are several conditions we could put on D. For example, one reasonable requirement is that confirmation measures scores can be arbitrarily close to each other according to D.

[6]The proof that only linear functions obey (1) is trivial and omitted.

Taking Expectations of Confirmation Measures

As we shall see later, several applications of Bayesian confirmation theory involve calculating the mathematical expectation of some confirmation measure. In general, the *expected value* of some quantity (random variable), $\mathfrak{x}$, that can take values $\mathfrak{x}_1$, $\mathfrak{x}_2$, ..., $\mathfrak{x}_n$, relative to a probability distribution $\mathfrak{p}$, is defined as follows: $E[\mathfrak{x}] = \sum_i \mathfrak{x}_i \mathfrak{p}(\mathfrak{x}_i)$.

Taking the expectation of a confirmation measure presupposes that the confirmation measure is not interpreted as a mere ordinal measure, even if we only care about the ordinal properties of the expectation. This is because the fact that two confirmation measures are ordinally equivalent does not entail that their *expectations* will be ordinally equivalent.

To see why this is the case, suppose more generally that we are interested in the expected value of quantities, $\mathfrak{x}$, $\mathfrak{y}$, $\mathfrak{z}$, etc. What kind of scale must $\mathfrak{x}$, $\mathfrak{y}$, $\mathfrak{z}$, etc. be on in order for us to be able to draw the *ordinal* conclusion that, for example, $E[\mathfrak{x}] \geqslant E[\mathfrak{y}]$? Clearly, in order for us to be able to draw the conclusion that the expected value of $\mathfrak{x}$ really is greater than or equal to the expected value of $\mathfrak{y}$, it must be the case that, for every admissible transformation, $\mathfrak{f}$, of $\mathfrak{x}$ and $\mathfrak{y}$, it is also the case that $E[\mathfrak{f}(\mathfrak{x})] \geqslant E[\mathfrak{f}(\mathfrak{y})]$. Hence, in order for us to draw merely ordinal conclusions from the expected values of $\mathfrak{x}$ and $\mathfrak{y}$, the class of admissible transformations must satisfy the following requirement:

$$E[\mathfrak{x}] \geqslant E[\mathfrak{y}] \implies E[\mathfrak{f}(\mathfrak{x})] \geqslant E[\mathfrak{f}(\mathfrak{y})] \tag{3}$$

But the class of all strictly increasing functions does not satisfy the above require-

ment.[7] In general, if $\mathtt{f}$ is a strictly increasing function, then the following will of course be true:

$$\mathrm{E}[\mathtt{x}] \geqslant \mathrm{E}[\mathtt{y}] \implies \mathtt{f}(\mathrm{E}[\mathtt{x}]) \geqslant \mathtt{f}(\mathrm{E}[\mathtt{y}]) \tag{4}$$

However, (4) does not entail (3) unless the following condition also holds:

$$\mathtt{f}(\mathrm{E}[\mathtt{x}]) \geqslant \mathtt{f}(\mathrm{E}[\mathtt{y}]) \implies \mathrm{E}[\mathtt{f}(\mathtt{x})] \geqslant \mathrm{E}[\mathtt{f}(\mathtt{y})] \tag{5}$$

But there are many strictly increasing functions that violate (5). Hence $\mathtt{x}$, $\mathtt{y}$, $\mathtt{z}$ cannot be on a mere ordinal scale even if we want to draw merely ordinal conclusions from their expected values. In general, we can guarantee that (5) (and therefore also (3)) holds if the class of admissible transformations satisfies the following requirement:

$$\mathtt{f}(\mathrm{E}[\mathtt{x}]) = \mathrm{E}[\mathtt{f}(\mathtt{x})] \tag{6}$$

As it happens, the class of linear functions satisfies (6). Hence, if $\mathtt{x}$, $\mathtt{y}$, and $\mathtt{z}$ are on an interval scale, then that is sufficient for us to be able to draw ordinal conclusions from their expected values.[8]

[7] Here is a simple counter-example. Suppose we have the following probabilities: $\mathtt{p}(\mathtt{H}_1) = 0.5$, $\mathtt{p}(\mathtt{H}_1|\mathtt{E}) = 0.6$, $\mathtt{p}(\mathtt{H}_1|\neg\mathtt{E}) = 0.2$, $\mathtt{p}(\mathtt{E}) = 0.625$, $\mathtt{p}(\mathtt{H}_2) = 0.4$, $\mathtt{p}(\mathtt{H}_2|\mathtt{E}) = 0.2$, $\mathtt{p}(\mathtt{H}_2|\neg\mathtt{E}) = 0.7333$. As can be verified, we have: $\mathrm{E}[\mathtt{d}(\mathtt{H}_1, \mathtt{E})] = 0 = \mathrm{E}[\mathtt{d}(\mathtt{H}_2, \mathtt{E})]$. However, $\mathrm{E}[\mathtt{d}(\mathtt{H}_1, \mathtt{E})^3] < \mathrm{E}[\mathtt{d}(\mathtt{H}_2, \mathtt{E})^3]$.

[8] Indeed, under several reasonable conditions, the class of linear functions is the *only* class that satisfies (6).

2.5 Applications of Bayesian Confirmation Measures that Rely on a Rejection of OET

As I pointed out in Section 2.2, many philosophers have adopted OET, either explicitly or implicitly. At the same time, however, there are also many examples in the literature of applications of Bayesian confirmation theory that implicitly rely on a rejection of OET. To the extent that one wants to make arguments of the sort discussed in this section, one must therefore reject OET.

Case 1: Schlesinger's argument against the difference measure

In Section 2.4, I explained that adopting OET prevents one from being able to set thresholds that can be used to determine whether a difference in confirmation scores is large, moderate or insignificant. As it happens, there are examples of arguments in the literature that implicitly rely on the assumption that such thresholds can be set. In particular, (Schlesinger, 1995, p. 211) presents an argument (repeated and endorsed in Zalabardo (2009)) against the difference measure and in favor of the ratio measure of confirmation. The argument asks us to compare a change in probability from $1/10^9$ to $1/100$ with a change from 0.26 to 0.27. According to Schlesinger and Zalabardo, the first probability shift is intuitively "much greater" than the second one. The ratio measure gets the "right" verdict here, but the difference measure does not. I will briefly return to this example in Section 3.5 of the next chapter,

where I will give my own analysis of the example. The only takeaway from the example that I wish to draw right now is that – as the argument in Section 2.4 shows – Schlesinger and Zalabardo cannot say that the ratio measure judges the shift from $1/10^9$ to 1/100 to be "much greater" than the shift from 0.26 to 0.27 unless the ratio measure is interpreted as something more than just an ordinal measure. Schlesinger and Zalabardo are consequently tacitly rejecting OET.

Case 2: Myrvold's Bayesian account of the virtue of unification

In Section 2.4, I also explained that OET prevents one from being able to say that two confirmation scores are "approximately the same"; only confirmation theses at least as strong as IET enable one to say this (if the absolute distance metric is used to measure distance). However, there are arguments in the literature that rely on the assumption that it is legitimate to talk about two confirmation scores being approximately the same. In particular, Myrvold (2003) (or, more recently, Myrvold (2016)) gives a Bayesian account that purports to show how a unifying hypothesis can sometimes be confirmed more by evidence than a non-unifying hypothesis, and he applies his account to several examples. Myrvold's explanation of the examples relies on the use of both a confirmation measure, c, and a measure of unification U, and he requires that the measures jointly exhibit the following property: If $\mathsf{c}(H_1, E_1) \approx \mathsf{c}(H_2, E_1)$, $\mathsf{c}(H_1, E_2) \approx \mathsf{c}(H_2, E_2)$, and $\mathsf{U}(E_1, E_2; H_1) > \mathsf{U}(E_1, E_2; H_2)$, then $\mathsf{c}(H_1, E_1 \& E_2) > \mathsf{c}(H_2, E_1 \& E_2)$. As argued earlier, the use of approximation signs requires that the confirmation measures not be interpreted as mere ordinal measures.

Myrvold's account can be salvaged even with OET if the approximation signs are replaced by equality signs. But in that case the unrealistic assumption must be made that H_1 and H_2 are independently confirmed to *exactly* the same degree by the evidence.

The next case I will consider comes from Fitelson (1999). Fitelson shows how several arguments given in the literature are sensitive to the choice of confirmation measure because the arguments depend crucially on properties that some measures have but others lack. According to Fitelson, the problem is that these arguments rely on properties that vary between ordinally non-equivalent measures. In the following, I will show that one of the arguments also implicitly relies on a rejection of OET.

Case 3: The Gillies-Popper-Miller argument

Gillies's (1986) reconstruction of an argument due to Popper and Miller (1983) depends on the confirmation measure used having the following decomposition property: $\mathsf{c}(\mathsf{H}, \mathsf{E}) = \mathsf{c}(\mathsf{H} \vee \mathsf{E}, \mathsf{E}) + \mathsf{c}(\mathsf{H} \vee \neg \mathsf{E}, \mathsf{E})$. According to Gillies, this decomposition allows us to neatly separate H's confirmation score into a deductive part and an inductive part. Fitelson points out that the Gillies-Popper-Miller argument depends on the assumption that d is the correct measure, but we can make the further observation that there are measures ordinally equivalent to d that do not have the preceding decomposition property. For example, the measure d^3, which of course is ordinally equivalent to d, does not have the decomposition property. Thus, Gillies's argument does not merely rely on d being the correct measure; it implicitly relies on d^3 *not* being the correct measure. Since d and d^3 are ordinally equivalent, Gillies's argument

is implicitly rejecting OET.

The next case I will consider concerns how the Paradox of the Ravens has been handled in the literature.

Case 4: Solutions to the Paradox of the Ravens

The Paradox of the Ravens is a paradoxical conclusion that arises from the combination of two very reasonable premises: Nicod's Criterion and the Equivalence Condition. Nicod's Criterion says that universal generalizations of the form $\forall x(Ax \rightarrow Bx)$ are confirmed by instances of the form $Ac\&Bc$. The Equivalence Condition says that if e confirms S, then e confirms every sentence logically equivalent to S. Together, the Equivalence Condition and Nicod's Criterion entail a conclusion that seems counter-intuitive, namely that a non-black non-raven confirms the proposition that every raven is black. Since Nicod's Criterion and the Equivalence Condition are widely accepted, the standard solution[9] is to embrace the paradoxical conclusion while explaining it away by conceding that a non-black non-raven confirms the proposition that every raven is black, but only to a "minute degree" (Vranas, 2004).

Standard solutions that have been given to the Paradox of the Ravens clearly violate OET. For example, Fitelson and Hawthorne (2004, pp. 31-7) give a quantitative solution that depends crucially on the non-ordinal properties of the likelihood ratio, l. In particular, their Theorem 4 (p. 34) gives a bound on the ratio of two likelihood ratios that can be violated if we transform the two likelihoods into different but ordinally equivalent measures.

[9]Which, of course, is not the only solution. See Rinard (2014) for instance.

In general, quantitative solutions to the Paradox of the Ravens inevitably reject OET. However, there are also non-quantitative solutions to the Paradox of the Ravens. These solutions have the more modest goal of showing that a non-black non-raven confirms the proposition that all ravens are black *less* than a black raven does, without making the quantitative claim that the confirmation is much less. Since these solutions only make ordinal claims, they do not rely on a violation of OET. However, a proper solution to the Paradox of the Ravens arguably *should* be quantitative. As an analogy, suppose I ask you why the sun looks the size of a tennis ball even though it is so far away, while a tennis ball looks tiny from just 100 yards away. If you answer that it is because the sun is bigger than a tennis ball, you have given me relevant information, but you have not really provided an adequate explanation. Similarly, our intuition in the Paradox of the Ravens is that a non-black non-raven should (in most circumstances) barely, if at all, confirm the proposition that all ravens are black, or at least that it should confirm this proposition much less than a black raven does. A proper solution to the Paradox of the ravens should entail this conclusion, and therefore cannot be just ordinal.

Case 5: The use of mathematical expectations

Several applications of confirmation measures involve taking mathematical expectations of confirmation measures. Indeed, in Chapter 4 of this dissertation I will make extensive use of expectations of confirmation measures. In that chapter I will argue that how one ought to measure the opinionatedness of a probability distribution is intimately tied up with the expected value of various confirmation measures, and I

will argue that how opinionated a probability distribution is in turn partly determines whether the probability distribution is rational. Thus, in order to determine whether a probability distribution is rational, we need to take expectations of confirmation measures.

However, for the remainder of this section, I will focus on two other applications of confirmation measures that rely on taking expected values of confirmation measures.[10]

First, a confirmation score tells you how much a piece of evidence confirms a single hypothesis. However, it's also often interesting to know how much the evidence influences the whole partition of hypotheses; or, in other words, how big the divergence is between the posterior distribution and the prior distribution, given the evidence. The natural way to generalize a confirmation measure to a divergence measure is by taking an expectation. For example, Crupi and Tentori (2014) suggest the following definition:[11]

$$\text{InfDis}(\mathsf{p}(\mathbf{H}|\mathsf{E}), \mathsf{p}(\mathbf{H})) = \sum_{\mathsf{j}} \mathsf{c}(\mathsf{H}_\mathsf{j}, \mathsf{E}) * \mathsf{p}(\mathsf{H}_\mathsf{j}|\mathsf{E}) \tag{7}$$

Plugging different confirmation measures into (7) then gives rise to different divergence measures. For example, plugging in the log-ratio measure gives rise to the well known KL divergence (Kullback and Leibler, 1951). Conversely, any divergence measure may be regarded as an implicit generalization of a confirmation measure. Divergence measures such as the KL divergence have been applied in many ways in the Bayesian literature. For example, they form the foundations of one of the most

[10]For examples of yet other applications, see Good (1985).

[11]Independently, I thought of the same definition, which I make use of in Chapter 4.

prominent versions of objective Bayesianism (Bernardo, 1979b).

Crucially, confirmation measures that are ordinally equivalent will in general not give rise to divergence measures that are ordinally equivalent, for the reasons given in Section 2.4. Indeed, ordinally equivalent confirmation measures can give rise to very different divergence measures. Consider, for example, the log-likelihood measure and the Kemeny-Oppenheim measure. These are ordinally equivalent, but the log-likelihood measure judges distances between probabilities that are close to 0 or 1 to be much larger than does the Kemeny-Oppenheim measure, because the log-likelihood measure is unbounded while the Kemeny-Oppenheim measure is bounded between 0 and 1. Thus, the log-likelihood measure and the Kemeny-Oppenheim measure give rise to divergence measures that will often ordinally disagree if the probabilities involved are extreme (close to 0 or close to 1).[12] Hence, if we want to be able to draw merely ordinal conclusions from Bayesian divergence measures, we cannot treat the confirmation measures on which they are based as mere ordinal measures.

Second, as has recently been pointed out by Brössel and Huber (2014), confirmation measures also have an application in experimental design. More precisely, from a Bayesian point of view, the best experiment to conduct is the one that can be expected to have the greatest confirmational impact, where the expectation is calculated over the prior probabilities of the possible evidence, given the candidate experimental design. The confirmation measure that is standardly used (implicitly) for this purpose in the literature on Bayesian experimental design is the log-ratio measure. Brössel and Huber instead use as their illustration the Kemeny-Oppenheim measure of

[12]Numerical examples are easy to come up with, but tedious. Note also that if there are many hypotheses, then at least some of the probabilities *must* be small.

confirmation. I. J. Good, on the other hand, advocated using the log-likelihood measure for the same purpose (Good, 1985). Interestingly, as was just pointed out, the Kemeny-Oppenheim measure and the log-likelihood measure are ordinally equivalent, and are for that reason generally regarded as equivalent in the philosophical literature. However, as we have seen, the fact that the Kemeny-Oppenheim measure and the log-likelihood measure are ordinally equivalent does not imply that their *expectations* will be ordinally equivalent; hence, the experiment that maximizes expected confirmation with respect to the log-likelihood measure will in general not be equivalent to the experiment that maximizes expected confirmation with respect to the Kemeny-Oppenheim measure.

2.6 On the Goal-Relative Nature of Bayesian Norms

Carnap (1962) first drew the distinction between the "comparative" and "quantitative" questions of confirmation. As the previous sections make clear, we can draw finer distinctions than that. In particular, the interval level occupies an intermediate position between the merely comparative (ordinal level) and the fully quantitative (ratio level). For whatever reason, the comparative question analyzed on the ordinal level has become the question that occupies philosophers' attention. Why is that? One possibility is that some philosophers simply believe the ordinal level to be the most interesting level of analysis. I do not think this belief is warranted, but even if it is granted, the arguments in the previous sections show that the quantitative levels

of analysis are by no means uninteresting. Several well known arguments that make use of confirmation measures implicitly rely on a rejection of OET. Moreover, many conclusions that we want to draw from the output of a confirmation measure may only be legitimately drawn if the measure is assumed to be more than a mere ordinal measure.

A different reason why philosophers may have focused on the ordinal level of analysis is that they think the question of which confirmation measure is ordinally correct must be settled before the more fine-grained question of which confirmation measure has the right quantitative structure can be approached. Although this idea seems intuitive, it is mistaken. Indeed, if we instead start with the desideratum that we want a confirmation measure that we can interpret as, say, an interval measure and not just an ordinal measure, then that puts significant restrictions on the functional form that the confirmation measure can take, as we will see in the next chapter.

Indeed, focusing on the interval level leads to a very different perspective on confirmation measures. By definition, each class of interval equivalent confirmation measures is a proper subset of a class of ordinally equivalent measures; but even so, it is possible for two ordinally non-equivalent confirmation measures to exhibit quantitative behavior that is more similar than the quantitative behavior exhibited by two measures that are ordinally equivalent. For example, from a quantitative point of view, the log-likelihood measure and the log-ratio measure are arguably "more similar" to each other than the log-likelihood measure is to the Kemeny-Oppenheim measure, even though the latter two measures are ordinally equivalent and the first two are not, because the log-likelihood measure and the log-ratio measure will often

have numerically similar outputs.[13] One consequence of this is that the log-ratio and log-likelihood measures arguably give rise to divergence measures that are more similar to each other than are the divergence measures derived from the log-likelihood measure and Kemeny-Oppenheim measure. Focusing only on the ordinal level of analysis therefore leads us to neglect quantitative similarities and dissimilarities that cut across ordinally equivalent classes.

A final probable reason why philosophers have focused their attention on the ordinal level of analysis and have implicitly accepted OET is that many philosophical Bayesians are subjective Bayesians who hold that an agent's probability function is supposed to represent the degrees of belief of the agent. It is already controversial whether agents' degrees of belief have the kind of quantitative structure that probability functions have. Several philosophers have endorsed an "anti-realism" about probabilistic representations of belief states (e.g. Easwaran (2016)).

To reject OET for confirmation measures is apparently to contend that agents' epistemic states have an even more fine-grained structure than is attributed to agents according to probabilism. If, for example, IET is accepted, then not only

[13] In particular, if the hypothesis space is large, it will generally be the case that $\mathsf{p}(\mathsf{E}|\neg\mathsf{H}) \approx \mathsf{p}(\mathsf{E})$, for most H's, and hence the log-likelihood measure and log-ratio measure will have numerically similar outputs. Indeed, if the hypothesis space is parameterized by a continuous parameter, Θ, then, for every $\theta \in \Theta$, we have $\mathsf{l}(\theta, \mathsf{E}) = \log \frac{\Pr(\mathsf{E}|\theta)}{\Pr(\mathsf{E}|\neg\theta)} = \log \frac{\Pr(\mathsf{E}|\theta)}{\int_{\Theta*} \Pr(\mathsf{E}|\theta)\Pr(\theta)\mathrm{d}\theta}$, where $\Theta*$ is Θ with θ taken out. But removing a single point from the parameter space will not have any effect on the integral, so $\int_{\Theta*} \Pr(\mathsf{E}|\theta)\Pr(\theta)\mathrm{d}\theta = \int_{\Theta} \Pr(\mathsf{E}|\theta)\Pr(\theta)\mathrm{d}\theta = \Pr(\mathsf{E})$. Therefore, $\mathsf{l}(\theta, \mathsf{E}) = \log \frac{\Pr(\mathsf{E}|\theta)}{\Pr(\mathsf{E}|\neg\theta)} = \log \frac{\Pr(\mathsf{E}|\theta)}{\Pr(\mathsf{E})} = \mathsf{lr}(\theta, \mathsf{E})$, and so $\mathsf{l}(\theta, \mathsf{E})$ is actually *identical* to $\mathsf{lr}(\theta, \mathsf{E})$ when the hypothesis space is continuous. As far as I know, this fact has not been noted before. On the other hand, the fact that the Kemeny-Oppenheim measure and the log-likelihood measure are ordinally equivalent means that they will always agree on whether $\mathsf{c}(\mathsf{H}, \mathsf{E}) > \mathsf{c}(\mathsf{H}', \mathsf{E})$, but they will often strongly disagree on whether the *difference* between $\mathsf{c}(\mathsf{H}, \mathsf{E})$ and $\mathsf{c}(\mathsf{H}', \mathsf{E})$ is small, large, or trivial; their *interval* judgments are in other words quite different.

do rational agents have degrees of belief that are representable by probabilities; all differences between differences (according to some confirmation measure) of an agent's probability function also represent actual features of the epistemic state of the agent. For Bayesians who already worry about the psychological realism of probabilistic degrees of belief, the complex structure seemingly attributed to agents' epistemic states according to IET may be a bridge too far.

