Framework theories in science

by

Leah Henderson
B. Sc., University of Auckland (1994)
M. Sc., University of Auckland (1997)
DPhil, University of Oxford (2000)

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2010

Signature of
author.....

.....
Department of Linguistics and Philosophy
1 September, 2010

Certified by.....

.....
Robert Stalnaker
Laurance S. Rockefeller Professor of Philosophy
Thesis supervisor

Accepted by.....

.....
Alex Byrne
Chairperson, Department Committee on Graduate Students

Framework theories in science

by

Leah Henderson

Submitted to the Department of Linguistics and Philosophy
on 1st September 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis consists of three papers on the nature of scientific theories and inference.
In many cases, scientific theories can be regarded as organised into hierarchies, with higher levels sometimes called `paradigms', and lower levels encoding more specific or concrete hypotheses. The first chapter uses hierarchical Bayesian models to show that the assessment of higher level theories may proceed by the same Bayesian principles as the assessment of more specific hypotheses. It thus shows how the evaluation of higher level theories can be evidence-driven, despite claims to the contrary by authors such as Kuhn. The chapter also discusses how hierarchical Bayesian models may help to resolve certain issues for Bayesian philosophy of science, particularly how to model the introduction of new theories.
The second chapter discusses the relationship between Inference to the Best Explanation (IBE) and Bayesianism. Van Fraassen has raised the concern that the explanatory considerations in IBE go beyond the Bayesian formalism, making IBE incompatible with Bayesianism. The response so far has been that the explanatory considerations can be accommodated within the Bayesian formalism by stipulating that they should constrain the assignment of the probabilities. I suggest a third alternative, which is that the extra explanatory considerations have their origins in the relationship between higher and lower level theories and can be modelled in Bayesian terms without directly constraining the probabilities.
The third chapter discusses an aspect of the debate over scientific realism. The No Miracles argument and the Pessimistic Induction are often seen as the primary arguments for and against scientific realism. Yet recently it has been alleged that both of these arguments commit the base-rate fallacy. I argue that both arguments can be formulated in a non-fallacious manner, so neither should be dismissed on the grounds of faulty form alone.

Thesis supervisor: Robert Stalnaker
Title: Laurance S. Rockefeller Professor of Philosophy

Acknowledgements

I would like to thank Bob Stalnaker, my primary advisor, for his guidance with this thesis. It has been enormously improved by his insightful criticisms. Thanks also to my other advisors -- to Agustin Rayo for raising the big questions, and to Roger White for his very helpful comments.

Chapter 1 is the result of a collaborative effort with Josh Tenenbaum, Noah Goodman and Jim Woodward. I was very fortunate to work with this very stimulating team. I owe a particular intellectual debt to Josh Tenenbaum for inspiring many interesting thoughts. The project was funded in part by the James S.McDonnell Foundation Causal Learning Collaborative and was supported and encouraged by Alison Gopnik. Thanks to Zoubin Ghahramani for providing the code which we modified to produce the results and figures in the section on Bayesian curve-fitting. Thanks to Charles Kemp for his contributions, especially helpful discussions of HBMs in general as well as in connection to philosophy of science. I am grateful to Franz Huber, John Norton, Ken Schaffner and Jiji Zhang for reading versions of the manuscript and making helpful criticisms.

Ideas gleaned from Kevin Kelly's seminar on simplicity have made their way into this thesis. I would also like to thank Jonah Schupbach, Richard Holton and Sylvain Bromberger for useful discussions. Many thanks to Judy Thomson, Ned Hall, Steve Yablo, Alex Byrne and Selim Berker who all read early drafts of Chapter 2 and gave very helpful comments. Special thanks to Bernhard Nickel who read drafts of both Chapter 2 and 3 and provided general guidance as well as specific comments. I am grateful to Robert Nola for arranging that I could present a version of Chapter 3 in Auckland.

My husband Patrick van der Wel assisted with some of the figures. But he did far more than that. He discussed the ideas, read drafts and gave detailed comments. Most importantly though he provided the support, encouragement and interest which has allowed this thesis to be written.

I owe special thanks to my mother for her heroic efforts with childcare while I worked on this thesis, and to my father for enduring her absence while she was in the US helping me.

To my parents

Contents

Chapter 1

Framework theories and Bayesian inference

1.1 Introduction

Although there has been considerable disagreement over specifics, it has been a persistent theme in philosophy of science that scientific theories are hierarchically structured, with theoretical principles of an abstract or general nature at higher levels, and more concrete or specific hypotheses at lower levels. This idea has been particularly emphasised by such historically oriented writers as Kuhn, Lakatos and Laudan, who have used terms such as 'paradigms', 'research programs' or 'research traditions' to refer to higher levels in the hierarchy (Kuhn (1996); Lakatos (1978); Laudan (1978)). In this tradition, the mutual dependence and interactions of different levels of theory in the process of theory change has been explored in a predominantly qualitative way.

Meanwhile, confirmation theories have tended to ignore the hierarchical structure of theories. On a Bayesian view, for example, as in other formal accounts,

scientific theories have typically been regarded as hypotheses in an unstructured hypothesis space of mutually exclusive alternatives, and there has been a tendency to focus exclusively on confirmation and testing of specific hypotheses.

However, Bayesian models with a hierarchically structured hypothesis space are now widely used for statistical inference (Gelman et al. (2004)), and have proved particularly fruitful in modelling the development of individuals' 'intuitive theories' in cognitive science.[1] In this paper, we suggest that such Hierarchical Bayesian Models (or HBMs) can be helpful in illuminating the epistemology of scientific theories.[2] They provide a formal model of theory change at different levels of abstraction, and hence help to clarify how high level theory change may be rational and evidence-driven. This has been a central topic of debate following the appearance of Kuhn's *Structure of Scientific Revolutions.*

HBMs also help to resolve a number of philosophical worries surrounding Bayesianism. They can explain why logically stronger or simpler theories may be preferred by scientists and how learning of higher level theories is not simply parasitic on learning of lower level theories, but may play a role in guiding learning of specific theories. They also give a new and more satisfactory Bayesian model of the introduction of new theories.

In this paper, we first introduce HBMs in section 1.2, and argue that they capture essential features of the evaluation of scientific theories. The following three sections explain how HBMs may be used to resolve issues in Bayesian philosophy of science. Section 1.3 discusses the objection that Bayesians cannot account for a preference for logically stronger theories. Section 1.4 deals with the Bayesian

[1] Tenenbaum et al. (2007); Griffiths and Tenenbaum (2007); Mansinghka et al. (2006); Kemp et al. (2004); Kemp (2007); Kemp and Tenenbaum (2008)

[2] Parallels between intuitive theories and scientific theories are explicitly drawn in Carey and Spelke (1996), Giere (1996) and Gopnik (1996).

treatment of simplicity. Section 1.5 explains how HBMs can overcome many of the problems that the introduction of new theories presents to Bayesians. As well as discussing particular issues, two of these sections also introduce different examples of HBMs, in order to illustrate the variety of scientific theories to which HBMs may be applicable. Section 1.4 gives the example of curve-fitting while section 1.5 shows how HBMs may be used for learning about causal relations. In the final section 1.6, we consider the implications of HBMs for some general aspects of theory change.

1.2 Hierarchical Bayesian Models (HBMs)

The Bayesian model standardly used in philosophy of science operates with a hypothesis space $\mathcal{H}$, which is just a set of mutually exclusive alternative hypotheses. A 'prior' probability distribution is defined over the hypothesis space $p(T)$, $T \in \mathcal{H}$. On observing data D, the prior distribution is updated to the posterior distribution according to the rule of conditionalisation:

$$p(T) \rightarrow p(T|D) \tag{1.1}$$

The posterior distribution can be calculated using Bayes' rule to be

$$p(T|D) = \frac{p(T)p(D|T)}{p(D)} \tag{1.2}$$

Here $p(D|T)$ is the 'likelihood' of theory T, given data D, and $p(D)$ is the prior probability of the observed data D which serves as a normalisation constant ensuring that $p(T|D)$ is a valid probability distribution that sums to 1.[3]

[3]This may be expressed as:

$$p(D) = \sum_{T \in \mathcal{H}} p(D|T)p(T) \tag{1.3}$$

In a hierarchical Bayesian model, or HBM, the hypothesis space has a hierarchical structure. Given a particular theory at the $i+1$th level, one has a hypothesis space $\mathcal{H}_i$ of hypotheses or theories at the ith level which are treated as mutually exclusive alternatives. One defines a prior probability for a theory $T_i \in \mathcal{H}_i$ at level i which is conditional on the theory at the next level up, as $p(T_i|T_{i+1})$ for $T_i \in \mathcal{H}_i$, and $T_{i+1} \in \mathcal{H}_{i+1}$. This distribution is updated by conditionalisation in the usual way to give a posterior distribution, again conditional on T_{i+1}

$$p(T_i|T_{i+1}) \rightarrow p(T_i|D, T_{i+1}) \tag{1.4}$$

As in the non-hierarchical case, the posterior can be found using Bayes' rule as

$$p(T_i|D, T_{i+1}) = \frac{p(D|T_i, T_{i+1})p(T_i|T_{i+1})}{p(D|T_{i+1})} \tag{1.5}$$

In many cases, one can assume that $p(D|T_i, T_{i+1}) = p(D|T_i)$, that is, T_{i+1} adds no additional information regarding the likelihood of the data given T_i. (T_{i+1} is 'screened off' from D, given T_i).[4]

Theories at higher levels of the hierarchy may represent more abstract or general knowledge, whilst lower levels are more specific or concrete. For example, the problem of curve-fitting can be represented in a hierarchical model. Finding the curve which best represents the relationship between two variables X and Y, involves not only fitting particular curves from some given hypothesis space to the data, but also making 'higher' level decisions about which general family or functional form (lin-

(the sum is replaced by an integral if the hypotheses T are continuously varying quantities).

[4]The normalisation constant is calculated in a similar way to before as:

$$p(D|T_{i+1}) = \sum_{T_i \in \mathcal{H}_i} p(D|T_i)p(T_i|T_{i+1}) \tag{1.6}$$

Again, it is assumed that T_{i+1} is screened off from D, given T_i.

ear, quadratic etc.) is most appropriate. There may be a still higher level allowing choice between expansions in polynomials and expansions in Fourier series. At the lowest level of the hierarchical model representing curve-fitting, theories T_0 specify specific curves, such as $y = 2x + 3$ or $y = x^2 - 4$, that we fit to the data. At the next level of the hierarchy, theories T_1 are distinguished by the maximum degree of the polynomial they assign to curves in the low level hypothesis space. For instance, T_1 could be the theory Poly_1 with maximum polynomial degree 1. An alternative T_1 is Poly_2 with maximum polynomial degree 2, and so on. At a higher level, there are two possible theories which specify that T_1 theories are either polynomials, or Fourier series, respectively. The model also specifies the conditional probabilities $p(T_0|T_1)$ and $p(T_1|T_2)$. At each level of the HBM, the alternative theories are mutually exclusive. In this example, Poly_1 and Poly_2 are taken to be mutually exclusive alternatives. We will see soon how this should be understood.

We now suggest that HBMs are particularly apt models in certain respects of scientific inference. They provide a natural way to represent a broadly Kuhnian picture of the structure and dynamics of scientific theories.