On the other hand, it is undoubtedly the case that we sometimes do have quantitative intuitions about confirmation, so there is a basis in human epistemological experience for looking at the quantitative structure of confirmation measures. For example, in the case of the paradox of the ravens, our intuition is that a white shoe confirms the the proposition that all ravens are black *much less* than a black raven does. And we often feel that a piece of evidence fails to really discriminate between two hypotheses, so that the hypotheses are intuitively confirmed to roughly the same extent.

Of course, the fact that we sometimes have strong quantitative intuitions about confirmation does not mean we always do. But nor, should it be added, do we always have strong ordinal intuitions. The Bayesian framework idealizes away these human limitations, but the norms of Bayesianism presumably still hold for more limited agents whenever the norms are applicable. Indeed, Bayesian norms, more generally, can fruitfully be construed as conditional norms. For example, even though human beings lack a completely ordered set of degrees of beliefs, Bayesian Dutch book arguments tell you that, provided you do have degrees of belief and you intend to use them in order to choose which bets to accept or reject, and you want to avoid

sure losses, then your degrees of belief need to be probabilistic. Accepting this norm does not entail believing that your degrees of belief are always probabilistic or even that it is always rationally required of you to have credences that are probabilistic.[14] However, if you make it your goal to use a set of credences to choose how to act, then accepting the norm implies you have to accept that the credences that are relevant to your actions, at least, ought to be probabilistic.

Accuracy-based arguments for probabilism (e.g. Joyce (1998)) may reasonably be construed as establishing a similar conditional norm: if your goal is to have accurate credences in a set of propositions, then your credences in those propositions need to be probabilistic.[15] But if you do not care about the truth value of some proposition, or you are not attending to the proposition, then the conditional norm does not apply to you with respect to that proposition.

In the same way, the arguments in this chapter establish the following conditional norm: if you have a set of confirmation scores and you intend to interpret the scores in a certain way (e.g. to say that some of them are approximately the same) or use them in a certain way (e.g. take their expectations), then your confirmation scores cannot be on just an ordinal scale. This conditional norm only "kicks in" if you use your confirmation measure in certain ways, and accepting the norm does not entail accepting that your confirmation judgments always will be or always ought to be on an interval scale. The norm therefore does not make unrealistic presuppositions about human psychology, nor does it make unreasonable demands.

[14]Of course, many Bayesians want to argue for this stronger unconditional norm as well.

[15]Of course, philosophers often want to go further; they want to say, for example, that you *ought* to have the goal of avoiding sure losses or having accurate credences.

2.7 Conclusion

I have argued that a commonly accepted thesis in Bayesian confirmation theory – the Ordinal Equivalence Thesis – must be rejected if confirmation measures are to be applied in many of the ways we want to apply them. This shows that how a Bayesian confirmation measure is applied is relevant to whether a normative thesis is rationally binding or not. I have moreover argued that all Bayesian norms can fruitfully be construed to be goal-relative in just this way. The upshot of this chapter is therefore a mild sort of pragmatic encroachment. In the next chapter, we will see that the general goal of having a confirmation measure that is interpretable as an interval measure puts constraints on which confirmation measure it is rational for you to use. In Chapter 4 and Chapter 5, we will see more radical examples of pragmatic encroachment.

3 A New Characterization of Confirmation Measures

3.1 Goals of the Chapter

In the previous chapter I showed how several applications of confirmation measures implicitly rely on a rejection of the Ordinal Equivalence Thesis – the thesis that says that two confirmation measures are equivalent if they rank-order all hypothesis-evidence pairs in the same way. Instead, these applications require that our confirmation measure, c, be at least an interval measure, so that quantities such as $\mathsf{c}(\mathsf{H}, \mathsf{E}) - \mathsf{c}(\mathsf{H}', \mathsf{E})$ are interpretable.

Suppose we want a confirmation measure that we are epistemically justified in interpreting at least as an interval measure. What kinds of requirements does this desideratum place on our measure? In this chapter I will show that, in fact, the desideratum places strict requirements on the functional form that the confirmation measure may take. In particular, there are three natural ways in which we may quantify what counts as a small shift in probability, and given that we want our measure to be an interval measure, each of these ways of quantifying a probability shift entails a well

known confirmation measure. In the next chapter, I will show that two of the three ways of measuring probability shifts are appropriate given two different purposes. It follows that the two purposes call for different confirmation measures. However, in this chapter, the main goal is simply to show that three natural ways of measuring a shift in probability entail the three most well known confirmation measures, provided we want the confirmation measures to be interpretable as interval measures. Consequently, the chapter can also be regarded as providing a new characterization of the three most well known confirmation measures. This supplements the characterization of the same confirmation measures provided by Crupi et al. (2013).

I start by laying out my background assumptions in Section 3.2. In Section 3.3, I make the requirements on the confirmation measure, c, more precise. In Section 3.4, I show how these requirements entail that c is either the difference measure, d, the log-ratio measure, lr, or the log-likelihood measure, l, depending on how we quantify what counts as a small probability shift. In Section 3.5, I discuss the implications of the argument and consider a couple of objections.

3.2 Background Assumptions

Recall that a confirmation measure, c is a measure of the distance between the posterior probability of a hypothesis, $\mathsf{p}(\mathsf{H}|\mathsf{E})$, and the prior probability of the same hypothesis, $\mathsf{Pr}(\mathsf{H})$, that is supposed to measure the degree to which E confirms H. In this chapter, the following three confirmation measures will be particularly important, so I list them again for the reader's convenience:

The log-ratio measure, $\mathsf{lr}(\mathsf{H},\mathsf{E}) = \log \mathsf{r}(\mathsf{H},\mathsf{E})$

The difference measure, $\mathsf{d}(\mathsf{H},\mathsf{E}) = \mathsf{Pr}(\mathsf{H}|\mathsf{E}) - \mathsf{Pr}(\mathsf{H})$

The log-likelihood measure, $\mathsf{l}(\mathsf{H},\mathsf{E}) = \log \frac{\mathsf{Pr}(\mathsf{E}|\mathsf{H})}{\mathsf{Pr}(\mathsf{E}|\neg\mathsf{H})}$

Since Bayesians analyze confirmation in terms of probability, and since the probability distribution over the algebra generated by H and E is determined by $\mathsf{Pr}(\mathsf{H}|\mathsf{E})$, $\mathsf{Pr}(\mathsf{H})$, and $\mathsf{Pr}(\mathsf{E})$, it has become standard to assume that any confirmation measure can be expressed as a function of $\mathsf{Pr}(\mathsf{H}|\mathsf{E})$, $\mathsf{Pr}(\mathsf{H})$, and $\mathsf{Pr}(\mathsf{E})$. The preceding assumption is essentially the requirement that Crupi et al. (2013) call "formality." A strong case can however be made for not allowing our measure of confirmation to depend on $\mathsf{Pr}(\mathsf{E})$. As Atkinson (2009) points out, if we let $\mathsf{c}(\mathsf{H},\mathsf{E})$ be a function of $\mathsf{Pr}(\mathsf{E})$, then $\mathsf{c}(\mathsf{H},\mathsf{E})$ can change even if we add to E a piece of irrelevant "evidence" E' that is informationally independent of H and E, and of their conjunction. To see this, suppose that $\mathsf{c}(\mathsf{H},\mathsf{E}) = \mathsf{f}(\mathsf{Pr}(\mathsf{H}), \mathsf{Pr}(\mathsf{H}|\mathsf{E}), \mathsf{Pr}(\mathsf{E}))$. Let E' be any proposition whatsoever that is informationally (and hence probabilistically) independent of H, E, and $\mathsf{H}\&\mathsf{E}$. Then $\mathsf{c}(\mathsf{H},\mathsf{E}\&\mathsf{E}') = \mathsf{f}(\mathsf{Pr}(\mathsf{H}), \mathsf{Pr}(\mathsf{H}|\mathsf{E}\&\mathsf{E}'), \mathsf{Pr}(\mathsf{E}\&\mathsf{E}')) = \mathsf{f}(\mathsf{Pr}(\mathsf{H}), \mathsf{Pr}(\mathsf{H}|\mathsf{E}), \mathsf{Pr}(\mathsf{E})\mathsf{Pr}(\mathsf{E}'))$. If f depends on the third argument, we can find some probability function Pr that respects the above independences such that $\mathsf{f}(\mathsf{Pr}(\mathsf{H}), \mathsf{Pr}(\mathsf{H}|\mathsf{E}), \mathsf{Pr}(\mathsf{E})\mathsf{Pr}(\mathsf{E}')) \neq \mathsf{f}(\mathsf{Pr}(\mathsf{H}), \mathsf{Pr}(\mathsf{H}|\mathsf{E}), \mathsf{Pr}(\mathsf{E}))$, and thus such that $\mathsf{c}(\mathsf{H},\mathsf{E}\&\mathsf{E}') \neq \mathsf{c}(\mathsf{H},\mathsf{E})$. However, this is clearly counterintuitive, since E' is informationally independent of H and E and therefore should not have any impact on the confirmation of H. So we conclude that f should not depend on $\mathsf{Pr}(\mathsf{E})$.

Since I find the preceding argument convincing, I will assume that the confirmation measure we are looking for is of the following form: $\mathsf{c}(\mathsf{H},\mathsf{E}) = \mathsf{f}(\mathsf{Pr}(\mathsf{H}),\mathsf{Pr}(\mathsf{H}|\mathsf{E}))$. Since there is no *a priori* restriction on what credences an agent may have except that these credences must lie somewhere in the interval $[0,1]$, I will assume that f is defined on all of $[0,1]*[0,1]$. Note that, as Huber (2008) points out, this is not the same as assuming that any particular probability distribution $\mathsf{Pr}(*)$ is continuous.

The preceding two assumptions are summed up in the following requirement:

Strong Formality (SF). *Any confirmation measure is of the following form:* $\mathsf{c}(\mathsf{H},\mathsf{E}) = \mathsf{f}(\mathsf{Pr}(\mathsf{H}),\mathsf{Pr}(\mathsf{H}|\mathsf{E}))$*, where* f *is a function defined on all of* $[0,1]*[0,1]$*.*

It should be noted that (SF) excludes some of the confirmation measures that have been offered in the literature.[1] I briefly address lingering objections to (SF) in Section 3.5.

Finally, I will also adopt the following convention:

Confirmation Convention (CC).

$$\mathsf{c}(\mathsf{H},\mathsf{E}): \begin{cases} > 0 & \textit{if } \mathsf{Pr}(\mathsf{H}|\mathsf{E}) > \mathsf{Pr}(\mathsf{H}), \\ = 0 & \textit{if } \mathsf{Pr}(\mathsf{H}|\mathsf{E}) = \mathsf{Pr}(\mathsf{H}), \\ < 0 & \textit{if } \mathsf{Pr}(\mathsf{H}|\mathsf{E}) < \mathsf{Pr}(\mathsf{H}). \end{cases}$$

(CC) is sometimes taken to be part of the definition of what a confirmation measure is (e.g. by Fitelson (2001)). (CC) has the role of setting 0 as the number that signifies confirmation neutrality.

[1] In particular, the alternative difference measure.

3.3 The Main Requirement on c

Suppose we witness a coin being flipped 10 times, and our task is to assign a credence to the proposition that the coin comes up heads on the 11th flip. If we do not in advance know anything about the coin's bias, it is reasonable to guess that the coin will come up heads with probability $H/10$ on the 11th flip, where H is the number of times the coin comes up heads in the 10 initial flips.[2] In making this guess, we are setting our credence in the coin landing heads equal to the observed frequency of heads. This move is reasonable in part because the law of large numbers guarantees that the observed frequency of heads converges in probability to the coin's actual bias. The observed frequency of heads does not necessarily equal the coin's bias after just 10 flips, however. In fact, classical statistics tells us that the confidence interval around the observed frequency can be approximated by $\hat{p} \pm z\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$, where $\hat{p}$ is the observed frequency, n is the sample size (in this case, 10 coin flips), and z is determined by our desired "confidence level."[3]

For example, suppose we witness 4 heads in 10 coin flips and we set our confidence level to 95%. In that case, $z = 1.96$, $\hat{p} = 0.4$, and the calculated confidence interval is approximately $[0.1, 0.7]$. Clearly, the confidence interval in this case is rather large. Given our evidence, it is reasonable to estimate the coin's bias as 0.4. However, we also need to realize that if the 10 flips were repeated, we would probably end up with a slightly different value for $\hat{p}$: we should acknowledge that credences are bound to

[2] This assumes $0 < H < 10$.

[3] "Confidence" here should not be interpreted in the (standard) Bayesian sense as a degree of belief: rather, the "confidence level" represents the degree of error due to variability that we are willing to tolerate.

vary with our varying evidence.

The above example illustrates one way that variability can sneak into our credences: if our credence is calibrated to frequency data, then our credence inherits the variability intrinsic to the frequency data. However, even if we set our credence by other means than frequency data, we must admit that rational credences are intrinsically somewhat variable. For example, if the sky looks ominous and I guess that there is a 75% chance that it is going to rain (or perhaps my betting behavior reveals that this is my credence that it is going to rain), I must concede that another agent whose credence (or revealed credence) is 74% or 76% is just as rational as I am: I do not have either the evidence nor the expertise to discriminate between these credences. And even if I do have good evidence as well as expertise, I must admit that I am almost never in a position where I have *all* the evidence, and had I been provided with somewhat different evidence, I would have ended up with a somewhat different credence.

The fact that our credences are variable is a fact of life that any rational agent must face squarely. It is not hard to see that this fact also affects Bayesian confirmation theory. Bayesian confirmation measures are defined in terms of credences, and are therefore infected by the variability inherent in credences. If Bayesian confirmation measures are necessarily affected by variable credences, I contend that we should want a confirmation measure that is affected by such variability in a systematic and predictable way. This is particularly so if we want our confirmation measure to be interpretable as an interval measure. Recall that if c is an interval measure, then we want to be able to draw conclusions from differences of the sort $\mathsf{c}(\mathsf{H}, \mathsf{E}) - \mathsf{c}(\mathsf{H}', \mathsf{E}')$.

However, if c is very sensitive to small shifts in the priors or posteriors of H

and H', then the quantity $c(H, E) - c(H', E')$ is unstable: it could easily have been different, since our priors or posteriors could easily have been slightly different (for instance if we calibrated our priors to frequency data). We are therefore only justified in interpreting the difference $c(H, E) - c(H', E')$ if c is relatively *insensitive* to small shifts in the priors and posteriors.

Suppose, moreover, that slight shifts in small priors (or posteriors) have a larger effect on c's output than do slight shifts in larger priors. Then we cannot compare the quantity $c(H, E) - c(H', E)$ to the quantity $c(H'', E) - c(H', E)$ unless our prior credences in H'' and H are approximately the same. In order for us to be able to compare $c(H, E) - c(H', E)$ to $c(H'', E) - c(H', E)$ in cases where our prior credences in H'' and H are very different, we need c to be *uniformly* insensitive to small shifts in the prior (and the posterior). We can sum up the preceding two remarks as follows:

Main Requirement (MR). *We are justified in interpreting and drawing conclusions from the quantity* $c(H, E) - c(H', E')$ *only if* c *is uniformly insensitive to small variations in* $Pr(H)$ *and* $Pr(H|E)$.

As it stands, (MR) is vague. What counts as a small variation in a credence? Moreover, what does it mean, concretely, for c to be *uniformly insensitive* to such variations? To get a better handle on these questions, let us formalize the important quantities that occur in (MR). Following (SF), we are assuming that c is of the form $c(H, E) = f(Pr(H), Pr(H|E))$. For simplicity, let us put $Pr(H) = x$ and $Pr(H|E) = y$, so that $c = f(x, y)$. According to (MR), we require that f be uniformly insensitive to small variations in x and y. I will use $\nu(p, \epsilon)$ to capture the notion of a small variation in the probability p, where ϵ is a parameter denoting the size of the variation.

Moreover, I will use $\Delta^{x}_{\epsilon}c(x,y)$ to denote the variation in c that results from a variation of size ϵ about x. That is to say,

$$\Delta^{x}_{\epsilon}c(x,y) = f(\nu(x,\epsilon),y) - f(x,y) \tag{1}$$

Similarly, I will use $\Delta^{y}_{\epsilon}c(x,y)$ to denote the variation in c that results from a variation of size ϵ about y. Thus,

$$\Delta^{y}_{\epsilon}c(x,y) = f(x,\nu(y,\epsilon)) - f(x,y) \tag{2}$$

The next step is to get a better grip on (MR) by investigating the terms that occur in (1) and (2). In the next subsections, that is what I do.

What is uniform insensitivity?

First, the demand that c be *uniformly* insensitive to variations in the prior and the posterior now has an easy formal counterpart: it is simply the demand that for different values x_1,x_2,y_1, and y_2 of x and y, we have $\Delta^{x_1}_{\epsilon}c(x_1,y_1) = \Delta^{x_2}_{\epsilon}c(x_2,y_2) = \Delta^{x_2}_{\epsilon}c(x_2,y_1) = etc.$ and $\Delta^{y_1}_{\epsilon}c(x_1,y_1) = \Delta^{y_2}_{\epsilon}c(x_2,y_2) = \Delta^{y_2}_{\epsilon}c(x_1,y_2) = etc.$ Thus, across different values of x and y, a small variation in c will mean the same thing. More importantly, this means we can consider $\Delta^{x}_{\epsilon}c(x,y)$ as purely a function of ϵ, and likewise for $\Delta^{y}_{\epsilon}c(x,y)$. From now on, I will therefore write:

$$g(\epsilon) := \Delta^{x}_{\epsilon}c(x,y) \tag{3}$$

$$h(\epsilon) := \Delta^{y}_{\epsilon}c(x,y) \tag{4}$$

In order to figure out what the requirement that $\mathfrak{c}$ be insensitive to *small* variations amounts to, we need to figure out how to quantify variations in credences. It is to this question that I now turn.

What is a small shift in a probability?

Given a probability $\mathfrak{x}$, what counts as a small shift in $\mathfrak{x}$? This question turns out to have a more subtle answer than one might expect. Using the notation from equations (1) and (2), what we are looking for is the form of the function $\mathfrak{v}(\mathfrak{x}, \epsilon)$. Perhaps the most natural functional form to consider is the following one: $\mathfrak{v}(\mathfrak{x}, \epsilon) = \mathfrak{x} + \epsilon$. On this model, a small shift in probability $\mathfrak{x}$ is modeled as the addition of a (small positive or negative) number to $\mathfrak{x}$. According to the additive model, whether ϵ counts as a small shift in $\mathfrak{x}$ does not depend on the size of $\mathfrak{x}$. For example, suppose $\epsilon = 0.05$ is a small shift relative to 0.5. Then, according to the additive mode, 0.05 will also count as a small shift relative to, say, $\mathfrak{x} = 0.00001$. This feature of the additive model may not always be sensible.

An alternative to the additive model is to scale the size of the shift with the size of $\mathfrak{x}$. In other words, we might suggest the following form for $\mathfrak{v}$: $\mathfrak{v}(\mathfrak{x}, \epsilon) = \mathfrak{x} + \mathfrak{x}\epsilon$. This adjustment gets rid of the feature mentioned in the previous paragraph. According to the new $\mathfrak{v}$, a shift of size 0.025 relative to 0.5 is "equal" to a shift of 0.0000005 relative to 0.00001. In contrast to the previous additive model, $\mathfrak{v}(\mathfrak{x}, \epsilon) = \mathfrak{x} + \mathfrak{x}\epsilon$ is a "multiplicative" model, as we can see by instead writing it in the following form: $\mathfrak{v}(\mathfrak{x}, \epsilon) = \mathfrak{x}(1 + \epsilon)$

In Chapter 4, we will see that both of the above models of a probability shift make

sense given different goals that you may have. However, there is a third interesting model that is also worth considering, and this model is appealing on several grounds.

Note, first of all, that the additive and multiplicative model both have certain problems. One problem is purely mathematical. Since $\nu(\chi, \epsilon)$ is supposed to correspond to a small shift in probability, we should require that $0 \leqslant \nu(\chi, \epsilon) \leqslant 1$, for all values of χ and ϵ. However, $\chi + \chi\epsilon$ can easily be larger than 1, for example if $\chi = 0.9$ and $\epsilon = 0.2$.[4] The other problem is that $\nu(\chi, \epsilon)$ treats values of χ close to 0 very differently from values of χ close to 1. For instance, a shift where $\epsilon = 0.1$ will be scaled to 0.001 when $\chi = 0.01$. But when $\chi = 0.99$, the same ϵ will be scaled to just 0.099. This seems problematic, since for every hypothesis H in which we have a credence of 0.99, there corresponds a hypothesis in which we have a credence of 0.01, namely $\neg H$. But a small shift in our credence in H is necessarily also a small shift in our credence in $\neg H$, simply because $\mathsf{Pr}(\neg H) = 1 - \mathsf{Pr}(H)$: intuitively, H and $\neg H$ should therefore be treated symmetrically by ν.

Neither of the preceding problems is a "deal breaker," in my view. However, they motivate the search for a model of probability shifting that avoids the problems. And in fact, there is an easy fix to both of the preceding problems: if we scale ϵ by $\chi(1 - \chi)$ instead, then first of all we have $0 \leqslant \chi + \epsilon\chi(1 - \chi) \leqslant 1$, and thus $0 \leqslant \nu(\chi, \epsilon) \leqslant 1$. Second of all, H and $\neg H$ are now treated symmetrically. From the preceding considerations, we therefore end up with the following as our functional form for ν: $\nu(\chi, \epsilon) = \chi + \chi(1 - \chi)\epsilon$.

There is a completely different argument by which we can arrive at the same

[4]This is also a problem for the additive model.

functional form for ν. As I mentioned in the example at the beginning of Section 2.3, credences are sometimes calibrated to frequency data. This is for example usually the case if H is a medical hypothesis. Suppose H represents the hypothesis that a person P has disease X, for instance. The rational prior credence in H (before a medical examination has taken place) is then the frequency of observed cases of X in the population from which P is drawn. The frequency of observed cases of X can be modeled as the outcome of a binomial process having mean $\mathsf{Pr}(\mathsf{H})$ and variance proportional to $\mathsf{Pr}(\mathsf{H})(1-\mathsf{Pr}(\mathsf{H}))$. Suppose we observe the frequency $\mathsf{fr}(\hat{\mathsf{H}})$. Then the estimated variance is proportional to $\mathsf{fr}(\hat{\mathsf{H}})(1-\mathsf{fr}(\hat{\mathsf{H}}))$. The variance is maximal at $\mathsf{fr}(\hat{\mathsf{H}}) = 0.5$ and decreases as $\mathsf{fr}(\hat{\mathsf{H}})$ moves closer to 0 or to 1. Arguably, it makes a lot of sense to scale the size of the probability shift with the variance in the frequency data. But that is exactly what $\nu(\chi, \epsilon) = \chi + \chi(1-\chi)\epsilon$ does: it scales what counts as a small shift in a probability with the variance in the probability due to the variance of the data.

Thus the third model of probability shifting that I will consider has the following functional form:

$$\nu(\chi, \epsilon) = \chi + \chi(1-\chi)\epsilon \tag{5}$$

I will refer to this model as the "variance model" of probability shifting.

Uniform insensitivity to small variations in the prior and posterior

The next step is to understand what insensitivity amounts to. To say that c is *insensitive* to small shifts in the prior or posterior is to say that such shifts have a small effect on confirmation: the most natural way to formalize this requirement is in terms of continuity. Since $\mathsf{g}(\epsilon)$ represents the change in confirmation resulting from a change (by ϵ) in probability, a natural continuity requirement for c would be that g and h should be continuous at 0.

However, continuity is too weak a requirement. Even if a function is continuous, it is still possible for it to be very sensitive to small shifts. For instance, the function $\mathsf{f}(\mathsf{x}) = 1000000^{\mathsf{x}}$ is continuous (everywhere), but is at the same time very sensitive to small perturbations of x. Sensitivity to perturbations is most naturally measured by looking at how the derivative behaves. Minimally, we should therefore require that g and h be differentiable at 0. The next natural requirement would be to demand that the derivative of both g and h be *bounded* by some "small" number. Of course, pursuing such a requirement would require a discussion of what is to count as a "small" number in this context. Since I do not actually need a requirement of this sort in my argument in the next section, I will not pursue a discussion of these issues here. The only upshot from this section is therefore that g and h should be differentiable at 0.

3.4 The Main Result

Let me summarize where we are. Our desire to be able to draw conclusions from differences in confirmation, i.e. from expressions of the form $\mathsf{c}(\mathsf{H},\mathsf{E}) - \mathsf{c}(\mathsf{H}',\mathsf{E}')$, led us to the requirement that c be *uniformly insensitive to small shifts* in $\mathsf{Pr}(\mathsf{H})$ and $\mathsf{Pr}(\mathsf{H}|\mathsf{E})$. In the previous subsections, I made the various components of this requirement more precise. Putting all these components together, we have the following:

Formal Version of the Main Requirement (MR) 3.4.1. *We are justified in drawing conclusions from the difference* $\mathsf{c}(\mathsf{H},\mathsf{E}) - \mathsf{c}(\mathsf{H}',\mathsf{E}')$ *only if the following conditions are all met:*

1. $\mathsf{f}(\nu(\mathsf{x},\epsilon),\mathsf{y}) - \mathsf{f}(\mathsf{x},\mathsf{y}) = \mathsf{g}(\epsilon)$, *where:*
2. $\mathsf{g}(\epsilon)$ *does not depend on either* x *or* y
3. $\mathsf{g}(\epsilon)$ *is differentiable at 0*
4. $\nu(\mathsf{x},\epsilon)$ *is an appropriate measure of a probability shift*
5. $\mathsf{f}(\mathsf{x},\nu(\mathsf{y},\epsilon)) - \mathsf{f}(\mathsf{x},\mathsf{y}) = \mathsf{h}(\epsilon)$, *where:*
6. $\mathsf{h}(\epsilon)$ *does not depend on either* x *or* y
7. $\mathsf{h}(\epsilon)$ *is differentiable at 0*

Note that 5 - 7 are just 1 - 3 except that they hold for h instead of for g. Note also that (MR) is essentially *epistemic*. It says that "we" (i.e. agents interested in confirmation) are only justified in drawing conclusions (of any kind) from $\mathsf{c}(\mathsf{H},\mathsf{E}) -$

$\mathsf{c}(\mathsf{H}', \mathsf{E}')$ if certain formal conditions are met. These conditions ensure that $\mathsf{c}(\mathsf{H}, \mathsf{E})$ behaves reasonably well. Together with (SF) and (CC), the conditions in (MR) entail the main result of chapter (proof in the appendix):

Main Result 3.4.1. *If (MR) is true, (SF) is assumed, and (CC) is adopted as a convention, then the following hold:*

First, if the additive model of probability shifting is used, we have:

$$\mathsf{c}(\mathsf{H}, \mathsf{E}) = \mathsf{Pr}(\mathsf{H}|\mathsf{E}) - \mathsf{Pr}(\mathsf{H})$$

Second, if the multiplicative model of probability shifting is used, we have:

$$\mathsf{c}(\mathsf{H}, \mathsf{E}) = \log \frac{\mathsf{Pr}(\mathsf{H}|\mathsf{E})}{\mathsf{Pr}(\mathsf{H})}$$

Finally, if the variance model of probability shifting is used, we have:

$$\mathsf{c}(\mathsf{H}, \mathsf{E}) = \log \frac{\mathsf{Pr}(\mathsf{E}|\mathsf{H})}{\mathsf{Pr}(\mathsf{E}|\neg\mathsf{H})}$$

In each case, the equality is unique up to multiplication by a positive number.

3.5 Discussion and Objections

In the previous section, I showed that (MR), (SF), and (CC) jointly entail three different confirmation measures, given three different ways of measuring probability shifts. The proof entails each confirmation measure up to multiplication by a positive number. That is to say, if, say, $\log \frac{\mathsf{Pr}(\mathsf{E}|\mathsf{H})}{\mathsf{Pr}(\mathsf{E}|\neg\mathsf{H})}$ is a legitimate confirmation measure,

then so is $\mathfrak{a} * \log \frac{Pr(E|H)}{Pr(E|\neg H)}$, for $\mathfrak{a} > 0$; the argument does not establish that any particular logarithmic base is better than another. In Stevens's (1946) terminology, the measure is apparently *ratio*, meaning that we are justified in interpreting both intervals and ratios between outputs of the measure. Analogously, mass is also a ratio measure since it makes sense to say both that the difference between 2kg and 4kg is the same as the difference between 4kg and 6kg, and that 4kg is twice as big as 2kg. It therefore appears that my conclusion is stronger than what I set out to establish: in the introduction, I said that the goal was to find conditions that entail that a confirmation measure that can be interpreted as *at least* an interval measure. But the proof in the previous section actually establishes that the confirmations are ratio measure under the conditions specified.[5]

The reader may wonder what has been gained from this exercise: have we not just replaced the problem of figuring out which confirmation measure is appropriate with the problem of figuring out how to best measure a small shift in probability? As I will argue in Chapter 4, the confirmation measure that it is appropriate to use depends on your goals; having a characterization of confirmation measures in terms of probability shifts is useful because sometimes it's easy to see which measure of a probability shift is most appropriate given your goals.

As an example, and to give a taste of one of the arguments that will be given in Chapter 4, let's consider Schlesinger's argument against the difference measure again in a bit more detail.[6] Here is Zalabardo's (2009) description of the argument:

"Schlesinger asks us to compare two scenarios. In the first, we consider

[5]I owe thanks to a reviewer for *Philosophy of Science* who persuaded me of this.

[6]I briefly discussed this argument in Chapter 2, p. 31.

> a type of aircraft which is regarded as extremely safe, with a $1/10^9$ probability of crashing in a single flight. However, further inspection of the structure of the aircraft reveals a flaw as a result of which the probability of one of these planes crashing is actually 1/100. The second scenario concerns troops landing gliders behind enemy lines. We start from the assumption that someone taking part in one of these operations has a 26% chance of perishing, but one day the commander announces that owing to peculiar weather conditions the risk has increased from 26% to 27%" (pp. 631-632).

Schlesinger and Zalabardo claim that the change from $1/10^9$ to 1/100 in the first scenario is intuitively much greater than the shift from .26 to .27 in the second scenario. Should we agree with this assessment? One thing that complicates matters is the fact that the argument asks us to compare two different scenarios – the first scenario deals with plane crashes and the second scenario deals with casualties on the battlefield. To remove this complication, let's suppose that we are just dealing with battlefield casualties, so that we are interested in comparing a shift from $1/10^9$ to 1/100 in the risk that someone taking part in battlefield activities will perish vs a shift from .26 to .27 that someone taking part in battlefield activities will perish. Given this fixed context, which of the preceding probability shifts is bigger?

Suppose – as is plausible – that what you ultimately care about is the number of expected lives that will be lost, and suppose – for concreteness – that you have 100 soldiers and that you consider each soldier's life equally valuable. Then a shift in probability from $1/10^9$ to 1/100 entails that the expected number of lives lost will

increase by 1. Similarly, a shift in probability from .26 to .27 also entails that the expected number of lives lost will increase by 1. Thus, if what you care about is the expected number of lives lost, then these two shifts in probability will be equally big for you. Since the shifts are equally big – because they have the same effect on the expected loss calculation – it follows that the additive model of probability shifts is appropriate, and it follows in turn from the arguments provided in this chapter that the difference measure of confirmation is appropriate. However, if you have some different goal, then it's not necessarily the case that the preceding probability shifts will be equally big, as we will see in Chapter 4.

More generally, whether a shift in probability is "big" or "small" can only be determined given some specified goal. Once the goal is specified, it's often obvious how one should appropriately measure the size of probability shits; this in turn settles which confirmation measure it's appropriate to use, and – given the arguments in this chapter – the resulting confirmation measure will be interpretable as an interval measure.

I will end by considering a couple of objections to the argument presented in the previous section. First, my argument is obviously only sound if the assumptions in (MR) are correct. However, the assumptions in (MR) might remind the reader of assumptions made in Good (1960, 1984) and Milne (1996). These assumptions have been criticized by Fitelson (2001, 2006) as being "strong and implausible" (2001, 28-29n43) and for having "no intuitive connection to material desiderata for inductive logic" (2006, 7n13).

Why does my argument escape Fitelson's criticisms? How is my argument different

from the arguments made by Good and Milne? The answer is that, whereas Good and Milne are not interested in the *interval* properties of their confirmation measures, and the various mathematical assumptions they make therefore seem unmotivated, all the properties listed in (MR) arise naturally out of our wish to have a confirmation measure that is at least an interval measure. And recall from the previous chapter that several straightforward applications of confirmation measures do rely on the relevant confirmation measure being an interval measure. Hence, we see that the goal you have (namely how you intend to use the confirmation measure) can in fact turn what seem like "strong and implausible" quantitative assumptions into reasonable *rational* constraints. As I argued at the end of the previous chapter, these constraints are only conditionally applicable to you: if you do not intend to use the confirmation measure in a way that requires the measure to be interpretable as an interval measure, then the constraints do not apply to you.

Finally, one may object to some of the other background assumptions I make in Section 3.2. In particular, *Strong Formality* may be accused of being too strong since it excludes the alternative difference measure right off the bat. My reply to this objection is as follows: the argument in Section 3.4 can be carried out without *Strong Formality*, but the resulting analysis does not yield the alternative difference measure, nor any other recognizable confirmation measure. Thus, even if one rejects (SF), one cannot use the type of argument I have given in this chapter to argue for the alternative difference measure or other standard measures that depend non-trivially on $\mathsf{Pr}(\mathsf{E})$.[7]

[7]Such as Carnap's measure, $\mathsf{c}(\mathsf{H},\mathsf{E}) = \mathsf{Pr}(\mathsf{H}\&\mathsf{E}) - \mathsf{Pr}(\mathsf{H})\mathsf{P}(\mathsf{E})$.

3.6 Conclusion

I have argued that there is a set of conditions that any confirmation measure must meet in order to justifiably be interpreted as an interval measure. Furthermore, I have shown that these conditions, together with three plausible ways of modeling a small shift in probability, jointly entail the three most common confirmation measures: the difference measure corresponds to an additive model; the log-ratio measure corresponds to a multiplicative model; and the log-likelihood measure corresponds to a variance-scaled model. Each of these models of probability shifting may be sensible given different goals that we may have. In the next chapter, I will consider two different goals – learning the truth and ranking actions by their expected utility – and I will show that the log-ratio measure and the difference measure, respectively, are rational measures of confirmation given these two goals.

4 The Goal-Relativity of Prior Probabilities

4.1 Goals of the Chapter

The main goal of this chapter is ultimately to show that the nature of an agent's goal influences what probability distribution it is rational for the agent to have. In order to reach this conclusion I will first argue that the confirmation measure it is rational for the agent to use depends on the goal of the agent. Next I will show how information measures can be derived in a natural way from confirmation measures. Finally, I will show that whether a probability distribution is rational for the agent depends on the agent's information measure.

As a motivating example, to which I will return later in the chapter, suppose you are about to roll a six-sided die (with faces numbered one through six) and you want a probability distribution that represents how probable each of the six possible outcomes is.[1] I have rolled the die many times already, and I tell you that – on

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

average – the die has landed on 5. Clearly, the die is strongly biased towards landing on high numbers, and it seems intuitively probable that the die will land on a high number on the next roll as well. But how do you come up with precise probabilities for each of the possible outcomes? This is an instance of the so-called "problem of the priors": how do you translate background information into a probability distribution, and – in the absence of background information – how do you represent a lack of information probabilistically? I will argue that how you should answer these questions depends on what goals you have.

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

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

which probability distribution most faithfully represents an agent's epistemic state.

4.2 Information Measures and their Importance for Bayesianism

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

An "information measure" is a quantitative measure of how "informative" or "opinionated" a probability distribution is.[2] The most well known information measure is the Shannon entropy, which says that the information content in $p(\mathbf{H})$ is given by $-\sum_j p(H_j) \log p(H_j)$. The higher the Shannon entropy, the less informative and less opinionated is the probability distribution. The probability distribution with the highest Shannon entropy is the "flat" distribution that assigns the same probability to

[2] I will use "informative" and "opinionated" interchangeably.

every hypothesis in **H**. Intuitively the flat distribution is indeed the least informative and least opinionated probability distribution since it does not favor any hypothesis in **H** over any other. On the other hand, the distribution over **H** that has the lowest Shannon entropy and is therefore the most informative is the distribution that assigns all its probability mass to one of the hypotheses. This also seems intuitively reasonable. Indeed, we may view it as a sanity check on any proposed information measure that the measure deem a probability distribution that assigns all its probability to a single hypothesis maximally opinionated, and that it deem the flat probability distribution minimally opinionated.[3]

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

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

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

That information measures are also crucially important to subjective Bayesians is probably a more contentious claim. I defer a more thorough discussion of this issue to Section 4.8, since my discussion will rely on developments made in the chapter. However, the reason why information measures are also important to subjective Bayesians can be put briefly as follows. Subjective Bayesians hold that an agent's probability distribution should accurately represent the agent's epistemic state. Since most of us do not have numerical probabilities in our heads, this introduces a kind translation problem, because agents' subjective degrees of confidence must somehow be translated into numbers. How this translation problem should be solved will sometimes depend on what the goals of the agent are and what the correspondingly suitable information measure is.

4.3 Other Approaches to Measuring Information

Many information measures have been proposed in the statistical and information theory literatures.[4] Which of these many information measures is appropriate for Bayesian purposes? Most Bayesians who have thought about this issue have endorsed the aforementioned Shannon information measure. As was pointed out previously, the Shannon entropy has the intuitively appealing feature of declaring the flat distribution maximally uninformative and the distribution that assigns all its probability to a single hypothesis maximally informative. However, there are many other information measures that also have this feature,[5] so why go for the Shannon entropy rather than

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

[5]Including all Rényi and Tsallis measures.

one of these other measures?

The standard arguments in favor of Shannon's information measure have nothing in particular to do with Bayesian inference,[6] and it is therefore unclear why Bayesians should care about these arguments.

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

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

Some Bayesians have taken a more radical and pluralist approach to information measures. For example, Morris DeGroot (1962) defines "the value of information" as the difference that a piece of evidence makes to the expected utility calculation of an agent. This definition is used by Bernardo (1981) to define "minimally valuable"

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

priors. However, the "minimally valuable" prior is often not the flat distribution and is sometimes even the probability function that assigns all its probability mass to a single hypothesis. Hence, whatever the "minimally valuable" prior is supposed to be, it should not be interpreted as the prior that is maximally uninformative,[7] and DeGroot's measure is therefore not an appropriate measure of the informativeness of probability functions, since the measure clearly fails the previously mentioned sanity checks. The reason DeGroot's measure gives unintuitive results is because the measure depends on the utility function of the agent.