Let us first highlight some of the key features of the structure and dynamics of scientific theories to which historians and philosophers with a historical orientation (Kuhn (1996); Lakatos (1978); Laudan (1978)) have been particularly attentive and for which HBMs provide a natural model. It has been common in philosophy of science, particularly in this tradition, to distinguish at least two levels of hierarchical structure: a higher level consisting of a paradigm, research program or research tradition, and a lower level of more specific theories or hypotheses.

Paradigms, research programmes and research traditions have been invested with a number of different roles. Kuhn's paradigms, for instance, may carry with them a commitment to specific forms of instrumentation, and to general theoretical

goals and methodologies, such as an emphasis on quantitative prediction or a distaste for unobservable entities. However, one of the primary functions of paradigms and their like is to contain what we will call 'framework theories',[5] which comprise abstract or general principles specifying the possible alternative hypotheses which it is reasonable to entertain at the more specific level - for example, the possible variables, concepts, and representational formats that may be used to formulate such alternatives, more general classes or kinds into which more specific variables fall, and possible relationships, causal and structural, that may obtain among variables. More generally, framework theories provide the raw materials out of which more specific theories may be constructed and constraints which these must satisfy. We will summarise this idea by saying that the relation between levels of theory is one of 'generation' where a lower level theory T_i is said to be *generated* from a higher level theory T_{i+1} when T_{i+1} provides a rule or recipe specifying constraints on the construction of T_i .

Framework theories are generally taken to define a certain epistemic situation for the evaluation of the specific theories they generate, since they help to determine the alternative hypotheses at the specific level and how likely they are with respect to one another. Confirmation of theories is relative to the framework which generates them. This type of idea may be illustrated even in the simple case of

[5]In discussing the application of HBMs to what we call 'framework theories', we intend to suggest relevance to several related notions. In cognitive development, the label 'framework theory' has been used to refer to the more abstract levels of children's intuitive theories of core domains – the organizing principles that structure knowledge of intuitive physics, intuitive psychology, intuitive biology and the like (Wellman and Gelman (1992)). In an earlier era of philosophy of science, Carnap introduced the notion of a 'linguistic framework', the metatheoretical language within which a scientific theory is formulated, which is adopted and evaluated on pragmatic or aesthetic grounds rather than being subject to empirical confirmation or disconfirmation. To the extent that there is common ground between Carnap's 'linguistic frameworks' and the later notions of 'paradigms', 'research programmes', or 'research traditions', as some have suggested (Godfrey-Smith (2003)), the term 'framework theory' also recalls Carnapian ideas.

curve-fitting. We can think of a scientist who fits a curve to the data from the set of alternatives characterized by or generated from Poly_1 as in a different epistemic or evidential situation from an investigator who fits a curve from the set of alternatives generated by Poly_2, even if the same curve is selected in both cases. The first investigator selects her curve from a different set of alternatives than the second and has more free parameters to exploit in achieving fit. This in turn affects the evidential support the data provides for the curve she selects. In part, Kuhn's concept of incommensurability reflects the idea that scientists working in different paradigms are in different epistemic situations. But the epistemic difference in the two situations need not be realized only in the minds of two different scientists. It applies also when a single scientist approaches the data from the standpoint of multiple paradigms or higher-level theories, weighing them against each other consciously or unconsciously, or when a community of scientists does likewise as a whole (without any one individual committing solely to a single framework).

Our thesis that HBMs provide a suitable model for the structure and dynamics of scientific theories, and particularly of this Kuhnian picture, rests on three core claims about how HBMs represent the scientific situation. First, we claim that the hierarchical hypothesis space in an HBM is appropriate for modelling scientific theories with hierarchical structure. Second, the notion of generation between levels of theory can be modelled formally in terms of the conditional probabilities $p(T_i|T_{i+1})$ linking levels of theory in an HBM. The conditional probabilities $p(T_i|T_{i+1})$ reflect the scientific assumptions about how T_i is constructed out of T_{i+1}, explicitly marking how the subjective probability of a lower-level theory is specified relative to, or with respect to the viewpoint of, the higher-level theory that generates it. And third, updating of the conditional probabilities $p(T_i|T_{i+1})$ of theories at level i with respect to a particular theory at the $i+1$ level represents confirmation of the level

i theory with respect to the class of alternatives generated by the $i+1$ level theory.

Before developing these claims in more detail, we first consider a few motivating examples of how higher-level framework theories may be structured and how they function to constrain more specific theories. The constraints which framework theories provide may take a variety of more specific forms: for example, they may reflect causal, structural, or classificatory presuppositions, or assumptions about the degree of homogeneity or heterogeneity of data obtained in different circumstances.

In the causal case, a framework theory could provide a 'causal schema', representing more abstract causal knowledge, such as that causal relations are only allowed between relata of certain types. A biological example is provided by the abstract description of the general principles that are now thought to govern gene regulation (eg. see Davidson (2006)). For example, current biological understanding distinguishes between structural and regulatory genes. These are organised into networks in which the regulatory genes influence the expression of both structural and other regulatory genes. Regulatory genes are also capable of changing independently of structural genes (e.g. by mutation). This represents a causal schema, which needs to be filled in with particular regulatory genes in order to yield a specific theory about the expression of any particular structural gene. Any alternative to this abstract schema (e.g. an alternative according to which gene expression is controlled by some other biological agent besides regulatory genes) will be represented by a competing higher level theory, which is inconsistent with the regulatory gene schema.

Another biological example is the so-called Central Dogma of Molecular Biology, suggested by Crick (1958) as a heuristic to guide research. According to this principle (in its universal, unqualified form) information flows from DNA to RNA to proteins but not vice versa. This can be represented by the abstract schema

DNA $\rightarrow$ RNA $\rightarrow$ protein. Specific lower level theories would fill in the details of the precise molecules involved. Competing high level theories to the central dogma would include schemas which also allow information to flow from RNA $\rightarrow$ DNA or Protein $\rightarrow$ DNA. In fact, the discovery of reverse transcriptase led to the replacement of the central dogma with an alternative schema, allowing information flow from RNA to DNA in certain cases. An example of the application of HBMs to causal networks is given in section 1.5.

In other applications, the specific theories of interest may be classifications or descriptions of a certain domain. Then a framework theory might specify the structure of the classification or description, for example whether the entities are organised into a tree, a lattice, or clusters, etc. Classification of living kinds was once thought to be a linear structure – each kind was to be placed in the great chain of being. Later Linnaeus discovered that the data were better organised into a tree, with branching structure. The linear structure and the tree structure were competing higher level theories, which were compared indirectly via how well specific theories of each type could account for the data[6].

Higher level theories may also specify how homogeneous data obtained from different trials or experimental settings are expected to be. Homogeneity assumptions can be represented as a higher level theory which can be learned, and they can help to guide further inference. For example, to a surprising extent genetic and molecular mechanisms are shared among different species of animals. This helps to make it plausible that say, results about the molecular mechanisms underlying synaptic plasticity in the sea slug (*aplysia*) can be generalised to give an understanding of synaptic plasticity in humans.

[6]Kemp and Tenenbaum (2008) and Kemp (2007) discuss these and other examples of structural frameworks, as well as showing how they can be learned in a HBM.

These examples illustrate that framework theories may take a wide range of representational forms. For instance, they, and the theories they generate, may be directed graphs, structural forms like trees or lattices, or multidimensional spaces. In principle, HBMs may be applied to theories of any kind of representational form, and current research is making these applications practical for such diverse representations as grammars, first order logic, lambda calculus, logic programs and more.[7]

We now turn to a more detailed discussion of how HBMs represent the structure and dynamics of scientific theories. Any model of scientific inference will take certain assumptions as given in the set-up of the model. These assumptions are then used as fixed points or common presuppositions in the evaluation of rival theories. For example, standard non-hierarchical Bayesian models presuppose a hypothesis space of rival candidate theories. We may think of this space as specified by the background assumptions that characterise a particular problem area – for example, that the hypotheses under consideration have a particular representational form, such as polynomials in curve-fitting or directed graphs in causal contexts. In an HBM, what has to be fixed in the set-up of the model is a hierarchical structure comprising the highest level hypothesis space and the conditional probabilities $p(T_i|T_{i+1})$ at each level. As we shall see in section 1.4.3, the background assumptions behind the highest level hypothesis space can be considerably more general and abstract than would typically be the case for a non-hierarchical Bayesian model. For this reason, in many cases, these background assumptions will be less demanding than the presuppositions required by non-hierarchical Bayesian models. The conditional probabilities $p(T_i|T_{i+1})$ can be thought of as reflecting scientists' judgments about how likely various lower level theories T_i are, given the higher level theory T_{i+1}. As

[7] This is current research by J. B. Tenenbaum and N. D. Goodman at MIT.

we will see in an example discussed in section 1.5, the higher level theory might specify the types of entities or relations involved in the lower level theories, and the conditional probability $p(T_i|T_{i+1})$ may be put together out of the probabilities that each entity or relation will take some particular form. The overall probability $p(T_i|T_{i+1})$ then reflects scientists' understanding of the principles governing how the lower level theories are to be cognitively constructed from the higher level theories. In other words, some assumptions about how T_{i+1} generates T_i are built into the set-up of the HBM.

As we mentioned earlier, updating of the conditional probabilities $p(T_i|T_{i+1})$ of theories at level i with respect to a particular theory at the $i+1$ level may be thought of as representing confirmation of the level i theory with respect to the class of alternatives generated by the $i+1$ level theory. For instance, the probability $p(2x+1|\text{Poly}_1)$ tells us about how likely the curve $2x+1$ is relative to a hypothesis space of lines of the form $y = \theta_0 + \theta_1 x$. On the other hand, the probability $p(0x^2+2x+1|\text{Poly}_2)$ tells us about how likely $0x^2+2x+1$ is with respect to the hypothesis space of quadratic curves $y = \theta_0 + \theta_1 x + \theta_2 x^2$. The fact that $p(2x+1|\text{Poly}_1)$ and $p(0x^2+2x+1|\text{Poly}_2)$ may differ, even though we may recognise $2x+1$ and $0x^2+2x+1$ as representing the same curve, reflects the framework relativity of confirmation mentioned earlier, namely that evaluations of theories may depend on the background knowledge, or higher level theory which frames the inquiry.

Thinking of higher level theories as generators of lower level theories contrasts with a certain traditional picture of higher level theories. According to this traditional approach, a hierarchy of theories can be regarded as a hierarchy of nested sets. On this view, there is a base set of all possible lowest level hypotheses, such as the set of all possible curves. In this base set, curves such as $2x+1$ and $0x^2+2x+1$

are taken to be the same hypothesis, so that the set contains only mutually exclusive hypotheses. The base set can be grouped into subsets sharing some common feature, such as the set of all nth order polynomials. Such subsets are then regarded as 'higher-level theories'. Thus, the set LIN of all linear hypotheses of the form $y = \theta_0 + \theta_1 x$ could be one higher level theory, and the set PAR of all quadratic hypotheses of the form $y = \theta_0 + \theta_1 x + \theta_2 x^2$ would be another. On this view, higher level theories like LIN and PAR are not mutually exclusive. For example, the curve represented by $2x + 1$ would be contained in both sets LIN and PAR.