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

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

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

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

when compared to the prior. To quantify our answer to this question, it is natural to make use of confirmation measures. Recall that a confirmation measure, $\mathsf{c}(\mathsf{H}, \mathsf{E})$ is a measure of the extent to which E changes the probability of H.

4.4 Confirmation Measures as Measures of the Informativeness of Data

In the rest of the chapter, two specific confirmation measures will play a particularly important role, for reasons that will become clear. The *difference measure*, $\mathsf{d}(\mathsf{H}, \mathsf{E})$, measures the degree of confirmation that E confers on H as $\mathsf{p}(\mathsf{H}|\mathsf{E}) - \mathsf{p}(\mathsf{H})$. The *log-ratio measure*, $\mathsf{lr}(\mathsf{H}, \mathsf{E})$, measures the degree of confirmation as $\log \frac{\mathsf{p}(\mathsf{H}|\mathsf{E})}{\mathsf{p}(\mathsf{H})}$. Note that both the difference measure and the log-ratio measure have the property that 0 signifies that E confers no confirmation on H.

Importantly for our purposes, confirmation measures may naturally be interpreted as quantitative measures of how much information a piece of evidence provides with respect to a hypothesis.[8] For example, if $\mathsf{c}(\mathsf{H}, \mathsf{E})$ is a large number (either positive or negative), then that means that E provides us with a lot of information about H, since H greatly changes the probability of H; if, on the other hand, $\mathsf{c}(\mathsf{H}, \mathsf{E})$ is 0, then E provides us with no information about H.

It is immediately clear that different confirmation measures will in general disagree on how informative a given datum is, and sometimes the extent of disagreement can

[8]That confirmation measures may be interpreted in this way is not to deny that they may also be interpreted in other ways. For example, one prominent strand of confirmation theory (e.g. Crupi et al. (2007)) regards confirmation as a generalization of logical entailment. I thank Jan Sprenger for emphasizing this to me.

be extreme. For example, a change from $\mathsf{Pr}(\mathsf{H}) = 0.00001$ to $\mathsf{Pr}(\mathsf{H}|\mathsf{E}) = 0.01$ is trivial compared to a change from 0.5 to 0.6 if we use the difference measure; but according to the log-ratio measure, the first change is much greater than the second. How informative E is with respect to H therefore depends on which confirmation measure is used.

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

The Learning Situation and the Log-Ratio Measure

In the first situation I consider – let's call it the "learning situation" – the goal is to identify which hypothesis, H, in a partition of hypotheses, $\mathbf{H}$, is true. Translated into the Bayesian framework, the goal is for the posterior probability of the true hypothesis, H_0, to be as large as possible. Ideally, we want $\mathsf{p}(\mathsf{H}_0|\mathsf{E}) = 1$. Given that this is the goal, what is the best way to measure the extent to which E informs us about some H in $\mathbf{H}$?

One way to make the goal more explicit is by creating a "scoring rule" that more precisely encodes what our epistemic values are in the learning situation. A "scoring rule" is a function of the form $\mathsf{s}(\mathsf{p}, \mathsf{H}_0)$, where H_0 is the "ideal" hypothesis in the partition $\mathbf{H}$—which is typically interpreted to mean that H_0 is the *true* hypothesis,

although this interpretation is not strictly necessary. The score of $\mathfrak{p}$ is supposed to represent how well $\mathfrak{p}$ achieves our goals. The defining feature of the learning situation is that we want to assign as much probability to H_0 as possible. A reasonable way to formalize this goal is to require that a probability function, $\mathfrak{p}$, receive a higher score than a different probability function, $\mathfrak{q}$, if and only if $\mathfrak{p}(\mathsf{H}_0) > \mathfrak{q}(\mathsf{H}_0)$.

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

Out of the well-known and independently plausible scoring rules, the only local scoring rule is the log scoring rule, which assigns a score of $\log \mathfrak{p}(\mathsf{H}_0)$ to $\mathfrak{p}$. In fact, the log-scoring rule is the only local scoring rule that is *strictly proper* (Bernardo, 1979a), which is a property that many philosophers have argued any reasonable scoring rule ought to have (see, e.g. Oddie (1997), Gibbard (2007), Joyce (2009), and Horowitz (2014)). The log-scoring rule is therefore a *reasonable* scoring rule in the learning situation: it appropriately encodes the epistemic goal of learning the truth. Note that this does not mean that the log-scoring rule is the *uniquely* rational scoring rule in the learning situation.

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

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

$$\log \mathsf{p}(\mathsf{H}|\mathsf{E}) - \log \mathsf{p}(\mathsf{H}) = \mathsf{lr}(\mathsf{H}, \mathsf{E}). \tag{1}$$

Thus, we get the conclusion that the log-ratio measure is a *reasonable* measure of the informativeness of evidence in the learning situation, where the goal is to learn whether H is true.[9]

The above argument is not intended to be a knock-down argument for the log-ratio measure of confirmation; the argument is only intended to show that the log-ratio measure is *reasonable* in the learning situation, where the goal is to identify the true hypothesis in a partition of hypotheses. Indeed, although the log-ratio measure is

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

reasonable in the learning situation, it is not reasonable in all other situations; in the next subsection, I consider a different situation where the log-ratio measure is not reasonable, while another confirmation measure is.

The Decision Situation and the Difference Measure

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

For simplicity, I will assume in this chapter that the acts and states are independent according to p.[10] More importantly, I will also assume that the utility function does not depend on p or on possible evidence.[11] The "prior" expected utility of some act A_n is then defined[12] as:

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

[11]Hence, the utility function is not a scoring rule in the sense of the previous section. The learning situation as I presented it in the previous section may also be regarded as a kind of decision problem, but it is important to realize that it is a qualitatively very different decision problem from the kind of decision problem considered in this section, because the utility function (i.e. the scoring rule) in the learning situation depends on the agent's probability function and on the data.

[12]Following Savage (1954).

$$EU(A_n) = \sum_m p(S_m)U(S_m\&A_n) \tag{2}$$

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

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

$$p(S_m) = \sum_j p(S_m|H_j) * p(H_j) \tag{3}$$

Now, suppose we wanted to use a scoring rule to evaluate the prior probability distribution over **H**. Is the log-scoring rule appropriate in this context? By assumption, we do not really care about what the true value of the barometric pressure is; what we care about is how much it will rain in the next hour. The hypotheses about barometric pressure are therefore for us mere *predictive tools*. Clearly, if the goal is

to use the H_j's to predict which S_m is going to obtain, then we want to assign high probabilities to predictively accurate hypotheses (irrespective of whether they are true) and low probabilities to predictively inaccurate hypotheses. Clearly, the "ideal hypothesis" is the hypothesis that is most predictively accurate, not the hypothesis that is true – the true hypothesis only has a special status insofar as it can be expected to have the highest predictive accuracy.[13] But a probability function that assigns a high probability to the truth will not be good for predictive purposes if it also assigns high probabilities to hypotheses that are very predictively inaccurate, and, moreover, it will not be better than a probability function that assigns a low (even 0) probability to the truth, but at the same time only assigns high probabilities to predictively accurate hypotheses. But this means that a local scoring rule, such as the log-scoring rule, is inappropriate, because a local scoring rule scores probability functions only by the probabilities that they assign to the truth.

In particular, in the prediction of S_m (i.e. formula (3)), each H_j is in a sense equally important because each H_j is used in the weighted prediction, so a non-local scoring rule that takes into account the probability assigned to every hypothesis in the partition seems much more appropriate. The most well known non-local scoring rule that does this is the quadratic scoring rule, which assigns a score of $\sum_j (i(H_j) - p(H_j))^2$ to p, where $i(H_j)$ is the indicator function that assigns 1 to H_j if H_j is the ideal hypothesis and 0 otherwise. The quadratic scoring rule therefore seems more appropriate than the log-scoring rule for the purpose of evaluating our prior over $\mathbf{H}$ in the decision situation, where $\mathbf{H}$ is used as a predictive tool. Moreover,

[13] Which it can, if it does not contain adjustable parameters, but not if it does.

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

The above argument is rather sketchy, so here is a more detailed analysis that shows how the difference measure of confirmation naturally arises in the decision situation. First, note that we can plug (3) into (2) in order to make the dependence of the expected utility of A_n on H_j explicit:

$$EU(A_n) = \sum_m \sum_j p(S_m|H_j)p(H_j)U(S_m\&A_n) \tag{4}$$

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

$$EU(A_n|E) = \sum_m \sum_j p(S_m|H_j, E)p(H_j|E)U(S_m\&A_n) \tag{5}$$

Here, $p(S_m|H_j, E)$ represents the probability that it will rain S_m millimeters in the next hour, given that the barometric pressure in the next hour is H_j and the barometric pressure two hours ago was E. It is natural to assume here and in many other similar cases that E does not give us any information about S_m except insofar

as E provides us with information about H_j.That is, it is natural to assume that $p(S_m|H_j, E) = p(S_m|H_j)$.[14] If we make this assumption, then the posterior expected utility of A_n is simply:

$$EU(A_n|E) = \sum_m \sum_j p(S_m|H_j)p(H_j|E)U(S_m\&A_n) \tag{6}$$

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

$$\Delta EU(A_n; E) = EU(A_n|E) - EU(A_n) = \sum_i \sum_j p(S_i|H_j)[\mathbf{p}(\mathbf{H_j}|\mathbf{E}) - \mathbf{p}(\mathbf{H_j})]U(S_i\&A_j) \tag{7}$$

Or, in other words,

$$\Delta EU(A_n; E) = \sum_i \sum_j p(S_i|H_j)\mathbf{d}(\mathbf{H_j}, \mathbf{E})U(S_i\&A_j) \tag{8}$$

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

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

for some A_n. But $\Delta EU(A_n; E)$ depends on the data only through $d(H_j, E)$; hence, in the decision situation, $d(H_j, E)$ arises as a natural measure of the informativeness of E with respect to H_n.

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

One answer to this objection[15] is that we do not really have a choice, because the ratio between two expected utilities will in general not be a meaningful quantity. This is because utility functions are usually only defined up to arbitrary linear transformations. In other words, if U is the utility function of some agent, then $aU + b$ is usually an equally valid representation of the agent's utilities, for any real number b and positive real number a. For instance, Savage's (1954) famous representation theorem, and its various descendants, only define the utility function up to arbitrary positive linear transformations. As a result of this, utilities and expected utilities exist on an interval scale. But this means that the ratio of two utilities is not meaningful, because the ratio will change if you transform the utility scale with an arbitrary positive linear transformation. The difference between ordinal, interval, and ratio scales was discussed in Chapter 2. To refresh the reader's memory of what an interval scale is, consider measurements of temperature. Celsius and Fahrenheit are interval scale measurements of temperature: it is meaningful to say that the difference between 5 and 10 degrees Celsius is the same as the difference

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

between 15 and 20 degrees Celsius, because these differences remain equal if they are both transformed to the Fahrenheit scale. However, it is not meaningful to say that 10 degrees Celsius is "twice as large" as 5 degrees Celsius, because the ratio between these temperatures changes if the temperatures are transformed to the Fahrenheit scale. By contrast, Kelvin is a ratio measure of temperature, which means ratios of temperatures are meaningful.

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

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

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

Suppose, first, that you are in the learning situation, so that your goal is to figure

out which of H or $\neg\mathsf{H}$ is true. According to the odds version of Bayes's formula,

$$\frac{\mathsf{p}(\mathsf{H}|\mathsf{E})}{\mathsf{p}(\neg\mathsf{H}|\mathsf{E})} = \frac{\mathsf{p}(\mathsf{E}|\mathsf{H})}{\mathsf{p}(\mathsf{E}|\neg\mathsf{H})}\frac{\mathsf{p}(\mathsf{H})}{\mathsf{p}(\neg\mathsf{H})} \tag{9}$$

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

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

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

This numerical example, which has nothing to do with scoring rules and therefore

provides an argument independent of the one provided earlier, strongly suggests that the change from 0.0001 to 0.1001 is much more informative regarding H's truth value than is the change from 0.4 to 0.5. This is exactly the verdict delivered by the log-ratio measure.[16]. The difference measure, on the other hand, says that these changes in probability are equally informative, which does not seem reasonable.

The above argument is based on the idea that predictive accuracy is a mark of truth. If H is much more predictively accurate that ¬H, then that strongly supports the conclusion that H rather than ¬H is true. Note that if we were to increase the probability of H by multiplying it by $\mathfrak{m}$, then (9) shows that ¬H's prediction would have to be $\mathfrak{m}$ times better than H's prediction in order to "compensate." This strongly suggests that the multiplicative model is an appropriate model of probability shifting in this context, because multiplying a prior probability by $\mathfrak{m}$ has the same "meaning" or effect regardless of how big the prior probability is. And as the argument in Chapter 3 shows, the multiplicative model of probability shifting entails the log-ratio measure of confirmation, as long as we want the measure to be an interval measure—and the applications later in this chapter will require that the confirmation be at least an interval measure.

But now suppose you are instead in the decision situation, and suppose you are calculating the expected utility of some action A. Then, as the following calculation shows, the change in the expected utility of A is the same whether the probability

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

of H changes from 0.0001 to 0.1001 or from 0.4 to 0.5. For suppose first that the probability of H changes from 0.0001 to 0.1001. Then:

$$\Delta EU(A;E) = \sum_m p(S_m|H)U(S_m\&A)\big[p(H|E) - p(H)\big] + \sum_m p(S_m|\neg H)U(S_m\&A)\big[p(\neg H|E) - p(\neg H)\big] \tag{10}$$

$$= \sum_m p(S_m|H)U(S_m\&A)\big[0.1001 - 0.0001\big] + \sum_m p(S_m|\neg H)U(S_m\&a)\big[0.8999 - 0.999\big] \tag{11}$$

$$= \sum_m p(S_m|H)U(S_m\&A) * 0.1 - \sum_m p(S_m|\neg H)U(S_m\&A) * 0.1 \tag{12}$$

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

$$\Delta EU(A;E) = \sum_m p(S_m|H)U(S_m\&A)\big[0.5 - 0.4\big] + \sum_m p(S_m|\neg H)U(S_m\&a)\big[0.5 - 0.6\big] \tag{13}$$

$$= \sum_m p(S_m|H)U(S_m\&A) * 0.1 - \sum_m p(S_m|\neg H)U(S_m\&A) * 0.1 \tag{14}$$

The fact that (12) and (14) are identical implies that the change in the expected utility of A is the same whether the probability of H changes from 0.0001 to 0.1001 or from 0.4 to 0.5. Thus, in this context, a change in probability from 0.4 to 0.5 is exactly as informative as a change in probability from 0.0001 to 0.1001. And this

is the verdict delivered by the difference measure. In fact, as the above calculation makes clear, the appropriate way to quantify a shift in probability in this case is by way of the additive model, $\nu(x, \epsilon) = x + \epsilon$. And according to the argument in Chapter 3, the additive model entails the difference measure.

On the other hand, the log-ratio measure is *not* a reasonable measure of informativeness in the decision situation. In fact, for every $\epsilon > 0$, no matter how small, and every $B > 0$, no matter how large, it is easy to come up with examples[17] such that the degree of confirmation (or disconfirmation) conferred by E according the log-ratio measure is greater than B, while at the same time, for every n, $|E(A_n; E) - E(A_n)| < \epsilon$ and $|E(A_n; E)/E(A_n) - 1| < \epsilon$, i.e. the difference that E makes to the expected utility of every action under consideration is arbitrarily small, regardless of whether you measure the impact that E has on the expected utility ranking of actions as a difference or as a ratio.

Arguably, in the decision situation, where what you care about is the expected utility ranking of the actions under consideration, it does not make sense to say that E provides you with a lot of information if E has essentially no influence on the expected utility of any action. But that is what you have to say if you measure informational impact with the log-ratio measure. The log-ratio measure is therefore not a reasonable measure of informativeness in the decision situation.

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

4.5 How to Derive Information Measures From Confirmation Measures

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

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

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

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

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

Thus, I contend, the following is a reasonable (though not necessarily uniquely

reasonable) quantification of the information distance between the prior distribution and the posterior distribution, given some piece of evidence E:

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

$$\text{InfDis}(\mathsf{p}(\mathbf{H}|\mathsf{E}), \mathsf{p}(\mathbf{H})) = \sum_{\mathsf{j}} \mathsf{c}(\mathsf{H}_{\mathsf{j}}, \mathsf{E}) * \mathsf{p}(\mathsf{H}_{\mathsf{j}}|\mathsf{E}) \tag{15}$$

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

$$\text{InfDis}(\mathsf{p}(\mathbf{H}|\mathbf{E}), \mathsf{p}(\mathbf{H})) = \sum_{\mathsf{i}} \text{InfDis}(\mathsf{p}(\mathbf{H}|\mathsf{E}), \mathsf{p}(\mathbf{H})) * \mathsf{p}(\mathsf{E}_{\mathsf{i}}) \tag{16}$$

(16) tells us how much information, on average, the partition of evidence $\mathbf{E}$ can be expected to provide us about the partition of hypotheses $\mathbf{H}$. A trick due to Jose Bernardo (1979) is now all we need in order to derive information measures.[18]

[18] I emphasize that my interpretation of Bernardo's trick differs significantly from Bernardo's own. For more faithful presentations of Bernardo's views, see Bernardo (1979b), Berger et al. (2009), or Sprenger (2012).

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

More precisely, a prior is intuitively non-informative to the extent that it is distant from most posteriors that are *heavily influenced* by data. To formalize this idea, imagine that we are going to receive a large amount of evidence $\mathbf{E}^1, \mathbf{E}^2, \ldots\ldots \mathbf{E}^{\mathfrak{n}}$. As the amount of evidence increases ($\mathfrak{n} \rightarrow \infty$), the posterior distribution will gradually become increasingly informed by the evidence, and in the limit of infinite evidence, the posterior distribution will be maximally informed and maximally opinionated; that is, some hypothesis (we do not know which one) will have a probability of 1.[19] A prior distribution is then non-informative in proportion to how informationally distant, on average, it will be from the maximally informative posterior distribution, whatever the maximally posterior distribution turns out to be. Using the definition of InfDis (16), we can formally quantify the preceding ideas, and define the information content of the prior distribution, $\mathfrak{p}(\mathbf{H})$, as follows:

$$\mathrm{Inf}(\mathfrak{p}) = \lim_{\mathfrak{n}\rightarrow\infty} \mathrm{InfDis}(\mathfrak{p}(\mathbf{H}|\mathbf{E}^{\mathfrak{n}}), \mathfrak{p}(\mathbf{H})) \tag{17}$$

It is very important to note that we do not need an actual sequence of evidence in order to make sense of (17). The imagined sequence of evidence, $\mathbf{E}^1, \mathbf{E}^2, \ldots\ldots \mathbf{E}^{\mathfrak{n}}$, merely functions as a way of formalizing the idea that the posterior gets increasingly

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

informed as more evidence comes in. The derivation in the appendix shows that, when the hypothesis space is finite, properties of the sequence of imagined evidence (e.g. the distribution of the evidence) do not make a difference to the information content of $\mathsf{Pr}(\mathbf{H})$.[20]

If we plug (15) and (16) into (17), we get the following alternative expression for Inf(p), which makes the dependence on the choice of confirmation measure explicit:

$$\mathrm{Inf}(\mathsf{p}) = \lim_{n\to\infty} \sum_{i} \sum_{j} \mathsf{c}(\mathsf{H}_j, \mathsf{E}_i^n) * \mathsf{p}(\mathsf{H}_j, \mathsf{E}_i^n) \tag{18}$$

Now we can plug different confirmation measures into (18) and get different information measures. In the case of a finite hypothesis space, it is actually possible to explicitly calculate (18) for several well known confirmation measures, and in particular for the difference measure and the log-ratio measure. More precisely, if we plug in the difference measure and the log-ratio measure, respectively, and perform the relevant calculations, we arrive at the following two alternative information measures (see the appendix for the derivations):

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

[20]When the hypothesis space is continuous, the situation is a bit more subtle—in this case, the information content depends on the *statistical model* in which the hypotheses are situated. This is reasonable because, in the continuous case, the hypotheses are generally indexed by continuous parameters, and it is those parameters that are assigned probabilities. But the meaning of a parameter generally depends on the statistical model of which it is a part. For example, the parameter $\mathbf{B}$ in the linear model $\mathsf{B}x + \mathsf{C}$ picks out the slope of a line; but in the quadratic model $\mathsf{A}x^2 + \mathsf{B}x + \mathsf{C}$, $\mathbf{B}$ does not pick out the slope of a line. Thus, it is not strange that the information content of $\mathsf{Pr}(\mathbf{B})$ should depend on which statistical model $\mathbf{B}$ is embedded in, since $\mathbf{B}$ arguably picks out a *different* partition of hypotheses if it is embedded in the quadratic model rather than the linear model.

information content of $\mathfrak{p}$ *according to the log-ratio measure is defined as,*

$$\mathrm{Inf}_{\mathfrak{lr}}(\mathfrak{p}) = -\sum \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \log \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \tag{19}$$

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

$$\mathrm{Inf}_{\mathfrak{d}}(\mathfrak{p}) = 1 - \sum \mathfrak{p}(\mathsf{H}_\mathfrak{i})^2 \tag{20}$$

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

4.6 Two Goal-Relative Non-Informative Priors

The general definition provided in (18) gives us a way of selecting maximally non-informative priors. More precisely, given some confirmation measure, a probability function that *maximizes* (18) is a natural candidate for a prior that is maximally non-opinionated. Both (19) and (20) are maximized by a single prior – namely the flat prior – so if the hypothesis space is finite, whether you use the log-ratio or the difference measure as the confirmation measure in (18) will not make a difference to the non-informative prior you select. In the next section, I consider what happens when (19) and (20) are maximized relative to constraints; it turns out they can then yield different priors, and so the confirmation measure you use makes a difference when you have background information.

However, if the hypothesis space is continuous, the confirmation measure you use makes a difference even if the maximization is not relative to any constrains. For concreteness, let us consider the problem of estimating the bias θ of a coin given n coin flips. In other words, the problem is to estimate the parameter θ in the binomial distribution. Then we have:[21]

The lr non-informative prior. *Given the problem of estimating the parameter* θ *of a binomial distribution, the maximally non-informative prior density function that*

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

corresponds to the log-ratio measure is

$$\text{NonInf}_{\text{lr}}(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} \tag{21}$$

The above prior is known as "the Jeffreys prior" after its discoverer, Harold Jeffreys (1946). We also have:[22]

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

$$\text{NonInf}_{\text{d}}(\theta) = 1 \tag{22}$$

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

[22] A proof is in the appendix.

4.7 Goal-Relative Priors Given Objective Background Information

As noted earlier, the information measures $-\sum \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \log \mathfrak{p}(\mathsf{H}_\mathfrak{i})$ and $1 - \sum \mathfrak{p}(\mathsf{H}_\mathfrak{i})^2$ are both uniquely maximized by the flat prior. However, if we have background information available, it is reasonable to maximize the two information measures relative to that background information. This is the procedure recommended by the objective Bayesians Jaynes (2003) and Williamson (2010), for example.[23]

If (19) and (20) are maximized relative to background information, they will in general not be maximized by the same priors. As a simple illustration, consider again the example provided in the introduction to this chapter.[24] The example was as follows: suppose you are about to roll a six-sided die and you want a probability distribution $\mathfrak{p}(\mathsf{X})$ over the possible outcomes $\mathsf{X} = 1, 2, 3, 4, 5, 6$. I have rolled the die many times already, and I tell you that – on average – the die has landed on 5. Let's first formalize the information that I give you. The natural way for you to formalize that the die has landed on 5 on average is to demand that the expected value of the die roll according to your prior should be 5. In other words, you should require that $\sum_{\mathfrak{i}=1}^{6} \mathsf{X}_\mathfrak{i} \mathfrak{p}(\mathsf{X}_\mathfrak{i}) = 5$. The additional constraints are, of course, that $\sum_\mathfrak{i} \mathfrak{p}(\mathsf{X}_\mathfrak{i}) = 1$ and that $\mathfrak{p}(\mathsf{X}_\mathfrak{i}) \geqslant 0$ for each $\mathsf{X}_\mathfrak{i}$ since probabilities must be non-negative and add up to 1.

[23] How are we to understand the learning of "background information"? This is a deep question that I do not have the space to discuss here. But, very briefly, the learning of background information cannot be the result of conditionalizing because conditionalizing requires that there already be a prior, but background information is supposed to be a constraint that is used in the construction of the prior and must therefore be "prior to the prior." For a discussion of these issues, see chapters 2 and 3 of Williamson (2010).

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

If you maximize $-\sum \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \log \mathfrak{p}(\mathsf{H}_\mathfrak{i})$ relative to all of the above three constraints, you end up with the probability distribution summarized in the following table:[25]

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

The distribution that maximizes $1 - \sum \mathfrak{p}(\mathsf{H}_\mathfrak{i})^2$, on the other hand, is as follows:[26]

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

Perhaps the most striking difference between the two tables is the fact that the second table has zeros in it whereas the first table does not.[27] This is not incidental to this example: whereas the prior that maximizes $-\sum \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \log \mathfrak{p}(\mathsf{H}_\mathfrak{i})$ will never

[25] I have omitted the very tedious calculation.

[26] I have again omitted the tedious calculation.

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

assign a probability of 0 to any hypothesis unless background information logically excludes the hypothesis, the prior that maximizes $1 - \sum \mathfrak{p}(\mathsf{H}_i)^2$ sometimes does assign 0 to hypotheses. Indeed, it is precisely for this reason that Jaynes rejects $1 - \sum \mathfrak{p}(\mathsf{H}_i)^2$ as a measure of non-informativeness, because he does not think that any hypothesis should ever be assigned a probability of 0 unless the hypothesis is logically excluded (Jaynes, 2003, p. 346).

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

The reader may object that assigning a probability of 0 to a hypothesis implies that you would be willing to accept absurd bets. For example, assigning a probability of 0 to H apparently implies that you would be willing to pay USD 1,000,000,000 for a bet that pays 1 cent if H is false. That does not seem rational. However, this objection implicitly assumes that every probability is a *betting* probability. But this assumption begs the question against the arguments made in this chapter. In fact,

as soon as I offer you a bet over a partition of hypotheses, your goal becomes to identify which hypothesis in the partition is true. In other words, you enter the learning situation with respect to that partition. However, according to the arguments presented here, you should only ever assign 0 to a hypothesis if you are in the decision situation, i.e. if you do not care about which of the hypotheses is true, but rather aim to use the hypotheses as a predictive tool in order to predict something else.

We may call the probabilities you assign to hypotheses in the decision situation "predictive probabilities"; thus, the predictive probability you assign to H_i reflects how much trust you put in H_i's prediction. Crucially, you can have trust in the predictions of a hypothesis, even if you are certain that the hypothesis is false. On the other hand, your betting probability in H_i is a reflection of the bets you would be willing to accept on whether H_i is true. Clearly you would not be willing to accept any bets on a hypothesis you are certain is false; your betting probability in a hypothesis you are certain is false is 0. Hence, predictive probabilities and betting probabilities are very different. In general, you should not use predictive probabilities as your betting probabilities.[28]

4.8 Wider Implications for Bayesianism

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

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

Upshots for Objective Bayesianism

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

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

Upshots for Subjective Bayesianism

Whereas the upshots for objective Bayesians are, I think, relatively clear, the upshots for subjective Bayesians are likely to be more controversial. In contrast to

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

This is perhaps most easily seen in cases where you want to represent probabilistically a lack of opinion. Suppose, for example, that you are trying to determine which probability distribution most faithfully represents your opinions regarding the bias of some coin, and suppose, moreover, that you consider yourself completely uninformed and unopinionated, so that you would like your probability distribution over the possible biases of the coin to reflect your lack of an opinion. According to the calculation in Section 4.6, the probability distribution that is maximally unopinionated and that therefore most accurately reflects your epistemic state is relative to whether you are in the learning situation or the decision situation. If you are in the learning situation, the Jeffreys prior is the most faithful probabilistic representation of your lack of an opinion, but if you are in the decision situation, the flat prior more faithfully represents your epistemic state.

Of course, it is possible that you are in both the learning situation and in the

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

More generally, suppose you consider yourself both epistemically risk-averse and empirically-minded and that you therefore want your epistemic state to be as unopinionated as possible given objective background information, such as, e.g., publicly available frequency data. Naturally, you will want your probability distribution to accurately reflect your epistemic risk-averseness. According to the results in Section 4.7, you will need to take into account your goals when you are deciding how to translate your epistemic state into a probability distribution, because whether a probability distribution counts as unopinionated given background information can only be determined once a goal has been specified. Thus the upshots we saw for objective Bayesians also carry over to at least some agents, namely those agents who see themselves as epistemically risk-averse.

4.9 Conclusion

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

5 The Interpretive Problem in Bayesian Inference

5.1 Goals of the Chapter

The preceding chapter drew a distinction between the "decision situation" and the "learning situation" and showed that the probability that it is rational for an agent to use (or to attribute to the agent) depends on which of these situations the agent is in. However, the learning situation, as defined in the preceding chapter, is a highly idealized conception of the goal of pure inquiry. Scientists and Bayesian statisticians often study hypotheses that they know to be false. This creates an interpretive problem because the Bayesian probability of a hypothesis is typically interpreted as a degree of belief that the hypothesis is true. In this chapter, I present and contrast two solutions to the interpretive problem, both of which involve reinterpreting the Bayesian framework in such a way that pragmatic factors directly determine in part how probability assignments are interpreted and whether a given probability assignment is rational. I argue that there is an important sense in which the two solutions are equivalent, and I suggest that the two reinterpretations can help us do

Bayesian inference better. I also explore various features of the two reinterprations, including their relations to the standard Bayesian interpretation of probability and to the Law of Likelihood. In contrast to the preceding chapters, this chapter will be specifically concerned with Bayesian *statistical* inference rather than with Bayesian inference in general. My reason for focusing on Bayesian statistical inference is that the interpretive problem and its solutions are easier to grasp and discuss in this more specific context.

5.2 The Basics of Bayesian Statistical Inference

The basic objects of study in statistical inference are *statistical models*. Given a set of candidate hypotheses indexed by a *parameter*, $\theta \in \Theta$, and given some particular context in which the possible observations or outcomes are x_1, x_2, etc. – or X, for short – a statistical model is a set of conditional probability (density) distributions,[1] $p(x|\theta)$, that jointly specify the probability of each possible $x \in X$ given each possible $\theta \in \Theta$. Given a statistical model or a set of statistical models, Bayesians do inference by following the three-step procedure I discussed in the introduction to the dissertation, and which I repeat here for the convenience of the reader:

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

[1]From now on, I will for simplicity simply use "probability" although in practice probability densities are more common.

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

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

The preceding three-step description of Bayesian inference is intentionally abstract; it is a completely formal description of how Bayesians apply the probabilistic machinery when they do statistical inference – in what follows, I will refer to the preceding three-step procedure as "standard Bayesian inference." Of course, Bayesians have a standard interpretation of what it is that they are doing – in fact, there are at least two standard interpretations. According to the standard "subjective" interpretation, all the probabilities are *degrees of belief* of some sort. According to the standard "objective" interpretation, the probabilities instead represent *logical degrees of support* of some sort. For the purposes of this chapter, the differences between subjective and objective Bayesians will not matter. Importantly for our purposes, subjective and objective Bayesians agree on a more fundamental point, namely that $p(\theta)$ represents the degree of belief/degree of support that θ is *true*.[2] I will refer to this shared commitment of subjective and objective Bayesians as the "standard interpretation." To simplify the presentation, I will often refer to the standard interpretation as the "degree-of-belief" interpretation; but it is important to understand that everything I say applies equally well to objective varieties of Bayesianism.

The initial major goal of the chapter will be to show that the standard inter-

[2]Or more precisely, the degree of belief/degree of support that the hypothesis indexed by θ is true.

pretation is inconsistent with the way in which the formal Bayesian framework is often applied. The fact that there is this inconsistency has been noted in the past,[3] but the seriousness of the issue seems to be under-appreciated. In particular, many statisticians seem to think that the interpretive problem only arises on the level of model evaluation or model selection. Hence a second major initial goal will be to show that the problem is more pervasive than seems to be widely acknowledged; in particular, the problem also often arises on the level of parameter inference.

5.3 The Interpretive Problem in Bayesian Statistical Inference

As an example of how the interpretive problem arises in practice, suppose you are interested in the functional relationship between two quantities, X and Y. For concreteness, suppose X represents some measurement of a complex system, e.g. the minimal pressure of a tropical storm, and Y represents some quantity of interest, e.g. the maximal windspeed of the storm. The true functional dependence of Y on X is unknown and complex, in part because the functional dependence is mediated by the geometry and rotation of the earth. Nonetheless, it is very common in such cases to restrict attention to classes of simple functional relationships, such as the set of linear hypotheses, which models the relationship between Y and X as follows:

$$Y = \alpha X + \beta + \epsilon \tag{1}$$

[3] E.g. Forster and Sober (1994), Shaffer (2001), Sprenger (2009), and more recently Sprenger (2016).

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

Ideally, the statistical model should be justified on scientific grounds.[4] In the case of the relationship between max windspeed and min pressure, various idealized assumptions (see Knaff and Zehr (2007)) justify the model $Y = \alpha X^n + \beta + \epsilon$,[5] and $n = 1$ is a reasonable choice if the minimal pressure is in an intermediate range. Note, however, that the fact that the model is based on idealized assumptions (i.e. assumptions that are known to be violated in practice) implies that the statistical model in fact is known to be false.

In order to perform the standard three-step inference Bayesian procedure on the preceding model, each value of α must be assigned a prior probability, and according to the standard interpretation each probability assignment represents a degree of belief. But what exactly does it mean for a given value of α to be "true" or "false"? Well, α indexes a set of hypotheses, namely $Y = \alpha X + \beta + \epsilon$, so to say that α_0 is "true" in this case is the same as saying that there *exist* values of β and d such that the hypothesis $Y = \alpha_o X + \beta + \epsilon$ is the true functional relationship between X and Y.

Note, however, that the fact that the statistical model is based on idealized assumptions means that every straight line picked out by α is known to be false –

[4] In the social sciences, the model is often not justified on scientific grounds, but is instead chosen based on convenience and tradition. In those cases, one can be very sure that the resulting model is false.

[5] Strictly speaking, the model is of the following form: $Y = \alpha(1010 - X)^n + \epsilon$. I've changed it slightly for the sake of streamlining the presentation.

none of them describe the true relationship between pressure and windspeed. If you know that all the lines are false and if the probabilities are supposed to represent your degrees of belief, then you should assign the minimal probability – i.e. 0 – to every value of α. But this is not what Bayesian statisticians do.[6]

This practice is what leads to the interpretive problem, which may be phrased in the form of a question: what does it mean to assign a model or hypothesis that is known to be false a non-zero probability? To more precisely diagnose the problem, it helps to state the probability axioms with the standard (subjective[7]) Bayesian interpretation made explicit:

> Suppose $\mathbf{H}$ is a set of hypotheses $\{H_1, H_2, \ldots, H_n\}$. Then
>
> 1S. $p(\mathbf{H}) = 1$. Interpretation: you are certain that one of the hypotheses in $\mathbf{H}$ is true.
>
> 2S. $p(H_i) \geqslant 0$ for all $H_i \in \mathbf{H}$. Interpretation: degrees of belief are non-negative.
>
> 3S. $p(\bigvee H_i) = \sum p(H_i)$, when the H_i are mutually exclusive. Interpretation: degrees of belief in a set of hypotheses are additive when it's impossible for more than one of the hypotheses to be true.

Here we can see that the interpretive problem is really a problem with the standard interpretation of the first probability axiom. That is, for many of the hypothesis sets

[6] And for good reason, since assigning a hypothesis a probability of 0 is tantamount to excluding the hypothesis from any further consideration.

[7] Again, nothing hinges on using a subjective degree-of-belief interpretation here, which I've chosen for the sake of simplifying the presentation; the same problems would arise if we instead used an objective interpretation.

that scientists study, it will not be the case that they are certain that one of the hypotheses is true. Hence, strictly speaking, many hypothesis sets will not satisfy axiom 1S. Axioms 2S and 3S, on the other hand, will generally be satisfied by the kinds of hypothesis sets that Bayesian statisticians study.

5.4 The Pervasiveness of the Interpretive Problem

At this point, we should be a bit more careful. The mere fact that you do Bayesian inference on a statistical model that you know to be false does not necessarily mean that you are faced with the interpretive problem. Suppose, for example, that you wish to estimate the mass $\mathfrak{m}$ of some object. A good way of getting an estimate of $\mathfrak{m}$ is by embedding $\mathfrak{m}$ in a statistical model. Even if the statistical model is false because it is based on known false (auxiliary) assumptions, e.g. the assumption that measurement error is normally distributed, probabilistic statements about $\mathfrak{m}$ can still be interpreted in the standard way. For example, a statement like "the probability that $\mathfrak{m}$ is 2kg is 0.5" can be interpreted as a degree of belief that $\mathfrak{m} = 0.5$, even though the statistical model is based on false auxiliary assumptions. If those auxiliary assumptions are seriously wrong, the probability may well be misleading;[8] however, the probability can still sensibly be interpreted as a degree of belief.

The preceding example shows that the fact that a model is known to be false does not necessarily mean that you face the interpretive problem when you do parameter

[8] "Misleading" in the sense of p.12n9 in the Introduction.

inference inside the model. The fact that the model is known to be false does, however, mean that you face the interpretive problem if you try to use Bayesian inference to compare and evaluate different statistical models.

Partly for the preceding reasons, George Box (1980) famously recommended a reconciliation between Bayesian and frequentist inference. According to Box, frequentist methods should be used to identify a "useful" (albeit false) statistical model; Bayesian inference can then be used to infer plausible parameter values inside the assumed statistical model. Bernardo and Smith (1994), Key et al. (1999), and Gelman and Shalizi (2013) recommend other similar two-step procedures, in which a model is first picked using a non-Bayesian method (e.g. cross-validation), and then standard Bayesian parameter inference is performed inside the selected model. All these authors seem to have the view that the interpretive problem prevents us from sensibly using the standard three-step Bayesian procedure to compare and evaluate statistical models, but that once you have chosen a (false) model, you can use standard Bayesian inference to do parameter inference inside the assumed model.

This view both underestimates and concedes too much to the interpretive problem. It concedes too much because giving up on standard Bayesian inference when you do model evaluation seems like a steep cost. More seriously, it underestimates the interpretive problem because the interpretive problem very often arises even on the level of parameter inference inside a fixed statistical model. In particular, it will typically be the case that the false auxiliary assumptions of a statistical model must be assigned probabilities in order for a Bayesian analysis to be possible. For example, if you use the normal distribution to measure measurement error, then you will

in general need to introduce a so-called "nuisance parameter" $\mathtt{d}$ representing the standard deviation of the distribution. In order to do Bayesian inference, each possible value of $\mathtt{d}$ must be assigned a prior probability. But if the normal distribution is known to be a false idealization, then every value of $\mathtt{d}$ will also be known to be false. And thus the interpretive problem usually arises whenever probabilities need to be assigned to nuisance parameters. Moreover, the fact that the statistical model is false often implies that the parameters inside the model also pick out hypotheses that are false. This was the case, for example, in the preceding example concerning the relationship between windspeed and pressure, where every parameter picked out a straight line that was known to be false.

The windspeed/pressure example is a problem in regression analysis, which is one of the major areas of statistical inference. In regression analysis, the goal is to estimate the functional relationship between two or more quantities. Almost invariably, the hypotheses under consideration will be restricted to very simple functional relationships, such as the set of lines, parabolas, exponentials, etc. However, most functional relationships in the world cannot realistically be expected to belong to one of these sets of simple functional relationships. Hence, scientists will generally face the interpretive problem when they use Bayesian inference in regression problems.

Another major area of statistical inference where scientists face the interpretive problem is in Bayesian phylogenetics. Phylogeneticists in both biology and linguistics use trees to represent family relationships between species or between languages. In both cases, the trees investigated omit known relationships and introduce false idealizations. For example, a tree phylogeny for a language family is premised on

the (false) idea that languages bifurcate instantaneously and are forever separated thereafter. Yet, even though all phylogenetic trees are clearly false, linguistic phylogeneticists often use Bayesian inference and are often interested in discovering which tree has the highest posterior probability. These probabilities cannot comfortably be interpreted as probabilities that the trees are literally true, and thus scientists who use Bayesian inference on problems such as this are faced with the interpretive problem.