By contrast, on the generative picture, higher level theories are mutually exclusive alternatives – this is a point stressed by Kuhn (Kuhn (1996), Ch. IX). This is also the case in an HBM, where theories at level i are treated as mutually exclusive alternatives, given a particular theory at the $i + 1$th level. For instance, the model Poly_1, together with the conditional probability $p(T_0|\text{Poly}_1)$, represents one way that scientists might think of specific theories T_0 as being constructed, or 'generated', whereas the model Poly_2 and probability $p(T_0|\text{Poly}_2)$ represents an alternative, and quite distinct way of producing specific theories. It is true that the sets of curves that each generates may overlap. However, the higher level theories Poly_1 and Poly_2 are not identified with the subset of curves that they generate. In this particular case, the HBM may be thought of as assigning probabilities to a tree-like hierarchy of theories, with arrows indicating a generation relation between a higher level theory and lower level theories that it generates (see Fig. 1).In some circumstances, one wants to evaluate theories without reference to a particular higher level theory. In the curve-fitting example, one might want to assign probabilities to specific curves from the base set of all possible curves. These form a mutually exclusive set. This can be done using the HBM by summing over the

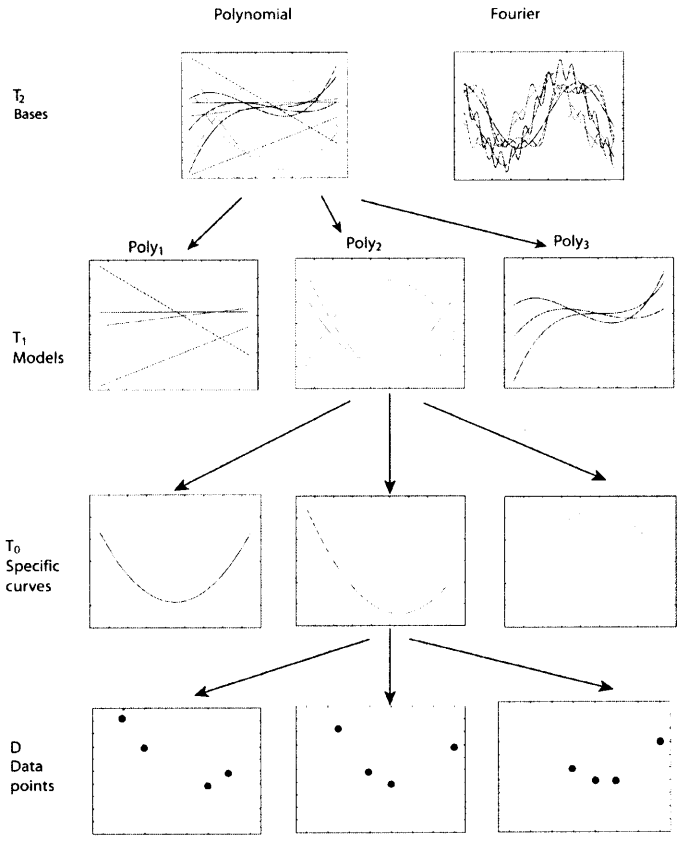

Figure 1-1: Hierarchical Bayesian Model for curve-fitting. At the highest level of the hierarchy, T_2 may represent either an expansion in a Fourier basis or in a polynomial basis. The polynomial theory generates theories, or models, T_1 of different degrees. Each of these then generates specific curves T_0 - quadratic curves are depicted. And each specific curve gives rise to possible data sets.

higher level theories which may generate the particular low level theory

$$p(T_0) = \sum_{T_1,..T_U} p(T_0|T_1)p(T_1|T_2)...p(T_{U-1}|T_U)p(T_U) \tag{1.7}$$

Here U indexes the highest level of the HBM. Probabilities for subsets of the base set, which on the traditional view comprise higher level theories, can also be calculated in this way.

1.3 Preference for stronger theories

We now turn to ways in which HBMs help to resolve certain challenges to Bayesian philosophy of science. The first problem we will consider was originally posed by Karl Popper (Popper (1959)). It has recently been repeated by Forster and Sober in the context of curve-fitting (Forster and Sober (1994); Forster (1995)).

The problem is the following. If one theory, T_1, entails another, T_2, then the following inequalities are theorems of the probability calculus:

$$p(T_1) \leq p(T_2) \tag{1.8}$$

$$p(T_1|D) \leq p(T_2|D) \tag{1.9}$$

for any data D. It would seem then that the Bayesian would always have to assign lower probability to the logically stronger theory. However, arguably scientists often do regard stronger theories as more probable.

Forster and Sober advance the argument in the context of curve-fitting (Forster and Sober (1994)). They define LIN to be the family of equations or curves of the

form:

$$Y = \alpha_0 + \alpha_1 X + \sigma U \tag{1.10}$$

and PAR to be the family of equations:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \sigma U \tag{1.11}$$

where σU is a noise term. The family LIN is then a subset of PAR, since 'if the true specific curve is in (LIN), it will also be in (PAR)' (p. 7, Forster and Sober (1994)). Forster and Sober claim that since LIN entails PAR, the Bayesian cannot explain how LIN can ever be preferred to PAR, because prior and posterior probabilities for LIN must always be less than or equal to the probabilities for PAR.

As we saw in the previous section, there are two ways to think of higher level theories: a set-based way, and a generative way. Forster and Sober assume that when scientists show a preference for a stronger theory, they are comparing sets of specific theories, such as LIN and PAR. However, the picture of high level theories involved in HBMs offers an alternative. The theories Poly_1 and Poly_2 considered at the T_1 level are mutually exclusive polynomial models, so it is quite legitimate to assign higher probabilities, whether prior or posterior, to Poly_1 as opposed to Poly_2. Therefore it is possible to prefer the linear theory Poly_1 over the quadratic theory Poly_2.

This is not in conflict with the assignment of lower probability to the theory LIN as opposed to PAR. Suppose Poly_1 has probability 0.6 in the HBM, and Poly_2 has probability 0.4 (assuming for the sake of simplicity that Poly_1 and Poly_2 are the only alternatives). The probability of LIN is the probability that the system is described by a linear hypothesis. A linear hypothesis could either be generated

by Poly_1, with probability 1, or by Poly_2, with some probability $p < 1$ depending on what weight Poly_2 gives to linear hypotheses (i.e. those quadratic hypotheses with $\beta_2 = 0$). Suppose $p = 0.1$. Then the probability for LIN is given by summing the probabilities for each generating model multiplied by the probability that if that model was chosen, a linear hypothesis would be drawn. Thus in this example, $p(\text{LIN}) = 0.6 \times 1 + 0.4 \times 0.1 = 0.64$. Similarly, the probability for PAR is $p(\text{PAR}) = 0.6 \times 1 + 0.4 \times 1 = 1$, since no matter whichever way the lower level hypothesis is generated, it will be a quadratic curve. Thus $p(\text{LIN}) \leq p(\text{PAR})$, as expected. However, the theories which are compared in an HBM are not LIN and PAR, but Poly_1 and Poly_2. This is because higher level theories are not regarded simply as sets of lower level possibilities, but are alternative generative theories.

The alleged failure of Bayesians to account for a preference for stronger theories has been associated with another alleged failure: to account for the preference for simpler theories. This is because the stronger theory may be the simpler one, as in the curve-fitting case. In the next section, we will argue that not only do HBMs allow preference for simpler theories, they actually automatically incorporate such a preference.

1.4 Curve-fitting

In fitting curves to data, the problem of fitting parameters to a function of a specified form, such as a polynomial of a certain degree, can be distinguished from the problem of choosing the right functional form to fit. There are statistical techniques of both Bayesian and non-Bayesian varieties for the latter problem of 'model selection'. It has already been suggested in the philosophy of science literature that particular versions of these methods may give a precise formalisation of the role

of simplicity in theory choice.[8] This section will give a more detailed account of Bayesian inference in the curve-fitting HBM introduced in section 1.2, describing inference at the three levels depicted in Figure 1. We will also show that Bayesian model selection, and hence a certain kind of preference for simplicity, arises automatically in higher level inference in HBMs.

At each level of the hierarchy, the posterior distribution is computed for hypotheses in the hypothesis space generated by the theory at the next level up the hierarchy.

1.4.1 Inference at lowest level: Bayesian model-fitting

At the lowest level of the hierarchy, the hypothesis space $\mathcal{H}_0$ comprises specific curves T_0 of the form

$$f_{\theta,M}(x) = \theta_0 + \theta_1 x + ... + \theta_M x^M + \varepsilon \tag{1.12}$$

(where $\varepsilon \sim N(0, \sigma^2)$ is the noise term), generated by Poly_M at the T_1 level. Let $\tilde{\theta} = (\theta_0, ..., \theta_M)$ be a vector representing the parameters of the curve to be fitted. For simplicity, we treat the variance σ^2 as a fixed quantity, rather than as a parameter to be fitted.

The posterior probability for the parameters $\tilde{\theta}$ is

$$p(\tilde{\theta}|D, Poly_M) = \frac{p(D|\tilde{\theta}, \text{Poly}_M)p(\tilde{\theta}|\text{Poly}_M)}{p(D|\text{Poly}_M)} \tag{1.13}$$

[8] Forster and Sober (1994) suggest this for the non-Bayesian method based on the Akaike Information Criterion (or AIC) and Dowe et al. (2007) for Minimum Message Length (or MML), which is a Bayesian form of model selection.

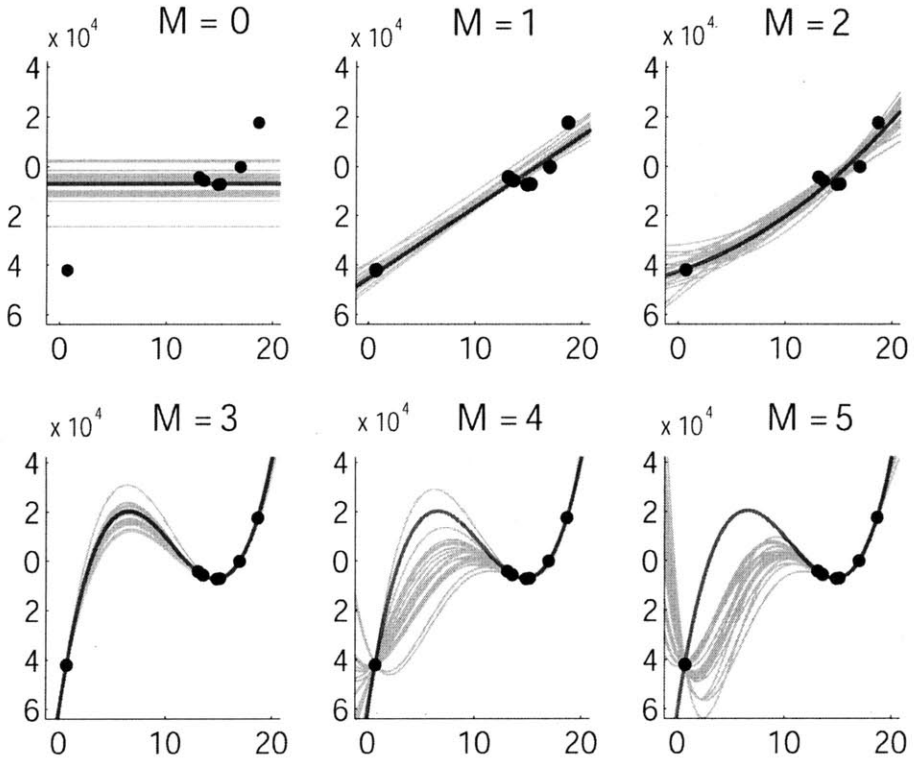

Figure 1-2: The polynomial of each degree with highest posterior (dark grey), with other polynomials sampled from the posterior (light grey). (Data is sampled from the polynomial $f(x) = 100(x-3)(x-12)(x-17)$, plus normally distributed noise.)