5.5 Solutions to the Interpretive Problem

Broadly speaking, there are three possible responses to the interpretive problem:[9]

1. You can give up on using standard Bayesian inference whenever you are faced with the interpretive problem.

2. You can offer a reinterpretation of the Bayesian framework that solves the interpretive problem.

3. You can claim that you do not need an interpretation of what you are doing in order to make reliable and useful inferences.

[9] A fourth response to the interpretive problem that initially seems attractive is to try to change the algebra over which the probability function p ranges. For example, some might be tempted to consider the algebra generated by the associated propositions, $<\mathsf{H_i}$ is the best hypothesis$>$, for each $\mathsf{H_i}$, or something similar. However, there is a fundamental reason why any such proposal will not work. Briefly, the reason is that if you want to do Bayesian inference on a statistical model that is parameterized by θ, then you need to assign probabilities to θ; you cannot instead assign probabilities to, e.g., propositions of the sort $<\theta = 2$ is the best parameter value$>$ or $<\theta = 2$ is the parameter value that is most predictively accurate$>$, because these propositions are not part of the statistical model and do not entail probabilities for possible observations. If you want to do statistical inference on the statistical model parameterized by Θ, you need to assign probabilities over Θ.

As we have seen, several statisticians seem to recommend option 1 in the case of Bayesian model evaluation and comparison. However, these statisticians apparently fail to see that the interpretive problem also often arises on the level of parameter inference. If you choose option 1, then the Bayesian machinery will have very limited applicability, so option 1 is not an appealing option for Bayesians.

To some extent, option 3 has to have something going for it, given that Bayesian statisticians routinely apply the Bayesian machinery over known false hypotheses spaces, and they often do so successfully.[10] Clearly, it is possible to do something well without understanding exactly what it is that you are doing. Option 3 is not satisfying for those who are attracted to Bayesianism on theoretical grounds, however. Moreover, intuitively, you should be able to make better inferences if you know what you are doing.

Hence, from both a theoretical and practical standpoint, option 2 is the best response to the interpretive problem. If you have an interpretation of what you are doing that is consistent with what you are doing, then you will be in a position to reach your goals more efficiently, or so I will try to argue.[11]

[10] "Successfully" in the sense that they get answers that "make sense" and predictions that are reasonably accurate.

[11] Morey et al. (2013) offer a solution to the interpretive problem that falls into either option 1 or option 2. According to Morey et al., "...scientific models, including statistical models, are neither true nor false" (p. 71) and that "Box's (1979) famous dictum... ...could be shortened to 'some models are useful' without any loss" (p. 71). They then recommend assigning odds rather than probabilities to models because a "Bayesian who employs odds is silent on whether or not she is in possession of the true model, and, in fact, need not acknowledge the existence of a true model at all" (p. 71). It is, however, unclear how using odds rather than probabilities is supposed to solve the interpretive problem. And it is not clear how refusing to assign truth values to models solves the problem either. What does it mean to say that your odds are 5 to 1 in a model that is neither true nor false as against another model that is also neither true nor false? The interpretive problem seems to be just as severe here as before.

For the remainder of the chapter, I will consider two option 2 solutions to the interpretive problem. One solution involves interpreting probabilities as counterfactual degrees of belief rather than standard degrees of belief, while the other interpretation involves interpreting probabilities as a degree of belief in verisimilitude. As we will see, both interpretations have the consequence that a probability distribution no longer is purely epistemic.

5.6 The Verisimilitude Interpretation of Probability

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

Below, I will argue that this idea can be accommodated by Bayesians in a fruitful way. For simplicity, I will from now on focus on parameter inference. That is, I will assume that the hypotheses are indexed by a parameter Θ inside some fixed statistical model, and that each $\theta \in \Theta$ picks out some hypothesis that does not itself contain adjustable parameters.

The idea that the goal of inference is to identify the θ that is closest to the truth leads to a natural reinterpretation of probability. Instead of interpreting $\mathfrak{p}(\theta)$ as a

degree of belief in θ, we interpret $p(\theta)$ as a degree of belief that θ is closest to the truth out of the hypotheses in Θ.[12] I will call this interpretation of probability the "verisimilitude interpretation."

To give the verisimilitude interpretation content, we need to say a bit about what is meant by a measure of "closeness to the truth." The study of verisimilitude was initiated by Popper (1963) and has by now accumulated a large literature.[13] The most influential contemporary approach in the study of verisimilitude – known in the literature as the "similarity approach" – understands verisimilitude as a particular kind of approximation. To say that something is a good approximation of something else is to say that the two things are similar in some relevant respect. Thus, to say that a hypothesis is close to the truth is to say that the hypothesis is similar to the true hypothesis.

This idea can be formalized if we suppose that there is a (context-appropriate[14]) verisimilitude measure, ν, that ranks hypotheses by how similar they are to the true hypotheses. If we presume that such functions are available, we can say that θ_1 is closer to the truth than θ_2 if and only if $\nu(\theta_1) > \nu(\theta_2)$. Below we will see that there are certain requirements that the verisimilitude measure ought to obey, but for now no such requirements will be made.

As a concrete example, a probabilistic divergence measure that has been suggested as a verisimilitude measure in a statistical context is the Kullback-Leibler divergence

[12] Again, we can be neutral on whether the "degree of belief" should be understood in a subjective or objective sense.

[13] See Niiniluoto (1998) for a survey.

[14] In general I agree with Northcott (2013) that there is little reason to assume a priori that there will be a single distance measure that appropriately measures approximate truth in all contexts.

(Forster and Sober, 1994). Supposing that q is the "true" probability distribution that governs the distribution of the data, then the verisimilitude (according to the K-L divergence) of some hypothesis θ (that does not contain adjustable parameters) is $KL(\theta) = -\int q(x) \log \frac{q(x)}{p(x|\theta)} dx$. Later in the chapter I will suggest a simple verisimilitude measure that makes sense in the earlier example concerning the relationship between windspeed and pressure.

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

> 1V. $p_\nu(\mathbf{H}) = 1$. Interpretation: You are certain that one of the hypotheses in $\mathbf{H}$ is a maximum of ν.

It is clear that the verisimilitude interpretation avoids the interpretive problem because, unless the hypothesis set or the verisimilitude measure are very pathological, at least one of the hypotheses under consideration will be maximally close to the truth, and hence 1V will be satisfied.

The verisimilitude interpretation has a couple of important features that distinguishes it from the standard degree-of-belief interpretation. Most importantly, on the verisimilitude interpretation, the probability assigned to a hypothesis is relative to a given way of measuring verisimilitude. The verisimilitude probability of a hypothesis is therefore not an absolute number; it is context-dependent and contrastive. This is in sharp contrast to the standard degree-of-belief interpretation, according to which

the probability assigned to a hypothesis represents your confidence in the hypothesis. Another feature of the verisimilitude interpretation is that the prior probability assigned to a hypothesis is not purely epistemic. According to the standard Bayesian interpretation of probability, the prior probability assigned to a hypothesis is supposed to reflect background information or background beliefs about whether the hypothesis is true; in other words, the probability is purely epistemic. However, a verisimilitude probability will be a reflection of not just background information, but also of the way in which verisimilitude is measured. Arguably, the way in which you ought to measure verisimilitude will depend on what goals you have. If that's correct, then the probability assignment that it is rational for you to use becomes relative to your goals. This feature of the verisimilitude interpretation will become clearer in the next section of the chapter.

5.7 The Verisimilitude Interpretation of Probability is Useful

Earlier I claimed that having an "option 2" solution to the interpretive problem would be useful because if you have an interpretation of what you are doing, then you will be able to reach your goals more efficiently. In this section I intend to make it clearer how the verisimilitude interpretation can be useful in this way.

As a preliminary step, it is helpful to step back for a moment and ask a fundamental question: what is it that makes Bayesian methods useful in the easy case where we do not face the interpretive problem and where the standard interpretation

therefore makes sense? If the benefits of Bayesian methods remain even when he standard interpretation of the probability axioms is replaced with the verisimilitude interpretation, then the verisimilitude interpretation is clearly useful.

Why Be a Bayesian?

Perhaps the greatest selling point of Bayesian statistical inference is that the prior distribution gives researchers a principled way of incorporating background information. For example, suppose you are estimating the mass of a small cup of water, and suppose you model the outcome of your measurement as a likelihood function $\mathfrak{p}(\mathfrak{x}|\mathfrak{m})$, where $\mathfrak{x}$ is the outcome of your measurement and $\mathfrak{m}$ is a possible value of the cup's mass. A standard classical ("frequentist") method of estimating the mass of the cup is to choose as your estimate the value of $\mathfrak{m}$ that maximizes the probability of the observed measurement. This estimation method is known as "maximum likelihood" estimation.

From a Bayesian point of view, maximum likelihood estimation is essentially equivalent to Bayesian inference with a flat (improper) prior probability function that assigns a non-zero and equal probability density to every possible value of $\mathfrak{m}$ from $-\infty$ to $+\infty$, because the maximum likelihood estimate will be equal to the estimate that has the highest posterior probability if and only if the prior is flat. Clearly, the prior implicitly used in maximum likelihood estimation neglects to incorporate common sense background information that we have about $\mathfrak{m}$, and is therefore – from a Bayesian and intuitive point of view – deficient. For example, the mass of an object cannot be a negative number, so no prior should assign any probability mass

to negative values of $\mathfrak{m}$. Furthermore, we can be absolutely certain that a small cup of water is not going to weigh more than, say, 1kg, so we can also assign a probability of 0 to all values of $\mathfrak{m}$ greater than 1kg. Thus, as a minimal requirement, any prior probability distribution we use should be restricted to the interval $[0, 1]$. Of course, we have additional common sense knowledge that allows us to restrict the class of sensible prior distributions further.

The above example shows how even very obvious background information can be incorporated in a Bayesian prior in order to improve the inference. Indeed, at least to Bayesian statisticians and scientists who make use of Bayesian methods, this is probably the single biggest advantage that Bayesianism has over its competitors. But how are you supposed to take into account your background information when you are trying to come up with a prior probability distribution over a class of false hypotheses? Do the advantages of Bayesianism carry over when the goal of inquiry changes from finding the truth to finding the hypothesis that is closest to the truth? In the next subsection, I will suggest that the answer is "yes." Scientists often have background knowledge that they can use to discriminate between false hypotheses in a principled way. And a good way of incorporating this background knowledge is through the construction of a Bayesian prior.

Verisimilitude and Background Knowledge

Consider again the example concerning the relationship between barometric pressure (X) and maximum windspeed (Y). Let's use f to denote the true (unknown) functional dependency of Y on X. Suppose the range of f you are interested has domain $\mathsf{X} \in (\mathfrak{p}, \mathsf{P})$.

Now, suppose one of the things you know about the relationship between barometric pressure and windspeed in this range is that changes in maximum windspeed are relatively insensitive to changes in barometric pressure. This background knowledge can be formalized as knowledge about the derivative of f with respect to X, for example that this derivative is bounded by some interval (a, b). Suppose, moreover, that you also know that $f(p) = K$.

So far, this is background knowledge about the actual, unknown function relating barometric pressure and windspeed. What consequences does this background knowledge about f have for inferences about the hypothesis set actually under consideration? Suppose, as before, that the hypothesis set you are considering is the set of linear functions and that you know that f is not in this set. That is, you model the relationship between windspeed and barometric pressure by the set of linear functions $L(Y) = \alpha X + \beta + \epsilon$, where ϵ is a normally distributed error term with 0 mean. Can you use your background knowledge about f to discriminate between the various false linear hypotheses in a principled way? The answer is yes, but how your background knowledge affects the inferences you are entitled to make will depend on how you measure verisimilitude.

Suppose that your ultimate goal is to build a structure that will be able to withstand strong winds.[15] In that case, it is very important that the maximal error you make when you estimate windspeed is as small as possible. In other words, Figure 5.1 is a natural measure of closeness to the truth given your goal; this is not to say that this is an appropriate way to measure closeness to the truth given other goals.

[15] I thank Michael Titelbaum for suggesting this example to me.

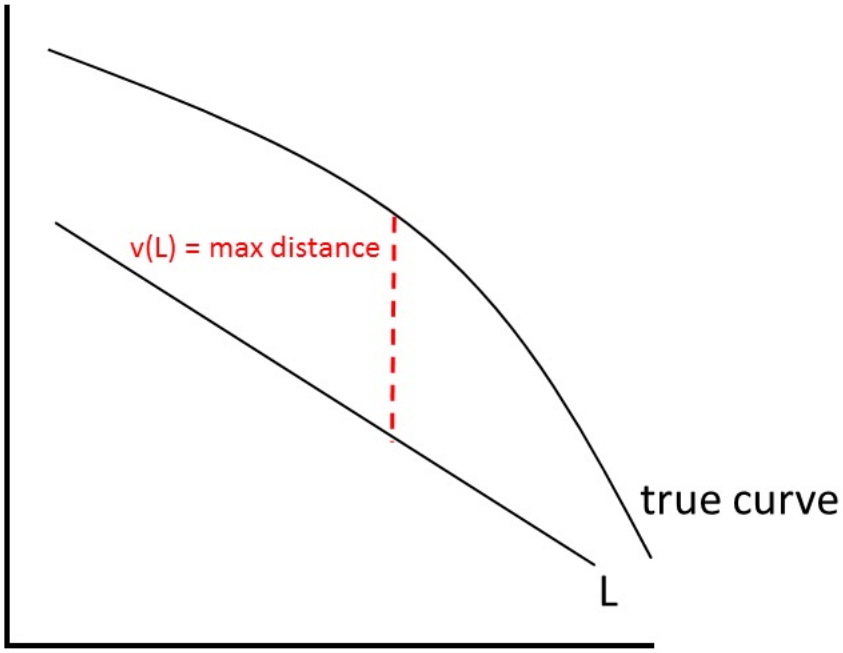

Figure 5.1: A measure of closeness to the truth

Mathematically, the verisimilitude of some straight line L is given by the formula $\nu(L) = -Max|f - L|$,[16] where the maximum is taken over the whole range of possible pressures, (p, P). Given that you use ν to measure verisimilitude, and given that you have restricted the analysis to the class of linear functions, your more immediate goal is to identify linear functions that are close to the truth according to ν. Suppose L is a linear function whose parameters α and β satisfy the following requirements: $\alpha > b$ and $\beta > K$. Then it is easy to show that your background knowledge about f together with your way of measuring verisimilitude entail that there exists some linear function L^1 such that $\alpha \in (a, b)$ and $\beta = K$ and such that L^1 is closer to the

[16]The minus sign is there to make sure that higher verisimilitude scores correspond to a smaller maximum difference between the line and true function.

truth than L (proof in the appendix). Thus, you can rule out in advance that L will be closest to the truth in the sense that you care about. But this straightforwardly leads to a rational requirement on the joint prior distribution $p(\alpha, \beta)$ over α and β, namely that $p(\alpha, \beta)$ should be 0 whenever $\alpha > b$ and $\beta > K$.

A prior distribution p that meets this requirement will be *more rational* than – say – a flat prior in the sense that a Bayesian inference that uses p as the prior can be expected to more quickly converge on the hypothesis under consideration that is closest to the truth in the relevant sense. This can be expected because p excludes from consideration hypotheses that cannot possibly be closest to the truth, whereas the flat prior excludes no hypotheses from consideration. The role played by p is exactly analogous to the role played by a prior over the possible values of the mass of a cup of water that assigns, say, 0 to all negative values.

However, crucially, if you measure closeness to the truth differently, you do not necessarily get the same rational requirement on the prior distribution. Suppose, for example, that you are instead very concerned with the minimal rather than maximal distance of each line from the truth. That is, you use $w(L) = \mathit{Min}|f - L|$ to measure the verisimilitude of each line. According to w, any line that intersects f will be maximally close to the truth, and so your goal is now to identify the lines that intersect f in the range (p, P). Clearly, lines that have a very steep slope in this range will stand a better chance of intersecting f than lines that do not, and thus if you use w to measure verisimilitude, then it is rational for you to use a prior distribution that assigns more probability to lines that have extreme values of α.[17]

[17]Some may feel that w is an irrational way of measuring verisimilitude; later we will see that, indeed, there are at least two senses in which w is arguably deficient as a measure of verisimilitude.

In general, how background knowledge interacts with a given measure of verisimilitude measure in order to give rational requirements on the prior distribution is a subtle and complex question. My goal in this section is not, however, to demonstrate in full generality how to best translate background information into reasonable requirements on prior distributions over sets of known false hypotheses. My goal is rather to show that it is possible to do so, and that it is plausibly useful. I defer a more thorough treatment of these issues to another time.

5.8 The Counterfactual Interpretation of Probability

The preceding section shows that the verisimilitude interpretation of the probability axioms is a potentially useful solution to the interpretive problem. The verisimilitude interpretation has the feature that the prior probability distribution incorporates not just your background information, but also what you hope to accomplish, formalized by way of a verisimilitude measure. In a very recent paper, Jan Sprenger (2016) proposes a new and different "option 2" solution to the interpretive problem. Sprenger's solution also involves reinterpreting the probability axioms, but he offers a reinterpretation that appears to be quite different from the verisimilitude interpretation. However, as we will soon see, given certain plausible assumptions, the verisimilitude solution and Sprenger's solution share many features in common and are even formally inter-translatable.

Sprenger's suggestion is that the probability of a false hypothesis can sensibly

be interpreted as a *counterfactual* degree of belief. More precisely, suppose α is a parameter that indexes a set of hypotheses, all of which are known to be false. Then any probability assigned to some particular α_0 should be construed as a degree of belief in α_0 that is *conditional* on the (false) supposition that one of the hypotheses indexed by α is true. In other words, the probability of α_0 is really the *conditional* probability $\mathrm{p}(\alpha_0 | \vee \alpha)$, where the condition $\vee\alpha$ is the false disjunction that says that one of the α's is true.

This idea is less abstract than it may seem at first blush. As an illustration, suppose I have a coin in a locked cabinet. The probability that the coin would land heads given that I *were* to toss the coin is 0.5, even if it is false that I ever toss the coin. Similarly, according to Sprenger, we can evaluate the probability that a hypothesis is true given that the false supposition that the world *were* such that one of the hypotheses under consideration is true.

According to Sprenger, the counterfactual interpretation of probability offers a simple solution to the interpretive problem that avoids the "muddy waters of verisimilitude." However, in order to actually evaluate counterfactual probabilities in a principled manner, it seems we have to enter waters that are at least as muddy as the verisimilitude waters. Consider again the example concerning the set of linear hypotheses relating X (barometric pressure) to Y (max windspeed). We have already agreed that your actual degrees of belief in all of these linear hypotheses is 0. Your degree of belief in some particular linear hypothesis conditional on the disjunction of all the linear hypotheses may still be different from 0, but how are you supposed to figure out what it is? You somehow have to figure out what your probabilities should

be on the assumption that the world *were* such that barometric pressure and max windspeed were perfectly linearly related. It is not easy to see how you are supposed to do this.

Hence, in order for the counterfactual interpretation of probability to have content, something arguably needs to be said about how we are to understand and evaluate counterfactual probabilities, in the same way that something needs to be said about verisimilitude in order for the verisimilitude interpretation to have content. Sprenger does not give us any guidance on how to evaluate or understand counterfactual probabilities. However, a natural thought is that counterfactual probabilities should be evaluated in a way that is analogous to the way counterfactual conditionals are evaluated. According to (a simplified version of) Lewis's (1973) analysis of counterfactuals, which is the standard analysis, in order to evaluate a counterfactual such as "If A were the case, then B would be the case," you have to consider the closest possible world in which A is true, and then see whether B is true in that world.[18] Crucially, Lewis's analysis depends on a ranking of possible worlds, where worlds are ranked by how similar they are to the actual world.