The denominator is given by

$$p(D|\text{Poly}_M) = \int p(D|\tilde{\theta}, \text{Poly}_M) p(\tilde{\theta}|\text{Poly}_M) d\tilde{\theta} \tag{1.14}$$

Figure 2 shows the polynomial of each degree with highest posterior probability for a small data set, together with samples from the posterior which illustrate the 'spread' of the posterior distribution.

The posterior is used by the Bayesian for the task of fitting the parameters to the data, given a model – the problem of 'model-fitting'. Strictly speaking, Bayesian assessment of hypotheses involves only the posterior probability distribution. However, one could also 'select' the best hypothesis, for example by choosing the one with the highest posterior probability.

1.4.2 Inference at second level: Bayesian model selection

At the next level of the hierarchy, the hypothesis space $\mathcal{H}_1$ consists of the polynomial models $\{\text{Poly}_M\}_{M=1}^{\infty}$ with different degrees M. These models may be compared by calculating their posterior probabilities, given by[9]

$$P(\text{Poly}_M|D) \propto P(\text{Poly}_M) P(D|\text{Poly}_M)$$

where

$$P(D|\text{Poly}_M) = \int_{\tilde{\theta}} P(D|\tilde{\theta}) P(\tilde{\theta}|Poly_M) d\tilde{\theta} \tag{1.15}$$

Computing the posterior distribution over models in this way is the way models at the second level of the HBM are assessed, and it is also the standard Bayesian

[9]Once the parameters $\tilde{\theta}$ of the polynomial are defined, so is the maximum degree of the polynomial. Therefore the screening off assumption mentioned after equation 1.5 holds and $p(D|\tilde{\theta},\text{Poly}_M) = p(D|\tilde{\theta})$.

approach to the problem of model selection (or 'model comparison', if the Bayesian strictly restricts herself to the posterior probability distribution). Although the posterior indicates the relative support for a theory Poly_M, the model is not directly supported by the data, but is indirectly confirmed through support for the specific functions $f_{\theta,M}(x)$ that it generates.

It has been observed by a number of authors that, with a certain natural choice of priors, the Bayesian posterior over models reflects a preference for simpler models, and Bayesian model selection involves a trade-off between the complexity of the model and fit to the data similar to that seen in other non-Bayesian approaches to model selection.[10]

We illustrate this in Figure 3, which shows the posterior probabilities for each model and how they change as data accumulates (this is shown for both polynomial and Fourier models). The prior probability over models has been assumed to be uniform, and the probability has also been distributed uniformly between specific polynomials in each hypothesis space. This choice does not imply equal probability for specific polynomials generated by different theories: individual polynomials have prior probability that decreases as degree increases, since they must 'share' the probability mass with more competitors.[11] With these priors, when the amount of data is small, the linear model Poly_1 is preferred over the higher order polynomial models. As the amount of data increases, the higher order models become more probable. If linear models may be regarded as 'simpler' than higher order models,

[10] Rosenkrantz (1977) discusses the role of simplicity in Bayesian evaluations of families of curves and other examples (see his discussion of what he calls 'sample coverage'). Similar ideas are discussed for a simple case in White (2005). Jefferys and Berger (1992) and MacKay (2003) highlight the trade-off between simplicity and fit in Bayesian model selection.

[11] Results shown in Fig. 3 were produced using a uniform prior over a finite number of models. If the number of model classes is countably infinite, one could use a geometric or exponential distribution over model classes.

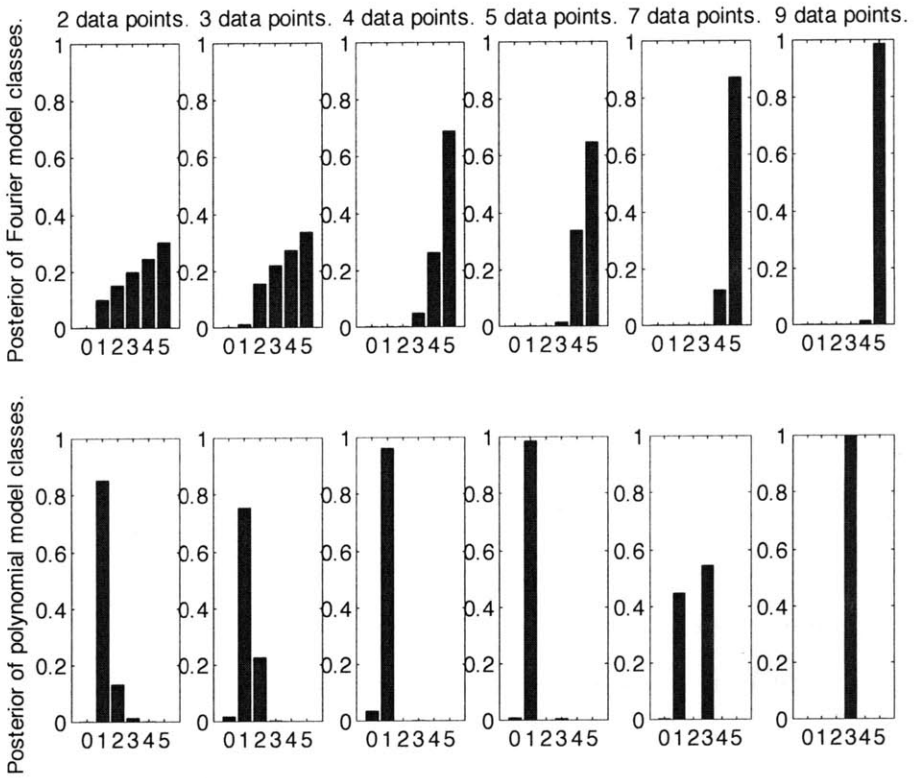

Figure 1-3: Posterior probability of models with different M (horizontal axis) for both the polynomial case, and the Fourier case discussed in section 1.4.3. The Bayesian Occam's razor is evident in the favouring of simpler models when the amount of data is small.

then the Bayesian posterior has a tendency to favour simpler models, at least until there is an accumulation of data supporting a more complex theory. This is a *Bayesian Occam's razor*: when there are only a few data points, the data can be fitted either by a line or by a quadratic (cubic, etc.), however the linear model Poly_1 is preferred because it is 'simpler'.

This simplicity preference arises because the posterior on models, Equation 1.15, involves an integral over *all* the polynomials generated by the model, not just the best fitting. Since there are more quadratics that fit poorly than there are lines (indeed, there are more quadratics than lines, period) the quadratic model is penalised.

This effect is manifested generally in the posterior probability for a higher level theory T_i. The likelihood $p(D|T_i)$ for this theory is obtained by integrating over all the possible specific models T_{i-1} that T_i generates:[12]

$$p(D|T_i) = \int p(D|T_{i-1})p(T_{i-1}|T_i)dT_{i-1} \tag{1.16}$$

That is, the likelihood of a high level theory is the expected likelihood of the specific theories that it generates. This will be large when there are relatively many specific theories, with high prior, that fit the data well—since complex higher level theories tend to have many specific theories which fit the data poorly, even when they have a few that fit the data very well, simplicity is preferred. For this preference, it is not essential that the priors $p(T_{i-1}|T_i)$ be exactly uniform, as they were in our illustration. All that is needed is that the priors for lower level theories are not weighted heavily in favour of those theories which fit the data best. Intuitively, the likelihood $p(D|T_i)$ penalises complexity of the model: if the model is more complex,

[12] If screening off does not hold, $p(D|T_{i-1})$ should be replaced by $p(D|T_{i-1}, T_i)$.

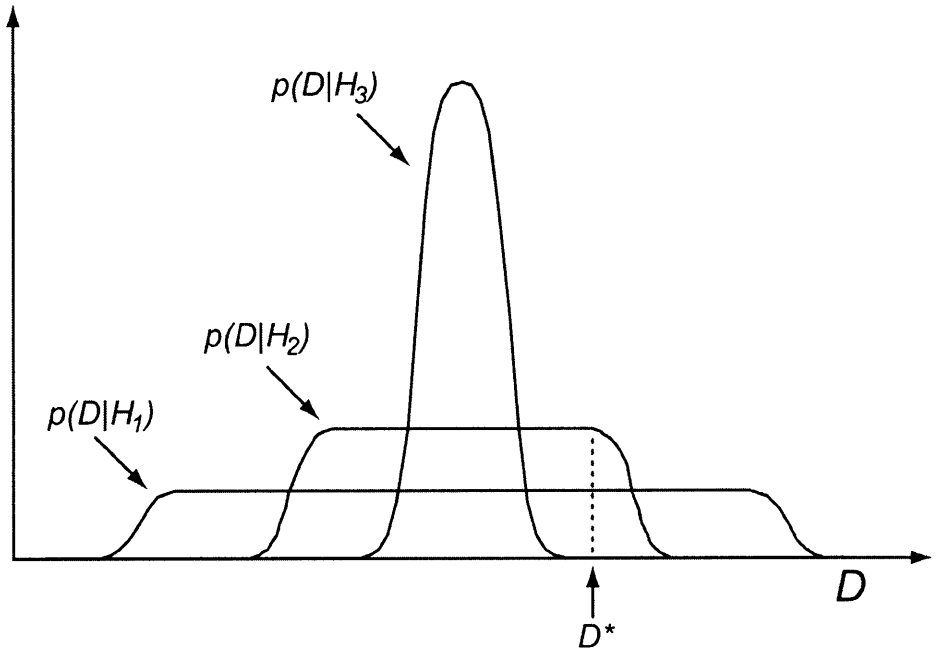

Figure 1-4: Probability distributions $p(D|H_i)$ over a one-dimensional data set D for three different theories. The more complex theory H_1 spreads the probability mass over a wider range of possible data sets than the simpler theories H_2 and H_3. For the observed data D^*, the complex theory H_1 has lower likelihood than H_2. The simpler theory H_3 doesn't spread its mass so widely, but misses the observed data D^*. In this case, the theory of intermediate complexity, H_2, will be favoured.

then it will have greater flexibility in fitting the data, and could also generate a number of other data sets; thus, the probability assigned to this particular data set will be lower than that assigned by a less flexible model (which would spread its probability mass over fewer potential data sets) - see Figure 4. This simplicity preference balances against fit to the data, rather than overriding it: as we see in Figure 3, an initial simplicity bias can be overcome by the accumulation of data supporting a more complex theory.

1.4.3 Inference at higher levels: Bayesian 'framework theory' selection

We have seen how at the second level of the HBM, we can compare or select the appropriate degree M for a polynomial model.

Each polynomial model Poly_M generates a set of specific hypotheses differing only in parameter values. All the polynomial models are expansions to different degrees in terms of polynomial functions. However, this is not the only way that models could be constructed. Models could also be expansions to degree M in terms of Fourier basis functions. The model Fouri_M, for example, would generate specific functions of the form $f_{\theta,M}(x) = \theta_0 + \theta_1 \sin(x) + ... + \theta_M \sin(Mx) + \varepsilon$.

In an HBM, comparison between the type of basis functions used can take place at a third level of the hierarchy. The principles are the same as those behind comparison of models at the second level. One finds the posterior probability:

$$P(\text{Basis}|D) \propto P(\text{Basis})P(D|\text{Basis})$$

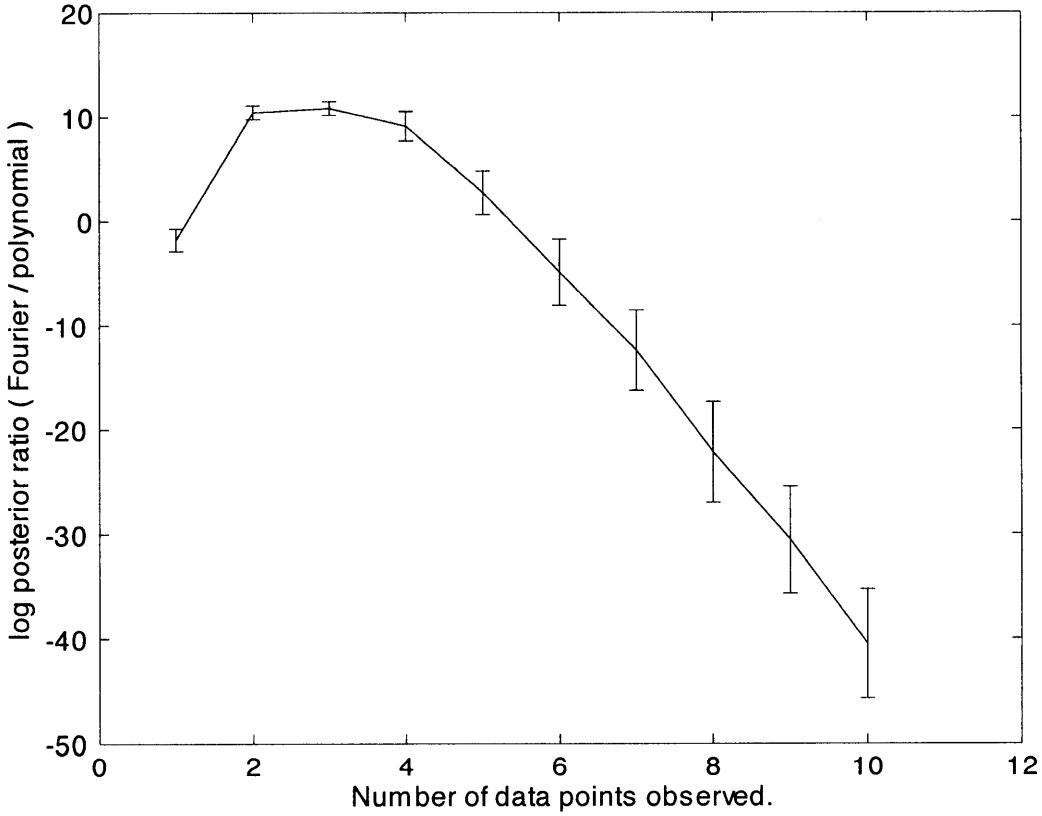

Figure 1-5: The log-posterior probability ratio between bases for curve-fitting (positive numbers indicate support for Fourier basis, negative for polynomial basis). Error bars represent standard error over twenty data sets randomly sampled from the polynomial $f(x) = 100(x-3)(x-12)(x-17)$ (plus normally distributed noise). When the number of observations is small the Fourier basis is favoured, eventually enough evidence accumulates to confirm the (correct) polynomial basis.

with

$$P(D|\text{Basis}) = \sum_{\text{Model}\in\text{Basis}} P(D|\text{Model})P(\text{Model}|\text{Basis})$$

where Model will be one of the Poly_Mor Fouri_M, depending on the basis.[13] Just as the models receive support from the evidence through the specific functions below them, the curve-fitting bases receive support through the models they generate. In Figure 5 the posterior probability values for each basis are plotted against the number of data points observed (the data is actually sampled from a cubic polynomial with noise). Note that there is a great deal of uncertainty when only a few data points are observed—indeed, the Fourier basis has higher posterior—but the correct (polynomial) basis gradually becomes confirmed. Since there are only two hypotheses at the highest level (polynomial or Fourier), we have made the natural assumption that the two are a priori equally likely: $P(\text{Basis}) = 0.5$.

In some respects, the choice of a basis in this simple curve-fitting example is analogous to a choice between 'framework theories' (see section 1.2). Framework theories frame the possibilities for how theories are expressed at lower levels. They may even be, as in Carnap's picture of linguistic frameworks, something like a language for expressing theories. In this example, we have a natural comparison between the 'language of polynomials', with a simple 'grammar' built from variables and constants, addition, multiplication and exponentiation and an alternative 'Fourier language', built from trigonometric functions, addition, constants, and variables. Since any function may be approximated arbitrarily well by polynomials or sinusoids (a standard result of analysis), the two languages are equally powerful in allowing fit to the data, so the main determinant of choice between them is simplicity as reflected in the likelihood of the framework theories. Simplicity here is a

[13] Once the model is specified, the basis is also given, so $p(D|\text{Model}) = p(D|\text{Model, Basis})$.

criterion which arises naturally from assessing the empirical support of a high-level hypothesis.

1.5 The problem of new theories

One of the most pressing challenges for Bayesian philosophy of science is to account for the discovery or introduction of new theories. When a genuinely new theory is introduced, the hypothesis space changes, and the Bayesian will have to reassign the prior distribution over the new hypothesis space. This has been called the 'problem of new theories' for Bayesians, because the adoption of a new prior is not governed by conditionalisation and so is strictly speaking a non-Bayesian process (Earman (1992)).

The main Bayesian proposal to deal with the problem has been to use a 'catch-all' hypothesis to represent as-yet-unformulated theories, and then 'shave off' probability mass from the catch-all to assign to new theories. This is an unsatisfactory solution since there is no particularly principled way to decide how much initial probability should be assigned to the catch-all, or how to update the probabilities when a new theory is introduced.

Given the inadequacy of this proposal, even would-be-full-time Bayesians like Earman have given up on a Bayesian solution and turned to a qualitative account of the introduction of new theories, such as that proposed by Kuhn. Earman appeals to the process of coming to community consensus, and suggests that the redistribution of probabilities over the competing theories is accomplished by a process of 'persuasions rather than proof' (Earman (1992), p. 197).

Difficulties in describing changes to the hypothesis space have also led to another alleged problem. Donald Gillies claims that Bayesians must limit themselves

to situations where the theoretical framework - by which he means the space of possible theories - can be fixed in advance. 'Roughly the thesis is that Bayesianism can be validly applied only if we are in a situation in which there is a fixed and known theoretical framework which it is reasonable to suppose will not be altered in the course of the investigation', where 'theoretical framework' refers to 'the set of theories under consideration' (Gillies (2001), p. 364). Gillies claims that this poses an enormous problem of practicality, since it would not be feasible to consider the 'whole series of arcane possible hypotheses' (p. 368) in advance. He thinks that for the Bayesian to 'begin every investigation by considering all possible hypotheses which might be encountered in the course of the investigation' would be a 'waste of time' (p. 376). This claim is motivated by consideration of the potentially enormous size of adequate hypothesis spaces, even for simple examples.

We will argue that both the problem of new theories and the practicality problem for large hypothesis spaces are alleviated if assignment of a prior probability distribution does not depend on an explicit enumeration of the hypothesis space. As we said in section 1.2, just as the application of non-hierarchical Bayesianism is restricted to a particular fixed hypothesis space, so HBM Bayesianism can be validly applied only if we are in a situation in which there is a fixed and known hierarchy which it is reasonable to suppose will not be altered in the course of the investigation. However, part of this hierarchy (the conditional probabilities $p(T_i|T_{i+1})$) represent background assumptions about how lower level theories are generated from higher level theories. Given these assumptions, there is no need to enumerate the lower level theories. In fact Bayesian inference in an HBM can be performed over very large, and even infinite hypothesis spaces. These considerations provide a solution to the problem of practicality which Gillies raises. Also, there can be theories implicit in the hypothesis space, initially with very low probability, which

come to get high probability as the data accumulates. This provides a way of effectively modeling the introduction of theories that are 'new' in the sense that they may be regarded as implicit in assumptions about how the lower level theories are generated although not explicitly enumerated or recognized as possible hypotheses.

To illustrate, we will use an example of an HBM that represents scientific theories about causal relations (Tenenbaum and Nigoyi (2003); Griffiths and Tenenbaum (2007)). The example also serves to illustrate another application of HBMs. Directed graphs where the arrows are given a causal interpretation are now a popular way to represent different possible systems of causal relationships between certain variables. These are called 'causal graphs'. More abstract causal knowledge may be represented by a 'causal graph schema' which generates lower level causal graphs.

Consider a family of causal graphs representing different possible systems of causal relationships among such variables as smoking, lung cancer, heart disease, headache and cough where an arrow from one variable or node X directed into another Y means that X causes Y. Compare the graphs in Figure 6. The three graphs N_0, N_1 and N_2employ the same set of variables. Although these graphs posit different causal links among the variables, they differ in a systematic way from graph N_3. In N_0, N_1 and N_2, the variables fall into three more general classes which might be described as behaviours, diseases and symptoms. Furthermore there is a more abstract pattern that governs possible causal relationships among variables in these classes. Direct causal links run only from behaviours to diseases and from diseases to symptoms. Other possible causal links (e.g. direct causal links from behaviours to symptoms or causal links between symptoms) do not occur. By contrast, N_3 does not follow this pattern – in this graph, for example, the disease variable, flu, causes the behaviour variable, smoking.