Perhaps counterfactual probabilities should be assessed in a similar manner. It is not hard to imagine very strange and fanciful possible worlds in which pressure and windspeed are linearly related, but presumably most of those possible worlds are not interesting or relevant. As is the case in the counterfactual analysis of conditionals, it is presumably the closest possible worlds that are the interesting ones. But which

[18]To simplify the presentation, my discussion in this and the next section will assume the so-called "Uniqueness Assumption" according to which, for every A, there is a unique closest possible world in which A is true. This is a strong and implausible assumption. However, nothing in the discussion will hinge on whether this assumption is true.

possible worlds are those? To answer this question, you need to be able to rank worlds in terms of their closeness or similarity to the actual world.

If we do have available a similarity ranking over possible worlds, $\mathfrak{s}$, then there are two different ways in which we can make more precise what a counterfactual degree of belief in α_0 is.[19] The first proposal is that $\mathfrak{p}(\alpha_0|\vee\alpha)$ should be understood as the actual degree of belief you would have in α_0 if you were in the closest possible world in which α is true. According to the second proposal, $\mathfrak{p}(\alpha_0|\vee\alpha)$ should be understood as your degree of belief that α_0 is true in the closest possible world in which α is true. Of these two proposals, the latter is arguably more appealing than the former. But we do not have to adjudicate between the two proposals here, because the important thing for us is that both proposals rely on the existence of a similarity ranking over possible worlds. This fact alone means that the counterfactual interpretation will have the same broad features as the verisimilitude interpretation, regardless of which more specific proposal is adopted.

In particular, on the counterfactual interpretation understood in the Lewisian way, every probability assignment becomes relative to the way you measure similarity between worlds. Thus, the counterfactual interpretation of probability makes probability assignments contrastive in the same way the verisimilitude interpretation makes probability assignments contrastive. Moreover, there are many ways of measuring similarity between worlds, but the way in which you ought to measure similarity between worlds is presumably relative to the features of the world that you care about. If that's correct, then the counterfactual interpretation, like the verisimilitude

[19] I thank Andrea Guardo for pressing me on this point.

interpretation, makes the rationality of agents' probability distributions goal-relative.

Thus, we see that the counterfactual interpretation shares with the verisimilitude interpretation the most interesting and important features of the latter. In the next subsection, we will see that there is also a formal relationship between the two interpretations.

Relationship Between the Verisimilitude and Counterfactual Interpretations

In general, any similarity ranking of possible worlds straightforwardly induces a natural verisimilitude ranking of hypotheses, and vice versa. More precisely, suppose we are given a similarity ranking on worlds $w_\alpha \geqslant w_1 \geqslant w_2 \geqslant \ldots$, where w_α is the actual world. Then we can define a verisimilitude ranking on hypotheses as follows: suppose w is the closest world in which H is true and w' is the closest world in which H' is true, then $\nu(H) \geqslant \nu(H')$ if and only if $w \geqslant w'$.[20]

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

[20]Hilpinen (1976) uses a similar approach to define a specific verisimilitude measure.

Thus, although they may appear different, the verisimilitude interpretation and the counterfactual interpretation of probability are, in a sense, formally inter-translatable. In the next section, I discuss the relationship of each interpretation to the standard Bayesian interpretation of probability.

5.9 Relationship Between the Verisimilitude, Counterfactual, and Standard Interpretations

According to the standard interpretation, $\mathsf{p}(\mathsf{H})$ is your degree of belief in H. This interpretation fails when all the hypotheses under consideration are known to be false. I have suggested that in those cases we need to move to either the verisimilitude interpretation or the counterfactual interpretation. Ideally, the verisimilitude and counterfactual interpretations should both be generalizations of the standard interpretation, so that both reduce to the standard interpretation in cases where the standard interpretation is applicable; i.e. in cases where we are certain that one of the hypotheses under consideration is true, so that 1S is satisfied. Is that the case?[21]

The answer is that it depends on characteristics of the verisimilitude and counterfactual ranking measures. Let's first consider the verisimilitude interpretation. Suppose one of the hypotheses under consideration is true, and let $\mathsf{p}(\mathsf{H})$ be your degree of belief that H is true. Let's call the true – but unknown – hypothesis t. Suppose ν is such that it has a *unique* maximum, and that the unique maximum is t. Then, according to the verisimilitude interpretation, $\mathsf{p}_\nu(\mathsf{H})$ is your degree of belief

[21] I thank Andrew Forcehimes for pressing me on this issue.

that H is a maximum of ν, which, under the conditions specified, means that $p_\nu(H)$ is your degree of belief that H is t; in other words, $p_\nu(H)$ is your degree of belief that H is true, so we have $p_\nu(H) = p(H)$. Hence, the verisimilitude interpretation reduces to the standard interpretation under the specified conditions. However, if ν has several maxima or if the truth is not among the maxima of ν, then clearly $p_\nu(H)$ will not necessarily equal $p(H)$. Hence, the verisimilitude interpretation reduces to the standard interpretation just in case the following conditions are all met: (1) one of the hypotheses under consideration is true, (2) ν has a unique maximum, (3) the truth is a maximum of ν.

Now let's consider the counterfactual interpretation of probability. Suppose the similarity ranking over possible worlds satisfies the following conditions: (1) there is a unique world that is closest to the actual world, (2) the actual world is closest to itself. Then, by essentially the same reasoning as above, it follows that we will have $p_c(H) = p(H)$. Hence, the counterfactual interpretation reduces to the standard interpretation just in case one of the hypotheses under consideration is true and the similarity ranking over possible worlds satisfies the constraint known in the counterfactuals literature as *strong centering*. If the similarity ranking does not satisfy strong centering, then the counterfactual interpretation and the standard interpretation will not be equivalent in the cases where both are applicable.

As I said above, it is natural to regard it as a desideratum that the verisimilitude and counterfactual interpretations reduce to the standard interpretation whenever the standard interpretation is applicable. If we do, then it follows that the preceding discussion gives us rational constraints on how to measure closeness to the truth or

similarity between worlds. In fact, these constraints are demanding enough that the second verisimilitude measure suggested in Section 5.8 – i.e. $w(L) = Min|f - L|$ – clearly does not satisfy the constraints, and even the first suggested verisimilitude measure – i.e. $v(L) = Max|f - L|$ – will not always satisfy the constraints.

5.10 The Law of Likelihood

Insofar as we want the verisimilitude and counterfactual interpretations to be generalizations of the standard interpretation, the ranking measures used by each need to satisfy the conditions described in the preceding section. The preceding section can therefore be understood as formulating reasonable constraints on the way in which similarity between hypotheses or worlds should be measured. In this section, I will mention another such constraint that arises from the fact that Bayesian inference is ultimately likelihood-driven. That is, given enough evidence, the likelihoods of the hypotheses under consideration will determine which hypothesis ends up having the highest posterior probability. The odds formulation of Bayes's formula makes this clear. Given E, H_1, and H_2, we have:

$$\frac{p(H_1|E)}{p(H_2|E)} = \frac{p(E|H_1)}{p(E|H_2)} * \frac{p(H_1)}{p(H_2)} \tag{2}$$

The odds formulation shows that Bayesian inference conforms to the *Law of Likelihood*. According to the Law of Likelihood, evidence x supports H_1 over H_2 if and only if $p(x|H_1) > p(x|H_2)$. The standard interpretation of the Law of Likelihood is as follows:

> **Law of Likelihood, standard interpretation:** evidence x supports *the proposition that* H_1 *is true* over *the proposition that* H_2 *is true* if and only if $p(x|H_1) > p(x|H_2)$.

On the standard interpretation, Bayesian inference automatically conforms to the Law of Likelihood. But if we instead move to the verisimilitude interpretation, then that is no longer the case. Here is the natural way to interpret the Law of Likelihood on the verisimilitude interpretation:

> **Law of Likelihood, verisimilitude interpretation:** evidence x supports *the proposition that* H_1 *is closest to the truth* over *the proposition that* H_2 *is closest to the truth* if and only if $p(x|H_1) > p(x|H_2)$.

The verisimilitude interpretation of the Law of Likelihood clearly imposes an extra constraint on how we ought to measure verisimilitude, because not every way of measuring verisimilitude will satisfy the verisimilitude reading of the Law of Likelihood. For example, the second measure of versimilitude that was suggested in Section 5.8 – namely, $w(l) = Min|f - l|$ – will not satisfy this reading of the Law of Likelihood because many of the lines that have a maximal verisimilitude will clearly have a very low likelihood. On the other had, the first measure of verisimilitude suggested in Section 5.8 does satisfy the constraint imposed by the Law of Likelihood.

5.11 What About In-Between Cases?

Before concluding the chapter, there is a final issue that the reader may have noticed and that deserves to be discussed. The standard interpretation of the first probability axiom states that you are certain that one of the hypotheses under consideration is true. Let's call cases in which this is satisfied the "classic case." The interpretive problem, as I have described it, occurs when you are certain that all the hypotheses under consideration are false. Let's call this the "problem case." The classic case and the problem case represent two extremes on a continuum. What should we say about cases that are in-between the extremes?

There is an analogy with conditionals that might shed light on these in-between cases. Suppose you are evaluating a standard conditional of the form $\mathsf{A} \rightarrow \mathsf{B}$. What I've called the "classic case" is analogous to a case where you are sure that the conditional you are evaluating is indicative; the problem case, on the other hand, is analogous to a case where you are sure that the conditional is counterfactual. In-between cases are analogous to cases where you are not sure whether the conditional should be evaluated indicatively or counterfactually.

To illustrate, consider the following conditional:

Someone killed Kennedy $\rightarrow$ Oswald killed Kennedy

This conditional has an indicative and a counterfactual reading. The indicative reading is as follows:

If someone killed Kennedy, then Oswald did it.

The counterfactual reading is as follows:

If someone were to have killed Kennedy, then it would have been Oswald.

Suppose we know that the antecedent is true. Then the natural reading of "Someone killed Kennedy → Oswald Killed Kennedy" is indicative, and the way we would evaluate whether this conditional is plausible would be by looking at where Oswald was at the time of Kennedy's death, whether Oswald had a motive for killing Kennedy, etc.

But suppose we know that the antecedent is false. Then we have to read "Someone killed Kennedy → Oswald Killed Kennedy" counterfactually, and the counterfactual conditional may plausibly be false even if its indicative counterpart is plausibly true, and vice versa. In evaluating the counterfactual conditional, some of the same considerations will still be relevant – e.g. whether Oswald has a motive for killing Kennedy – but other considerations will be completely irrelevant, such as where Oswald was at the time of Kennedy's death.

Suppose we do not know whether the antecedent of the conditional is true or false. Then we do not know whether to read it indicatively or counterfactually. In that case, we must presumably evaluate its overall plausibility by evaluating its plausibility given that it's read counterfactually and its plausibility given that it's read indicatively.

Similarly, it presumably sometimes happens that you are unsure whether you are in the classic case or the problem case; that is you are not sure whether one of the hypotheses under consideration is true. In this case, we can represent your situation as a mixture over the classic case, C, and the problem case, P. Thus, your probability in

some given hypothesis H can be represented as $\mathsf{p}(\mathsf{H}) = \mathsf{p}(\mathsf{H}|\mathsf{C}) * \mathsf{p}(\mathsf{C}) + \mathsf{p}(\mathsf{H}|\mathsf{P}) * \mathsf{p}(\mathsf{P})$. Here, $\mathsf{p}(\mathsf{H}|\mathsf{C})$, $\mathsf{p}(\mathsf{C})$, and $\mathsf{p}(\mathsf{P})$ can be interpreted in the standard way as degrees of belief in truth; but $\mathsf{p}(\mathsf{H}|\mathsf{P})$ must be interpreted as either a counterfactual or verisimilitude probability. Since $\mathsf{p}(\mathsf{H}|\mathsf{P})$ cannot be interpreted as a standard degree of belief, it follows that $\mathsf{p}(\mathsf{H})$ cannot be interpreted as a standard degree of belief either. Note that in most realistic cases, your degree of belief that you are in the classic case will be vanishingly small and your degree of belief that you are in the problem case will be very high. Hence, in general, it will be the case that $\mathsf{p}(\mathsf{H}) \approx \mathsf{p}(\mathsf{H}|\mathsf{P})$.

5.12 Conclusion

The interpretive problem arises whenever Bayesian probability distributions are assigned over sets of competing hypotheses, all of which are known to be false. I have discussed two solutions to the interpretive problem, both of which involve reinterpreting probability in such a way that the goals you have directly influence not just whether a given probability distribution is rational (as we also saw could happen in a different way in the preceding chapter), but even how the probability distribution is interpreted. Thus, there is an extreme kind of pragmatic encroachment on epistemic rationality whenever scientists are faced with the interpretive problem; moreover, I have argued that scientists are faced with the interpretive problem quite often.

Of course, the standard Bayesian interpretation also allows for pragmatic factors to play a role. According to the standard view, agents have both a probability

function and a utility function; any pragmatic factor – such as what the agent wants or is interested in – is relegated to the utility function. In this chapter, we have seen that this neat separation between the purely epistemic and the pragmatic fails in cases where we face the interpretive problem.

This chapter has largely been exploratory and some of the most interesting and difficult questions that arise from thinking about the interpretive problem have consequently been put aside. In particular, most of this chapter has focused on parameter inference inside a specific statistical model because few, if any, people seem to have realized that the interpretive problem even arises on this lower level. However, an arguably more interesting question is how to think of statistical model inference in a Bayesian framework.[22] Suppose we have two models $\mathfrak{m}$ and $\mathfrak{M}$, such that $\mathfrak{m}$ is nested inside $\mathfrak{M}$. In a sense $\mathfrak{m}$ cannot be closer to the truth than $\mathfrak{M}$ since any hypothesis in $\mathfrak{m}$ will also be in $\mathfrak{M}$. However, suppose the prior, $\mathfrak{p}$, over $\mathfrak{M}$ is centered on hypotheses that are very far from the truth and that the prior, $\mathfrak{q}$ over $\mathfrak{m}$ is centered on hypotheses very close to the truth. Clearly, in this case, the *pair* $(\mathfrak{m}, \mathfrak{q})$ is better than the *pair* $(\mathfrak{M}, \mathfrak{p})$.[23] These kinds of considerations lead me to think that the proper objects of verisimilitude measures in the context of model selection are model-prior pairs; it makes no sense to say that $\mathfrak{m}$ is closer to the truth than $\mathfrak{M}$ unless you take into consideration the priors over $\mathfrak{m}$ and $\mathfrak{M}$. The next natural step is

[22]Forster and Sober (1994) argue that Bayesians have particular difficulty making sense of model selection (see also Forster (1995)). They suggest that the key to model selection is predictive accuracy, and they suggest that predictive accuracy may be regarded as a kind of verisimilitude. Of course, there are different sorts of predictive accuracy (Forster, 2001, 2002), and consequently different kinds of verisimilitude. Given the verisimilitude interpretation presented here, these different kinds of verisimilitude can potentially all be accommodated in the Bayesian framework.

[23]Indeed, this is the case even if $\mathfrak{M}$ contains the truth and $\mathfrak{m}$ doesn't.

therefore to figure out how to measure the verisimilitude of model-prior pairs such as (M, p). This is work for the future.

6 Concluding Speculative Thoughts

I have argued that the Bayesian framework underwrites a kind of epistemic instrumentalism. More precisely, I have argued that the goals that agents have sometimes influence what Bayesian norms they ought to follow, what probability distributions they ought to use, and even how they ought to interpret the probability distributions that they use. If Bayesianism is a correct framework for epistemology and philosophy of science – as I tried to argue is at least a live option in the Introduction – then epistemic instrumentalism follows, by Modus Ponens.[1]

But what if Bayesianism is not a correct framework for epistemology and philosophy of science? After all, there are other quantitative frameworks for rationality—perhaps these other quantitative frameworks do not underwrite instrumentalism. On the other hand, maybe there *is* no correct *quantitative* framework for rationality—maybe rationality just cannot be quantified.

[1]Recall that when I say that Bayesianism is "a correct framework," I do not mean to say that the Bayesian framework is all there is to rationality, but only that it is a part of the story of what it takes to be rational. Indeed, the Bayesian framework is limited enough in its applicability that I personally think it cannot plausibly be the whole story. Even if we restrict ourselves to just statistical inference, Bayesian inference isn't plausibly the whole story.

The latter possibility seems implausible to me. Human beings are capable of being rational (sometimes); whatever it is that they do when they succeed in making a rational inference can surely be codified – it is not magic. And whatever that codification will turn out to be will surely in part be quantitative.

But what about the first possibility? There are several quantitative frameworks for rational belief change and decision making, and some of them may not support instrumentalism. This is a possibility, but again it is a possibility of which I am doubtful. The reason there is pragmatic encroachment in the Bayesian framework is (in part) because there are multiple ways of formally explicating important informal epistemic concepts, such as confirmation, information, opinionatedness, and so on, and pragmatic factors render some explications clearly better than others. My guess is that any quantitative framework that is rich enough to explicate all of our important epistemic concepts would at the same time allow for multiple reasonable formal explications. If that's the case, then there there would have to be some way of deciding between the different explications—and that's one place where pragmatic factors could come creeping in.

7 Appendix

7.1 Derivation of the Main Result of Chapter 3

For the sake of brevity, I will only derive the result that $\mathsf{c}(\mathsf{H},\mathsf{E}) = \mathsf{l}(\mathsf{H},\mathsf{E})$ when $\nu(x,\epsilon) = x + x(1-x)\epsilon$, since all three cases are very similar and this case is the most involved.

Proof. Starting with (1) from (MR), we have,

$$f(\nu(x,\epsilon),y) - f(x,y) = g(\epsilon) \tag{1}$$

If we divide each side by $x(1-x)\epsilon$, we get:

$$\frac{f(\nu(x,\epsilon),y) - f(x,y)}{x(1-x)\epsilon} = \frac{g(\epsilon)}{x(1-x)\epsilon} \tag{2}$$

Next, we let $\epsilon \to 0$:

$$\lim_{\epsilon \to 0} \frac{f(v(x,\epsilon),y) - f(x,y)}{x(1-x)\epsilon} = \lim_{\epsilon \to 0} \frac{g(\epsilon)}{x(1-x)\epsilon} \tag{3}$$

Since g is differentiable at 0 (from part (3) of (MR)), the right hand side of the above equation is just $\frac{1}{x(1-x)} g'(0)$. Since the limit exists on the right hand side of the equation, it must exit on the left side as well. But the left side is just $\frac{\partial}{\partial x} f(x,y)$. We therefore have,

$$\frac{\partial}{\partial x} f(x,y) = \frac{1}{x(1-x)} g'(0) \tag{4}$$

Next, we take the antiderivative of each side of (4) with respect to x. Since g and hence $g'(0)$ does not depend on x (from part (2) of (MR)), we have:

$$f(x,y) = g'(0)(\log x - \log(1-x)) + C \tag{5}$$

Here, C is a number that depends on y but not on x. If we perform the above calculations again starting instead with $f(0.5, v(y,\epsilon)) - f(0.5,y) = h(\epsilon)$ and using (5), we find that:

$$C = h'(0)(\log y - \log(1-y)) + K \tag{6}$$

Here, K is just a constant; i.e., it depends on neither x nor y. We therefore have:

$$f(x,y) = g'(0)(\log x - \log(1-x)) + h'(0)(\log y - \log(1-y)) + K \tag{7}$$

Now set $x = y = 0.5$. The second part of (CC) then entails that $K = 0$. Next, set

$x = y$. Then (CC) entails:

$$g'(0)(\log x - \log(1-x)) + h'(0)(\log x - \log(1-x) = 0 \tag{8}$$

This in turn entails that $g'(0) = -h'(0)$. Thus we have,

$$f(x,y) = -h'(0)(\log x - \log(1-x)) + h'(0)(\log y - \log(1-y)) \tag{9}$$

$$= h'(0) \log \frac{y}{1-y} * \frac{1-x}{x} \tag{10}$$

Remembering that $x = Pr(H)$ and $y = Pr(H|E)$, (9)-(10) together with (SF) entail:

$$c(H,E) = f(Pr(H), Pr(H|E)) \tag{11}$$

$$= h'(0) \log \frac{Pr(H|E)}{1-Pr(H|E)} * \frac{1-Pr(H)}{Pr(H)} \tag{12}$$

$$= h'(0) \log \frac{Pr(H|E)}{Pr(H)} * \frac{Pr(\neg H)}{Pr(\neg H|E)} \tag{13}$$

$$= h'(0) \log \frac{Pr(H|E) * Pr(E)}{Pr(H)} * \frac{Pr(\neg H)}{Pr(\neg H|E) * Pr(E)} \tag{14}$$

$$= h'(0) \log \frac{Pr(E|H)}{Pr(E|\neg H)} \tag{15}$$

Finally, (CC) entails that $h'(0)$ must be a positive number. Thus $c(H,E) = l$, up to multiplication by a positive number.

□

7.2 Derivations of (19) and (20) in Chapter 4

Proof. The first goal is to show that $\mathrm{Inf}_{\mathtt{lr}}(\mathfrak{p}) = -\sum \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \log \mathfrak{p}(\mathsf{H}_\mathfrak{i})$ under the condition that the posterior mass converges on some $\mathsf{H}_\mathfrak{i}$ as $\mathfrak{n} \to \infty$, for any imagined sequence $\mathsf{E}^\mathfrak{n}$ of evidence. In other words, for any $\mathsf{E}^\mathfrak{n}$, we require that there exists an $\mathsf{H}_\mathfrak{i}$ such that $\lim_{\mathfrak{n}\to\infty} \mathsf{P}(\mathsf{H}_\mathfrak{i}|\mathsf{E}^\mathfrak{n}) = 1$. To avoid clutter, I will suppress $\mathfrak{n}$ in the notation henceforth. Now, definition (18) in Chapter 4 with $\mathfrak{c} = \mathtt{lr}$ yields,

$$\mathrm{Inf}_{\mathtt{lr}}(\mathfrak{p}) = \lim_{\mathfrak{n}\to\infty} \sum_{\mathsf{E}} \sum_{\mathfrak{i}} \log \frac{\mathfrak{p}(\mathsf{H}_\mathfrak{i}|\mathsf{E})}{\mathfrak{p}(\mathsf{H}_\mathfrak{i})} \mathfrak{p}(\mathsf{H}_\mathfrak{i}, \mathsf{E}) \tag{16}$$

$$= \lim_{\mathfrak{n}\to\infty} \sum_{\mathsf{E}} \sum_{\mathfrak{i}} \log \mathfrak{p}(\mathsf{H}_\mathfrak{i}|\mathsf{E}) \mathfrak{p}(\mathsf{H}|\mathsf{E}) \mathfrak{p}(\mathsf{E}) - \lim_{\mathfrak{n}\to\infty} \sum_{\mathsf{E}} \sum_{\mathfrak{i}} \log \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \mathfrak{p}(\mathsf{H}_\mathfrak{i}, \mathsf{E}) \tag{17}$$

$$= \sum_{\mathfrak{i}} \lim_{\mathfrak{n}\to\infty} \sum_{\mathsf{E}} \mathfrak{p}(\mathsf{E}) \log \mathfrak{p}(\mathsf{H}_\mathfrak{i}|\mathsf{E}) \mathfrak{p}(\mathsf{H}|\mathsf{E}) - \sum_{\mathfrak{i}} \log \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \tag{18}$$

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

$$\mathrm{Inf}_{\mathtt{lr}}(\mathfrak{p}) = -\sum_{\mathfrak{i}} \log \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \mathfrak{p}(\mathsf{H}_\mathfrak{i}) \tag{19}$$

Which was the first thing to be proven. Note that no assumptions were made

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

about the sequence of evidence in the above derivation. This shows that the derivation does not depend on any such assumptions.

Now suppose that we instead plug $\mathfrak{c} = \mathfrak{d}$ into definition (18). Then the calculation becomes:

$$\mathrm{Inf}_{\mathfrak{d}}(\mathfrak{p}) = \lim_{\mathfrak{n}\to\infty} \sum_{E} \sum_{i} \big[\mathfrak{p}(H_i|E) - \mathfrak{p}(H_i)\big]\mathfrak{p}(H_i, E) \tag{20}$$

$$= \lim_{\mathfrak{n}\to\infty} \sum_{E} \sum_{i} \mathfrak{p}(H_i|E)\mathfrak{p}(H_i, E) - \lim_{\mathfrak{n}\to\infty} \sum_{E} \sum_{i} \mathfrak{p}(H_i)\mathfrak{p}(H_i, E) \tag{21}$$

$$= \lim_{\mathfrak{n}\to\infty} \sum_{E} \sum_{i} \mathfrak{p}(H_i|E)^2\mathfrak{p}(E) - \sum_{i} \mathfrak{p}(H_i)^2 \tag{22}$$

$$\tag{23}$$

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

$$\mathrm{Inf}_{\mathfrak{d}}(\mathfrak{p}) = \lim_{\mathfrak{n}\to\infty} \sum_{E} \mathfrak{p}(E) - \sum_{i} \mathfrak{p}(H_i)^2 \tag{24}$$

$$= 1 - \sum_{i} \mathfrak{p}(H_i)^2 \tag{25}$$

$$\tag{26}$$

And that was the second thing to be proven.

□

7.3 Derivation of (22) in Chapter 4

Proof. Here I will again suppress the k so that x in each case should really be read as x^k. Now, we are interested in evaluating the following expression:

$$\lim_{k\to\infty} E[d(\theta, x)] = \lim_{k\to\infty} \int_\Theta \int_X p(\theta, x)[(p(\theta|x) - p(\theta)]dxd\theta \tag{27}$$

$$= \lim_{k\to\infty} \int_\Theta [\int_X p(\theta, x)p(\theta|x)dx - \int_X [p(\theta, x)p(\theta)]dx]d\theta \tag{28}$$

$$= \int_\Theta \big[p(\theta)\big(\lim_{k\to\infty} \int_X p(x|\theta)p(\theta|x)dx\big) - p(\theta)^2\big]d\theta \tag{29}$$

Note that $\lim_{k\to\infty} \int_X p(x|\theta)p(\theta|x)dx$ does not depend on $p(\theta)$ because, as the amount of evidence goes to infinity, the posterior distribution, $p(\theta|x)$ becomes completely dominated by the likelihood $p(x|\theta)$. Therefore, (29) is of the form, $\int_\theta [p(\theta)g(\theta) - p(\theta)^2]d\theta$, where $g(\theta)$ does not depend on $p(\theta)$. Thus, the maximally non-informative prior is the $p(\theta)$ that maximizes the last integral subject to the constraint that $\int_\theta p(\theta) = 1$, since p is a probability density. This is a problem in the calculus of variations, and its solution is $p(\theta) = \frac{1}{2}(\lambda + g(\theta))$, for some constant λ.

Now, let us look at g. Suppose that the probability mass converges on some particular θ' as the evidence increases (which will happen in normal cases). That is, suppose there is some θ' such that for all θ not equal to θ', there exists for every $\epsilon > 0$ a k' such that for all $k > k'$, $p(\theta|x^k) < \epsilon$. Then it follows that, for $k > k'$,

$$g(\theta) = \int_X p(x|\theta)p(\theta|x)dx \tag{30}$$

$$< \int_X p(x|\theta)\epsilon dx \tag{31}$$

$$= \epsilon \tag{32}$$

Therefore, for all values of θ not equal to θ', $\lim_{k\to\infty} g(\theta) = 0$. Therefore, $p(\theta) = \frac{1}{2}(\lambda + g(\theta)) = \frac{1}{2}\lambda$. Thus, $p(\theta)$ is the flat prior that assigns $\frac{1}{2}\lambda$ to every value of θ.

□

7.4 Derivation of the Rational Requirement on the Prior in Chapter 5, Section 5.7

Proof. The goal is to show that if $f(p) = K$ and the derivative of f is bounded by (a, b), then, for every line $L = \alpha X + \beta$ such that $L(p) = k > K$ and $\alpha > b$, there will be another line $L^1 = \alpha^1 X + \beta^1$ such that $L^1(p) < k$, $\alpha^1 \in (a, b)$ and L^1 is closer to f than L is, using the verisimilitude measure $Max|f(X) - L(X)|$, where the maximum is taken over some range $X \in (p, P)$.

Suppose $L(X) = \alpha X + k - \alpha p$. Then $g(X) = L(X) - f(X)$ is strictly increasing, since $g'(X) = \alpha - f(X) > b - f(X) > 0$ and g is non-negative since $g(p) = 0$. Hence, the maximum of $|L(X) - f(X)|$ is simply $L(P) - f(P) = \alpha P + k - \alpha p - f(P)$, so that $\nu(L) = -\alpha(P - p) - k + f(P)$. By the same reasoning, the verisimilitude of

$L_1(X) = bX + K - bp$ is simply $\nu(L_1) = -b(P - p) - K + f(P)$. Thus, since $\alpha > b$ and $k > K$, it follows that $\nu(L_1) > \nu(L)$, i.e. L_1 has a higher verisimilitude than L.

□

references

Ahmed, Arif. 2012. Push the Button. *Philosophy of Science* 79(3):386–95.

Armendt, Brad. 2009. Stakes and Beliefs. *Philosophical Studies* 147(1):71–87.

Atkinson, David. 2009. Confirmation and Justification. A Commentary on Shogenji's Measure. *Synthese* 184(1):49–61.

Berger, James O., José M. Bernardo, and Dongchu Sun. 2009. The Formal Definition of Reference Priors. *The Annals of Statistics* 37(2):905–938.

Bernardo, José M. 1979a. Expected Information as Expected Utility. *The Annals of Statistics* 7(3):686–690.

———. 1979b. Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(2):113–147.

———. 1981. Reference Decisions. *Symposia Mathematica* XXV:85–94.

Bernardo, José M., and Adrian F. M. F. M Smith. 1994. *Bayesian theory*. Wiley, New York, NY.

Box, George E. P. 1979. Robustness in the Strategy of Scientific Model Building. In *Robustness in statistics: Proceedings of a workshop*, ed. R. L. Launer and G. N. Wilkinson. New York: Academic Press.

———. 1980. Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)* 143(4):383–430.

Brössel, Peter, and Franz Huber. 2014. Bayesian Confirmation: A Means With No End.

Cabrera, Frank. 2017. Dissertation manuscript. Ph.D. thesis, University of Wisconsin – Madison.

Carnap, Rudolph. 1962. *Logical foundations of probability.* 2nd ed. Chicago: University of Chicago Press.

Christensen, David. 1999. Measuring Confirmation. *Journal of Philosophy* XCVI: 437–61.

Crupi, Vincenzo, Nick Chater, and Katya Tentori. 2013. New Axioms for Probability and Likelihood Ratio Measures. *British Journal for the Philosophy of Science* 64: 189–204.

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

Crupi, Vincenzo, Katya Tentori, and Michel Gonzalez. 2007. On bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science* 74: 229–252.

DeGroot, Morris H. 1962. Uncertainty, Information, and Sequential Experiments. *The Annals of Mathematical Statistics* 33(2):404–419.

Easwaran, Kenny. 2016. Dr. Truthlove or: How I learned to Stop Worrying and Love Bayesian Probabilities. *Nous* 50(4):816–853.

Eells, Ellery. 1991. *Probabilistic causality*. Cambridge: Cambridge University Press.

van Enk, Steven J. 2014. Bayesian measures of confirmation from scoring rules. *Philosophy of Science* 81(1):101–113.

Fantl, Jeremy, and Matthew McGrath. 2002. Evidence, Pragmatics, and Justification. *Philosophical Review* 111(1):67–94.

Feldman, Richard. 2007. Reasonable Religious Disagreements. In *Philosophers without gods*, ed. L. Anthony, 194–214. Oxford: Oxford University Press.

Fitelson, Branden. 1999. The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science* 66:S362–78.

———. 2001. Studies in Bayesian Confirmation Theory. Ph.D. thesis, University of Wisconsin – Madison.

———. 2006. Logical Foundations of Evidential Support. *Philosophy of Science* 73: 500–12.

———. 2007. Likelihoodism, bayesianism, and relational confirmation. *Synthese* 156:473–489.

Forster, Malcolm R. 1995. Bayes and bust: Simplicity as a problem for a probabilist's approach to confirmation. *British Journal for the Philosophy of Science* 46(3):399–424.

———. 2001. The New Science of Simplicity. In *Simplicity, inference and modelling*, ed. Arnold Zellner, Hugo Keuzenkamp, and Michael McAleer, 83–117. Cambridge: Cambridge University Press.

———. 2002. Predictive Accuracy as an Achievable Goal of Science. *Philosophy of Science* 69:S124–S134.

Forster, Malcolm R., and Elliott Sober. 1994. How To Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science* 45(1):1–35.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology* 66: 8–38.

Gibbard, Allan. 2007. Rational Credence and the Value of Truth. In *Oxford studies in epistemology*, ed. Tamar Szabo Gendler and James Hawthorne. OUP Oxford.

Gillies, Donald. 1986. In Defense of the Popper-Miller Argument. *Philosophy of Science* 53:110–13.

Glass, David H., and Mark McCartney. 2015. A New Argument for the Likelihood Ratio Measure of Confirmation. *Acta Analytica* 30(1):59–65.

Glymour, Clark N. 1980. *Theory and Evidence.* Princeton University Press.

Good, I. J. 1960. Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society. Series B (Methodological)* 22(2):319–31.

———. 1984. The Best Explicatum for Weight of Evidence. *Journal of Statistical Computation and Simulation* 19:294–99.

———. 1985. Weight of evidence: A brief survey. In *Bayesian statistics 2*, ed. J. M Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M Smith, 249–270. Elsevier Science Publishers.

Hajek, Alan, and James Joyce. 2008. Confirmation. In *Routledge companion to the philosophy of science*, ed. Stathis Psillos and Martin Curd. Routledge.

Hawthorne, James. manuscript. A Better Bayesian Convergence Theorem. Manuscript.

Hilpinen, Risto. 1976. Approximate Truth and Truthlikeness. In *Formal methods in the methodology of empirical sciences*, 19–42. Synthese Library 103, Springer Netherlands.

Horowitz, Sophie. 2014. Immoderately Rational. *Philosophical Studies* 167(1):41–56.

Huber, Franz. 2008. Milne's Argument for the Log-Ratio Measure. *Philosophy of Science* 75:413–20.

Jaynes, E. T. 1989. Where Do We Stand on Maximum Entropy? In *Papers on probability, statistics and statistical physics*, ed. Roger Rosenkrantz, vol. 158 of *Synthese Library*, 210–314. Springer Netherlands.

———. 2003. *Probability theory: The logic of science*. Cambridge University Press.

Jeffrey, Richard. 1983. *The logic of decision*. 2nd ed. Cambridge University Press, Cambridge.

Jeffreys, Harold. 1946. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186(1007):453–461.

Joyce, James. 1998. A Non-Pragmatic Vindication of Probabilism. *Philosophy of Science* 65(4):575–603.

———. 1999. *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

———. 2009. Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In *Degrees of belief*, ed. Franz Huber and Christoph Schmidt-Petri. Synthese.

Kemeny, John G., and Paul Oppenheim. 1952. Degree of Factual Support. *Philosophy of Science* 19(4):307–324.

Key, Jane T., Luis R. Pericchi, and Adrian F. M. Smith. 1999. Bayesian Model Choice: What and Why? In *Bayesian statistics 6*, ed. José M. Bernardo, James O.

Berger, A. Phillip Dawid, and Adrian F. M. Smith, 343–370. Oxford: Oxford University Press.

Knaff, John A., and Raymond M. Zehr. 2007. Reexamination of Tropical Cyclone Wind–Pressure Relationship. *Weather and Forecasting* 22(1):71–88.

Knight, Frank H. 1921. *Risk, uncertainty and profit*. Boston, MA: Hart, Schaffner and Marx; Houghton Mifflin Co.

Kopec, Matthew, and Michael G. Titelbaum. 2016. The Uniqueness Thesis. *Philosophy Compass* 11(4):189–200.

Kullback, Solomon, and Richard Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* 22(1):79–86.

Lewis, David. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59(1):5–30.

Lewis, David K. 1973. *Counterfactuals*. Blackwell Publishers.

Milne, Peter. 1996. Log[P(hleb)/P(hlb)] Is the One True Measure of Confirmation. *Philosophy of Science* 63:21–6.

Morey, Richard D., Jan-Willem Romeijn, and Jeffrey N. Rouder. 2013. The Humble Bayesian: Model Checking From a Fully Bayesian Perspective. *British Journal of Mathematical and Statistical Psychology* 66(1):68–75.

Myrvold, Wayne. 2003. A Bayesian Account of the Virtue of Unification. *Philosophy of Science* 70(2):399–423.

———. 2016. On the Evidential Import of Unification. Unpublished manuscript.

Niiniluoto, Iikka. 1998. Verisimilitude: The Third Period. *British Journal for the Philosophy of Science* 49(1):1–29.

Northcott, Robert. 2013. Verisimilitude: A Causal Approach. *Synthese* 190(9): 1471–1488.

Oddie, Graham. 1997. Conditionalization, Cogency, and Cognitive Value. *British Journal for the Philosophy of Science* 48(4):533–541.

Okayasu, Emi. 2017. Justifying the Comparative Method in Historical Linguistics. Ph.D. thesis, University of Wisconsin – Madison.

Orzack, Steven Hecht, and Elliott Sober. 1993. A Critical Assessment of Levins's The Strategy of Model Building in Population Biology (1966). *The Quarterly Review of Biology* 68(4):533–546.

Pearl, Judea. 2009. *Causality: Models, reasoning, and inference*. 2nd ed. Cambridge: Cambridge University Press.

Pettigrew, Richard. 2016. *Accuracy and the Laws of Credence*. Oxford University Press.

Popper, Karl. 1963. *Conjectures and refutations: The growth of scientific knowledge*. London, Hutchinson.

Popper, Karl, and David Miller. 1983. The impossibility of inductive probability. *Nature* 302:687–88.

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

Rinard, Susanna. 2014. A new bayesian solution to the paradox of the ravens. *Philosophy of Science* 81(1):81–100.

Royall, Richard. 1997. *Statistical evidence: A likelihood paradigm*. CRC Press.

Savage, Leonard J. 1954. *The foundations of statistics*. New York: Dover Publications.

Schlesinger, G. 1995. Measuring degrees of confirmation. *Analysis* 55:208–12.

Schupbach, Jonah N. Forthcoming. Robustness Analysis as Explanatory Reasoning. *British Journal for the Philosophy of Science*.

Seidenfeld, Teddy. 1986. Entropy and Uncertainty. *Philosophy of Science* 53(4): 467–491.

Shaffer, Michael J. 2001. Bayesian Confirmation of Theories That Incorporate Idealizations. *Philosophy of Science* 68(1):36–52.

Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3):379–423.

Shogenji, Tomoji. 2012. The Degree of Epistemic Justification and the Conjunction Fallacy. *Synthese* 184(1):29–48.

Simpson, Edward H. 1949. Measurement of Diversity. *Nature* 163:688–688.

Sober, Elliott. 2008. *Evidence and evolution: The logic behind the science*. Cambridge University Press.

———. 2015. *Ockham's razors: A user's manual*. Cambridge University Press.

Sprenger, Jan. 2009. Statistics Between Inductive Logic and Empirical Science. *Journal of Applied Logic* 7(2):239–250.

———. 2012. The Renegade Subjectivist: Jose Bernardo's Objective Bayesianism. *Rationality, Markets and Morals* 3(50):1–13.

———. 2016. Conditional Degree of Belief. Manuscript.

Stanley, Jason. 2005. *Knowledge and Practical Interests*. Oxford: Oxford University Press.

Stevens, Stanley S. 1946. On the Theory of Scales of Measurement. *Science* 103(2684): 577–80.

Tentori, K, V Crupi, and D Osherson. 2007. Comparison of confirmation measures. *Cognition* 103:107–119.

Tsallis, Constantino. 1988. Possible Generalization of Boltzmann-Gibbs Statistics. *Journal of Statistical Physics* 52(1):479–487.

Vranas, Peter. 2004. Hempel's Raven Paradox: A Lacuna in the Standard Bayesian Solution. *British Journal for the Philosophy of Science* 42:393–401.

White, Roger. 2005. Epistemic Permissiveness. *Philosophical Perspectives* 19: 445–459.

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

Zalabardo, José. 2009. An Argument for the Likelihood Ratio Measure of Confirmation. *Analysis* 69:630–5.