The particular graphs N_0, N_1 and N_2 (but not N_3) are generated by a more

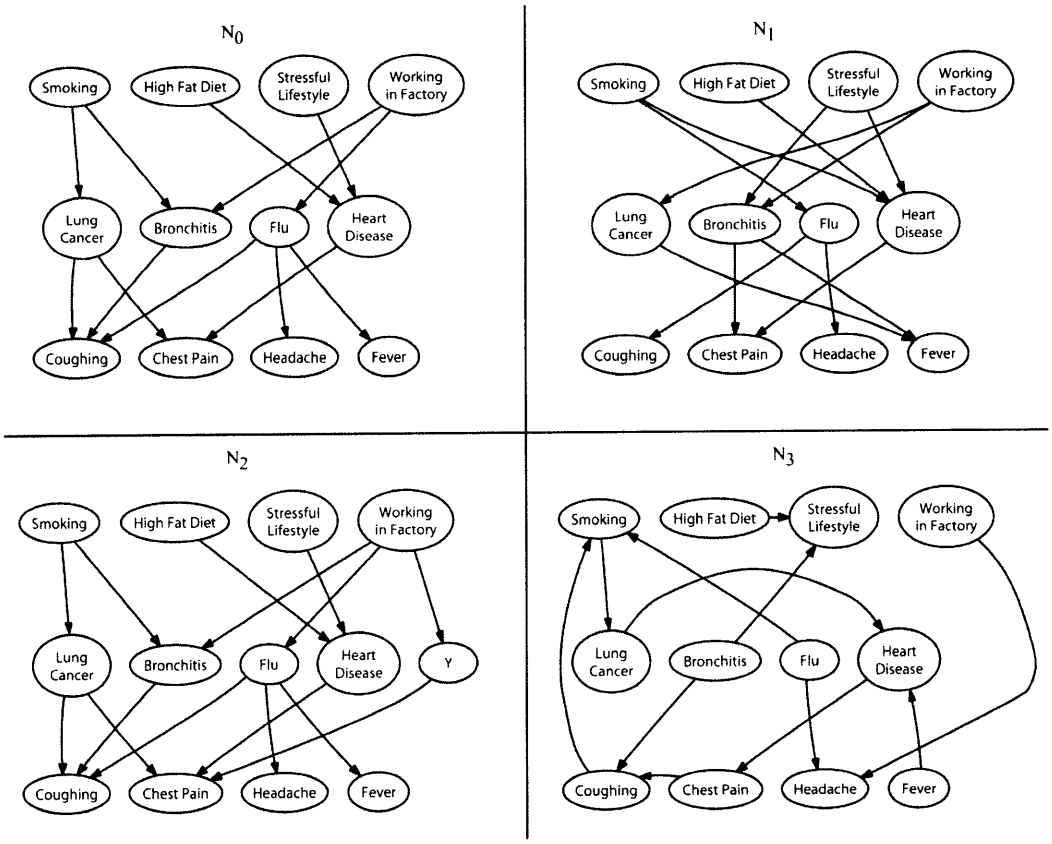

Figure 1-6: Causal networks illustrating different possible relationships between behaviours, diseases and symptoms. N_0, N_1 and N_2 are based on the same abstract graph schema G_{dis}, whereas N_3 is not. The network N_2 contains an extra disease node.

abstract graph schema G_{dis} which is characterised by the following features:

i) There are three node classes B, D and S into which specific nodes fall. Each node class is open in the sense that additional variables may be added to that class.

ii) Possible causal relationships take the form $B \rightarrow D$ and $D \rightarrow S$ only

i) and ii) thus represent structural features that characterise an entire class of more specific theories. These structural features have to do with causal relationships (or their absence) that are determined by the classes into which variables fall.

In an HBM we may regard G_{dis} as a general theory, T_1, generating specific networks N_i as specific theories T_0. It divides the variables of interest into classes or kinds and specifies that only a limited set of causal relationships may hold among these variables, in much the same way that the Central Dogma of molecular biology restricts the set of possible informational relationships among DNA, RNA, and proteins. In order to specify the HBM completely, we need to define the prior $p(N_i|G_{dis})$, (i.e. $p(T_0|T_1)$), which encapsulates probabilistic information about how the specific networks N_i depend on, or are generated by, the causal schema G_{dis}. As an illustration, in Griffiths and Tenenbaum (2007), the prior $p(N_i|G_{dis})$ was specified by giving probability distributions for the number of nodes in each class (B, D, or S) in the network N_i, and distributions for the probability of causal links between pairs of nodes from classes B and D, and from classes D and S. More specifically, the number of nodes in each class was assumed to follow a power law distribution $p(N) \sim \frac{1}{N^{\alpha}}$ with an exponent specific to each class (so that, other things being equal, graphs with more nodes have lower prior probability). There was also assumed to be a fixed probability η_{BD} of a causal link from b to d for any nodes $b \in B$, and $d \in D$, and a fixed probability η_{DS} of a causal link from d to s for any nodes $d \in D$, and $s \in S$. Thus the probability of a causal link depends only on the classes to which the nodes belong. A specific causal graph such as N_0 may

then be generated by randomly drawing individual nodes from each node class and then randomly generating causal links between each pair of nodes. The result is a probability $p(N_i|G_{dis})$ for each specific causal graph N_i, which is non-zero if and only if G_{dis} generates N_i.

At the outset of the investigation, the range of graphs to be considered need not be explicitly enumerated. The range of hypotheses is implicitly determined by the causal schema (or schemas) under consideration and the 'instructions' we have just given for building hypotheses and their prior probabilities at the lower level based on the schema. At first, a high probability is assigned to certain of the possible causal graphs – for instance those with fewer disease variables. However, a causal network containing a new disease can be discovered, given the right data, even though initially all the networks with non-negligible prior probability do not contain this disease. Suppose for example that a correlation is observed between a previously known behaviour b like working in a factory and a previously known symptom s like chest pain. To accommodate this correlation, the logically simplest possibility is simply to add another causal link directly from b to s, but the schema G_{dis} rules out this possibility: any link between a behaviour and symptom must pass through a disease as an intermediate link. Another possibility which is allowed by G_{dis} is to add links from b to one of the known diseases and from this disease to s. This has the advantage that no new disease node needs to be introduced. But it may be that any new links from working in a factory to existing disease nodes and from these to symptoms generate correlations that are inconsistent with what is observed. In such circumstances, one will eventually learn that the correct causal graph is one which introduces a new disease node Y which is causally between b and s as shown in N_2. The rules associated with the graph schema G_{dis} for constructing specific graphs tell us what the prior is for this new graph and as we update on the

basis of the data, this new graph may acquire a higher posterior than any of its competitors Griffiths and Tenenbaum (2007).

In general, HBMs can provide a Bayesian model of the introduction of new theories.[14] New theories which were implicit in the hypothesis space, but which initially received very low prior probability, can be explicitly introduced and come to receive high posterior probability as the appropriate supporting data accumulates. The example also illustrates how the higher level theory may play a role in guiding the construction of more specific theories. What G_{dis} in effect does is to provide a sort of abstract recipe for the construction or generation of more specific theories. By restricting the range of possible hypotheses among which the learner has to search, G_{dis} makes it possible to learn the correct hypothesis from a much smaller body of data than would be required if one were instead searching a much larger space of possible alternatives. So the adoption of the schema represented by G_{dis} greatly facilitates learning.

The lack of need to explicitly enumerate hypotheses also removes the practical problem for large hypothesis spaces posed by Gillies. In the context of HBMs, one might be concerned that the evaluation of posterior probabilities, although in principle possible, is too computationally challenging. However, Bayesian inference in large HBMs is made practical by the existence of algorithms for producing good

[14] Earman suggests distinguishing 'weak revolutions', which involve the introduction of theories where the new theory is a possibility that was within the space of theories, previously unarticulated, from revolutions proper, or 'strong revolutions', where a completely new possibility is introduced. HBMs provide a Bayesian treatment of weak revolutions. This is important for at least two reasons. First, given the ubiquity of weak revolutions in day-to – day science it would be a serious limitation if the Bayesian account could not deal with them without making the implausible assumption that all weakly new hypotheses need to be explicitly enumerated before inference begins. Second, it is far from clear how common 'pure' strong revolutions are. Detailed investigations of putative examples of such revolutions typically reveals a major guiding role from previously accepted frameworks, suggesting that at least some aspects of such episodes can be modeled as weak revolutions.

approximations to the posterior probabilities. Indeed, there are a number of ways to efficiently approximate Bayesian inference that appear, prima facie, very different from the usual method of explicit enumeration and computation that Gillies criticises. For instance, in Markov Chain Monte Carlo (MCMC) the posterior distribution is approximated by sequentially sampling hypotheses as follows. From the current hypothesis a 'proposal' for a new hypothesis is made using some heuristic—usually by randomly altering the current hypothesis. Next, the current hypothesis is compared to the proposed hypothesis, resulting in a stochastic decision to accept or reject the proposal. This comparison involves evaluation of the ratio of prior and likelihood functions, but not the (properly normalised) posterior probability. With a proper choice of proposals, the resulting sequence of hypotheses is guaranteed to comprise a set of samples from the posterior distribution over hypotheses which can be used to approximate the distribution itself.[15]

In the case of an HBM in which one level of theory generates the hypotheses of a lower level, each step of sequential sampling which changes the higher level can allow access to entirely different hypotheses at the lower level. Thus, while an arbitrary variety of alternative specific theories is available, only a small portion need be considered at any one time. Indeed, the sequence of local moves used to approximate posterior inference may never reach most of the hypothesis space – though in principle these hypotheses could be accessed if the evidence warranted.

It has been demonstrated that MCMC provides an effective way to implement Bayesian learning in a computational model of the disease-symptom example (Mansinghka et al. (2006)). The MCMC method is used to learn both the specific causal graph and the division of variables into the classes which appear in the higher level graph schema. For instance, to learn the causal schema G_{dis} it would have to

[15]For more details, see eg. MacKay (2003).

be discovered that the variables can be divided into three classes ('behaviours' B, 'diseases' D and 'symptoms' S) with causal links from B to D and from D to S. The size of the hypothesis space is extremely large in this example, but the model can still effectively find an appropriate theory in reasonable time.

The MCMC method can even be regarded as a suggestive metaphor for the process of scientific discovery. It highlights two ways in which the Bayesian approach to science may be more realistic than has often been assumed. First, as just described, it is possible to efficiently approximate Bayesian models, even over infinite hypothesis spaces, without 'wasting' an inordinate amount of time considering very unlikely hypotheses. These approximation methods provide for an orderly, rule-governed process by which new possibilities are introduced and considered. Second, such approximation can have a qualitative character that is very different from exact Bayesian computation: the approximate search may look locally arbitrary, even irrational, mixing elements of hypothesis testing and heuristic change, but it still arrives at the rational Bayesian answer in the long run.

1.6 Broader implications for theory change

We have argued so far that HBMs help to resolve certain issues for Bayesian philosophy of science. In particular, they give a Bayesian account of high level theory change and of the introduction of new theories. In addition, they allow us to resolve puzzles associated with the preference for stronger or simpler theories.

HBMs also have implications for general discussions of theory construction and theory change which are not specifically Bayesian. A number of traditional accounts of how abstract knowledge may be learned proceed 'bottom-up'. For instance, in the logical empiricist tradition, more 'observational' hypotheses must be learned first,

with the acquisition of the more theoretical level following, rather than guiding learning at the observational level. Such a bottom-up picture has led to puzzlement about why we need theories at all (Hempel (1958)). It has been alleged that this is a particularly pressing problem for a Bayesian, since a theory presumably should always receive lower probability than its observational consequences (Glymour (1980b), pp. 83-84).

This problem is dissolved in the HBM analysis, which validates the intuition – central in Kuhn's programme but more generally appealing – that higher level theories play a role in guiding lower level learning.[16] In section 1.5 we saw how higher level theories may guide the search through a large lower level hypothesis space by 'spot-lighting' the particular subset of lower level hypotheses to be under active consideration. In both the curve fitting and causal network problems discussed in previous sections, it is possible for a hierarchical Bayesian learner given a certain sample of evidence to be more confident about higher level hypotheses than lower level knowledge, and to use the constraints provided by these higher-level hypotheses to facilitate faster and more accurate learning at the lower level. In one representative case study, Mansinghka, Kemp, Tenenbaum, and Griffiths (Mansinghka et al. (2006)) studied learning of a causal network with 16 variables according to a simple 'disease theory' schema (variables divided into two classes corresponding to 'diseases' and 'symptoms', with causal links connecting each disease to several symptoms). A hierarchical Bayesian learner needed only a few tens of examples to learn this abstract structure. It was found that after only 20 examples, the correct schema dominated in posterior probability – most of the posterior probability was

[16] Also, since the relation between levels in a HBM is not logical entailment, but generation, probability assignments are not constrained by entailment relations between levels. Indeed, theories at different levels of an HBM are not in the same hypothesis space, and so are not directly compared in probabilistic terms.

placed on causal links from diseases to symptoms – even though specific causal links (between specific pairs of variables) were impossible to identify. After seeing a few more examples, the hierarchical Bayesian learner was able to use the learned schema to provide strong inductive constraints on lower level inferences, detecting the presence or absence of specific causal links between conditions with near-perfect accuracy. In contrast, a purely bottom-up, empiricist learner (using a uniform prior over all causal networks) made a number of 'false alarm' inferences, assigning significant posterior probability to causal links that do not exist and indeed should not exist under the correct abstract theory – because they run from symptoms to diseases, or from one symptom to another. Only the hierarchical Bayesian learner can acquire these principles as inductive constraints and simultaneously use them to guide causal learning.[17]

HBMs illuminate aspects of theory change which have been controversial in the aftermath of Kuhn's *The Structure of Scientific Revolutions.* A number of commentators have contended that on Kuhn's characterization, high level theory change, or paradigm shift, was a largely irrational process, even a matter of 'mob psychology' (Lakatos (1978), p. 91). Considerable effort was devoted to providing accounts which showed that such changes could be 'rational'. However, these accounts were handicapped by the absence of a formal account of how confirmation of higher level theories might work. HBMs provide such an account.

HBMs also help to resolve an ongoing debate between 'one-process' and 'two-process' accounts of scientific theory change (as described in Chap. 7, Godfrey-Smith (2003)). If scientific knowledge is organised into levels, this opens up the possibility that different processes of change might be operative at the different levels – for example, the processes governing change at the level of specific theo-

[17] See Mansinghka et al. (2006), particularly Figure 4.

ries or the way in which these are controlled by evidence might be quite different from the processes governing change at the higher levels of theory. Carnap held a version of this 'two-process' view – he held that changes to a 'framework' were quite different from changes within the framework. Similarly, Kuhn thought that the processes at work when there was a paradigm change were quite different from the processes governing change within a paradigm (that is choice of one specific theory or another). Part of the motivation for two-process views has been the idea that change at lower levels of theory is driven by empirical observations whereas change at higher levels is driven more by pragmatic, social or conventional criteria. Carnap, for example, thought that changes to a 'framework' were mostly governed by virtues like simplicity which were primarily pragmatic, not empirical.

On the other hand, there have been those who favour a single general account of theory change. Popper and Quine may plausibly be regarded as proponents of this 'one-process' view. According to Popper, change at every level of science from the most specific to the most general and abstract proceeds (or at least as a normative matter ought to proceed) in accord with a single process of conjecture and refutation (Popper (1959), Popper (1963)). According to Quine, all changes to our 'web of belief' involve the same general process in which we accommodate new experience via a holistic process of adjustment guided by considerations of simplicity and a desire to keep changes 'small' when possible (Quine (1986)).

HBMs allow us to make sense of valuable insights from both the one-process and two-process viewpoints, which previously seemed contradictory. Within the HBM formalism, there is a sense in which evaluation at higher framework levels is the same as evaluation at lower levels, and also a sense in which it is different. It is the same in the sense that it is fundamentally empirical, resting on the principle of Bayesian updating. This reflects the judgment of the one-process school that

all theory change ultimately has an empirical basis. Yet evaluation of framework theories is different from that of specific hypotheses in the sense that it is more indirect. In HBMs, framework theories, unlike more specific hypotheses, cannot be directly tested against data. Instead they are judged on whether the hypotheses they give rise to do well on the data – or more precisely, whether the specific theories they generate with high probability themselves tend to assign high probability to the observations. As we have seen, when this Bayesian principle of inference is applied to higher levels of a hierarchy of theories, it can lead to effects that would seem to depend on ostensibly non-empirical criteria such as simplicity and pragmatic utility. Thus, HBMs also reflect the judgment of the two-process school that criteria like simplicity can be the immediate drivers of framework change, although in this picture those criteria are ultimately grounded in empirical considerations in a hierarchical context. In place of the 'one process' versus 'two process' debate that animated much of twentieth-century philosophy of science, we might consider a new slogan for understanding the structure of scientific theories and the dynamics by which they change: 'Many levels, one dynamical principle'.

Chapter 2

Bayesianism and Inference to the Best Explanation

2.1 Introduction

There are different philosophical theories about how scientific theories are evaluated in the light of available evidence. Such accounts attempt both to capture key features of scientific practice and to identify principles on which scientific theory assessment should rest. In this chapter I will consider the relationship between two of the most influential philosophical accounts of theory assessment. One is Bayesianism, according to which we believe different hypotheses to differing degrees; these degrees of belief are represented by probabilities; and the probabilities are updated in the light of new evidence by a certain rule, known as 'conditionalisation'. Another is 'Inference to the best explanation' or 'IBE', according to which scientific theory assessment involves inferring the theory which provides the best explanation of the available evidence.

How are these two accounts related? Although it is traditional to characterise

IBE as specifying the theory in which one should believe, the comparison with Bayesianism can be facilitated by formulating IBE also in terms of degrees of belief. In this version, one has greater degrees of belief in hypotheses which provide better explanations of the given evidence. Is this version of IBE then compatible with Bayesianism?

Bas van Fraassen has claimed that the answer to this question is 'no'. IBE and Bayesianism are, on his view, incompatible alternatives, since IBE is a rule which adds extra 'bonus' probability onto more explanatory hypotheses, over and above what they would have received from Bayesian conditionalisation alone.

In response to van Fraassen, several authors have proposed what I call the 'composite view' of the relationship between IBE and Bayesianism (Okasha (2000), Lipton (2004), Weisberg (2009)). According to this view, IBE and Bayesianism are compatible, and the best account of scientific inference is a composite of elements of IBE and elements of Bayesianism. In particular, Bayesianism supplies the rule for updating probabilities, and explanatory considerations determine the probabilities that one starts with.

In this chapter, I will argue that van Fraassen has not established that Bayesianism is incompatible with IBE. However, I also claim that the composite view lacks independent motivation beyond the desire to establish compatibility between IBE and Bayesianism. I propose an alternative way in which IBE and Bayesianism could go together compatibly. This is based on isolating in the concept of IBE certain core considerations which determine how the explanationist assigns degrees of belief, and then showing how these core considerations can be represented in Bayesian terms. On this account of the relationship between IBE and Bayesianism, there is no need for explanatory considerations to explicitly constrain the assignment of Bayesian probabilities, as in the composite view.

2.2 IBE

Better explanation is often cited by scientists themselves as a reason for giving a certain theory extra weight. For example, in *On the origin of species*, Charles Darwin argues for his theory of natural selection on the grounds that it explains 'large classes of facts' in a more satisfactory manner than its rivals, in particular special creationism (Darwin (1962)). IBE appears prominently in recent philosophical accounts of scientific method[1], and is part of a long-standing tradition of describing the scientific method in qualitative terms.

Since IBE is intended as an account of theory assessment, it should be interpreted in the following way: out of a range of candidate theories each of which provides some explanation of the data, select the theory which provides the best explanation. The explanations provided by these candidate theories are, in Hempel's sense, 'potential explanations' (Hempel (1965), p. 338). That is, unlike in an 'actual explanation', the explanans does not have to be true.

In attempting to spell out the IBE account, an obvious first question is: what does it mean to provide the best, or a better, explanation? Peter Lipton, one of the main proponents of IBE, says that 'the best explanation is the one which would, if correct, be the most explanatory or provide the most understanding'. He calls this the 'loveliest' explanation (Lipton (2004), p. 59). To the extent that an account of what constitutes better explanation or loveliness is offered in the IBE literature, it is usually what might be termed the 'virtues view'. A theory is judged to provide a better explanation if it possesses some optimal combination of 'explanatory virtues' in relation to the phenomena. Explanatory virtues which are commonly mentioned are simplicity, unification, precision, scope, and fruitfulness.

[1]See for example Harman (1965), Day and Kincaid (1994), Niiniluoto (1999) and references therein.

Theories of inductive inference may be more or less restrictive in what they regard as permissible inductive inferences, given particular hypotheses and data. IBE is generally taken to be towards the more restrictive end of the spectrum, with not much room for intersubjective disagreement between explanationists over which theories provide better explanations.

In the following, I will follow the practice in the literature of using the term 'explanationist' to refer both to an agent who performs IBE and to someone who espouses IBE as an account of inference. A similar double usage applies in the case of 'Bayesian'.

2.3 Bayesianism

Bayesians assume that belief comes in degrees. A minimalistic version of Bayesianism, which I will call 'Min Bayes', tells you how to represent these degrees of belief and how to update them as more evidence is gathered. Min Bayes has two tenets. First, the *coherence* postulate says that degrees of belief in the different hypotheses should be represented by a probability distribution (or density) over the hypotheses. Second, updating is conducted according to the rule of *conditionalisation*. According to this rule, when evidence or data D is obtained, the probabilities are updated by replacing the 'prior' probability in hypothesis T, $p(T)$, with a 'posterior' probability which is given by the conditional probability, $p(T|D)$. The conditional probability $p(T|D)$ can be computed from the prior probability $p(T)$, the 'likelihood' $p(D|T)$, and the initial probability of observing the evidence, $p(D)$, using Bayes' rule:

$$p(T|D) = \frac{p(D|T)p(T)}{p(D)} \tag{2.1}$$

The probability $p(D)$ is essentially a normalisation constant which ensures that the posterior distribution is a valid probability distribution which sums to 1.[2]

Versions of Bayesianism which hold that Min Bayes is close to the complete story about rational constraints on inductive inference are known as 'subjective' forms of Bayesianism. According to subjective Bayesianism, there is a wide range of rationally permissible assessments of a set of hypotheses in the light of certain evidence.

Rather than being viewed as a complete account of inference, Min Bayes may also be viewed as an 'inductive framework', which can be filled in in various ways to produce a more restrictive account of what counts as a permissible inductive inference.[3] The inductive framework is filled in by adding constraints on the assignment of priors and likelihoods. The constraints may be provided by other accounts of inference which could in some cases be regarded as more fundamental.

Bayesians have also proposed further constraints of their own which result in what are known as more 'objective' forms of Bayesianism. One such constraint is that the Bayesian probabilities representing 'subjective' degrees of belief should satisfy a chance-credence principle, such as David Lewis's 'Principal Principle' (Lewis (1986)). A simple version of this is:

> The subjective probability that A is the case, on the supposition that the objective chance of A equals x, equals x. i.e. $p(A|ch(A) = x) = x$.[4]

[2]It can be computed from the 'likelihood function' which is a function of the hypotheses T, giving the likelihood for each T given particular D, as:

$$p(D) = \sum_{T \in \mathcal{H}} p(D|T)p(T) \tag{2.2}$$

(the sum is replaced by an integral if the hypotheses T are continuously varying quantities).

[3]The terminology here is used by Strevens (2004), who contrasts an 'inductive framework' with the filled in framework he calls an 'inductive logic'.

[4]Van Fraassen refers to this version of the Principal Principle as 'Miller's principle' (Van

In practice, there are few Bayesians of the completely subjective variety. Most would accept at least this principle as a constraint on probabilities.

More objective Bayesians have traditionally invoked the Principle of Indifference to constrain prior probabilities. This principle says that if there is no evidence to distinguish one hypothesis from another, they should be assigned equal prior probabilities.[5] Although objections have been raised to more objective forms of Bayesianism, there are also strategies and approaches which have been proposed to get around the alleged difficulties. I will mention a few of these in section 2.8.

2.4 The incompatibilist view

Van Fraassen interprets IBE as a rule for updating which is incompatible with Bayesian conditionalisation (Van Fraassen (1989)). Consider a scenario where we have hypotheses $T_1 ... T_n$ which are mutually exclusive and exhaustive, and some prior distribution over these. By conditionalising on evidence D we obtain a posterior probability distribution. Van Fraassen's claim is that IBE should be understood as a rule which assigns greater probability to the more 'explanatory' hypotheses than they would receive in the posterior distribution obtained by conditionalising. This is because the explanationist takes into account 'not only on the initial probabilities and the series of outcomes', but also a factor we might call 'explanatory success' (Van Fraassen (1989), p. 166). Van Fraassen does not fill in the details of how this rule is supposed to work. The claim is simply that it assigns extra bonuses to explanatory hypotheses over and above what they would receive from conditionalisation.

Fraassen (1989), p.82)

[5]Objective Bayesian positions are set forth in Jeffreys (1998), Jaynes and Bretthorst (2003), and Rosenkrantz (1977). They are widely criticised, for example in Seidenfeld (1979).

Van Fraassen casts his argument in the form of a dialogue between an 'orthodox Bayesian' whom he calls 'Peter', and an itinerant Preacher of IBE. The Preacher tries to convert Peter to infer to the best explanation. The discussion takes place surrounding the particular example of an 'alien die' (that is, a die which is discovered on another planet). There are a number of hypotheses $T_1 ... T_n$ concerning the bias of the die. The hypothesis T_i says that the chance of throwing an ace on any given toss is $\frac{i}{n}$. Peter assigns equal prior probabilities to the different bias hypotheses, and assigns likelihoods in accordance with the Principal Principle. That is, he takes $p(ace|T_i) = \frac{i}{n}$. He throws the die and observes a series of aces. He then forms a posterior probability distribution by conditionalisation on this evidence. The Preacher comes and tells Peter that hypotheses of high bias are a better explanation of throwing a series of aces than hypotheses of lower bias. Therefore, the Preacher says, 'you should raise your credence in the more explanatory hypotheses. "What?" exclaims Peter. "More than I would anyway?" "Yes" says the Preacher' (Van Fraassen (1989), p. 166).

How we interpret van Fraassen's argument here depends on exactly what kind of Bayesian Peter is. Suppose he is a follower of Min Bayes, but accepts no further constraints on his probability assignments. In this case, he has assigned a prior and likelihood in accordance with the Principle of Indifference and the Principal Principle, but he could just as easily have done something else and been none the less rational for it. Then if IBE involves assigning higher probabilities to more explanatory hypotheses, Peter could have assigned priors and likelihoods which would result in these probabilities on conditionalisation. There is no need to interpret IBE as a non-conditionalising rule.

On the other hand, suppose Peter is an objective Bayesian. It wasn't just a whim which made him assign the prior and likelihood that he did. He didn't think

there were any other rational options than to satisfy the Principle of Indifference and the Principal Principle. Peter has a strong sense of what he would 'do anyway', if the Preacher were not about. He is not just waiting on Preachers to tell him how to fix his probabilities. If Peter follows the Preacher's recommendations then, this will require him to violate some of his own principles: either conditionalisation, or one of the principles which he has used to fix priors and likelihoods.

The basic problem with van Fraassen's proposal is that it is actually not clear that there has to be a divergence between the Bayesian posterior and the probability distribution produced by IBE. Peter may not need to do anything differently in order to conform with IBE. Although it is logically possible for the probability distribution assigned by the Bayesian and by the explanationist to diverge in the way van Fraassen suggests, nothing about the example he gives inclines us to think they do. In fact, it seems rather easy to understand the alien die example without conflict between IBE and Bayesian conditionalisation. As more and more aces are thrown, the explanationist regards high bias hypotheses as increasingly clearly the best explanation. But this is just in accordance with the increase in Bayesian posterior probabilities. Why should we think that IBE involves assigning any more probability to the 'explanatory' hypotheses than conditionalisation would? One might regard van Fraassen's model as a misinterpretation of IBE in the probabilistic context.[6] Rather, IBE and Bayesianism could still go together compatibly. The

[6]One might think that van Fraassen has simply produced an 'idiosyncratic' interpretation of IBE in the probabilistic context (Okasha (2000), p. 703). However, in my view, van Fraassen's interpretation is not so much idiosyncratic as based on the way IBE is put to use in certain contexts as an argument against anti-realism. One of the key arguments against the underdetermination of theories by evidence is that explanatory considerations provide extra evidence for the truth of a theory which go beyond any evidence we may have for a theory's empirical adequacy (van Fraassen (1980), pp. 153-155). On this view, IBE would be expected to contain extra elements which cannot be fitted into a standard Bayesian story about confirmation. Van Fraassen claims that if this is what is meant by IBE, it will fall foul of the dynamic Dutch-book arguments which justify Bayesian conditionalisation, and will thus be unsuitable to use as a rule of inference (Van

question is how.

2.5 The composite view

To date, the most common view of how IBE and Bayesianism go together compatibly is the 'composite view'. According to this view, IBE and Bayesianism complement one another, both making a contribution to the correct account of inference. The view is predicated on the idea that Bayesianism is not a complete or adequate picture of inference on its own, and it needs to be helped out or supplemented by IBE.

The composite view appears to have several attractions. For proponents of IBE, it is attractive because it preserves a key role for explanatory considerations in inference. They can make use of the 'powerful and well-studied framework' that Bayesianism provides (Weisberg (2009), p. 14). On the other hand, the view is also supposed to be beneficial to Bayesianism, by filling in gaps in the Bayesian framework.

2.5.1 Explanatory considerations as a constraint on Bayesian probabilities

All versions of the composite view have in common the proposal that explanatory considerations play a role in assigning the Bayesian probabilities that enter into conditionalisation. They differ however in their claims about exactly which explanatory considerations constrain which probabilities. I will now briefly outline

Fraassen (1989), pp. 166-169). If one is not motivated to preserve the particular philosophical application of IBE to the realism argument, one is free to adopt a different interpretation of IBE than van Fraassen does.

the different approaches.

On Jonathan Weisberg's view, the priors and likelihoods are constrained by explanatory considerations via constraints on the conditional probabilities $p(T|D)$. These probabilities $p(T|D)$ are constrained to agree with the explanationist ranking of hypotheses in the light of evidence D. If T provides a better explanation of D than T', then $p(T|D) > p(T'|D)$. Weisberg says that explanationist thinking should fix these probabilities 'either in conjunction with, or in place of existing objectivist principles. Ideally, explanationist considerations would complement existing objectivist principles like the Principle of Indifference' (p. 14). He gives the following example of how explanatory constraints could go together with existing objective Bayesian constraints. He suggests that a theory's ability to systematize the evidence should be seen as an explanatory virtue. Suppose a theory T provides a better systematization of certain evidence D than another theory T'. If more systematized theories are preferred as more explanatory, this constrains the conditional probabilities such that $p(T|D) > p(T'|D)$. If the likelihoods are taken to be fixed by the Principal Principle, this inequality can be seen as imposing a certain constraint on the priors $p(T)$ and $p(T')$.

On Lipton's view, people actually employ explanatory considerations in order to determine what priors and likelihoods to assign.[7] He suggests that explanatory virtues, including simplicity, unification and scope 'guide' the assignment of the prior probability. He also suggests that 'although likelihood is not to be equated with loveliness, it might yet be that one way we judge how likely D is on H is by considering how well H would explain D'. He notes that this would require

[7]Lipton puts his claim in this way because he thinks that it is possible to regard theories such as Bayesianism and IBE as having a descriptive function which is distinct from their normative role. The descriptive function is to describe what people actually do when they make inferences. Lipton restricts his claims about the relationship between IBE and Bayesianism to the descriptive domain.

that loveliness is 'reasonably well correlated' with likelihood, so that we could 'use judgements of loveliness as a barometer of likelihood'. Specifically, he suggests that considering the mechanism linking cause and effect in a potential causal explanation may help us to form a judgment of how likely the cause would make the effect (Lipton (2004), p. 114).

Samir Okasha's account of how explanatory considerations play a role in assigning the probabilities is given with the aid of the following example:

> 'A mother takes her five-year old child to the doctor. The child is obviously in some distress. On the basis of the mother's information, the doctor forms two competing hypotheses: that the child has pulled a muscle, and that he has torn a ligament; call these H_1 and H_2 respectively. A keen advocate of IBE, the doctor examines the child carefully, and decides that H_2 offers the better explanation of the observed symptoms ...' (Okasha (2000), p. 703)

Okasha suggests that the doctor might justify her reasoning by appeal to two considerations. First, that 'preadolescent children very rarely pull muscles, but often tear ligaments'. And second, that the symptoms, 'though compatible with either diagnosis, are exactly what we would expect if the child has torn a ligament, though not if he has pulled a muscle'. Okasha claims that the first consideration is an aid to determining the prior probabilities. The prior probability of H_2 is higher than that of H_1 because H_2 is the more initially plausible hypothesis. The second consideration is a way to determine the likelihood. The likelihood of H_2 is also greater than that of H_1, because H_2 makes the observed evidence more expected than H_1.

In claiming that explanatory considerations determine the assignment of Bayesian probabilities, proponents of the composite view essentially treat Bayesianism as an

inductive framework which can be filled in to represent other forms of inference. If a minimalistic version of Bayesianism is regarded as an inductive framework, then it can be constrained by explanatory considerations in such a way as to yield a Bayesian representation of IBE. This is trivial in the sense that any form of inductive inference resulting in a probability distribution, including Inference to the Worst Explanation, can be represented by the Bayesian framework by appropriating setting constraints on the probabilities.

The general problem with the composite view is the lack of independent motivation for using explanatory considerations as constraints on priors and likelihoods rather than any other type of constraints. In fact, the idea that explanatory considerations play the role of constraining probabilities is actually quite unconvincing when it comes to the prior probabilities. Most explanatory considerations, including virtues like unification and simplicity, are not just properties of the hypothesis T, but of the relationship between T and the data D. Priors, on the other hand, are supposed to represent the epistemic situation before the new data D is taken into account. This makes it difficult to see how Lipton can be right that explanatory virtues like simplicity and unification are a guide to the priors.

One could, like Weisberg, argue that the priors are constrained indirectly via constraints on the conditional probabilities $p(T|D)$. As we saw, Weisberg's idea is that $p(T|D)$ is constrained to agree with the explanationist ranking on hypotheses, and this then sets some constraints on the priors and likelihoods. The story cannot be quite this simple, however, since one should be able to assign the prior before knowing which way the data will turn out. It could well be that different evidence D' would lead to different constraints on $p(T|D')$. It could be, for instance, that relative to D, T is a better systematization than T', but the reverse is true with respect to evidence D'. One could, of course, hypothetically consider each possible

way that the evidence might turn out, and then find prior probabilities which will agree with the explanationist assessment no matter which evidence they are conditionalised on. However, it is no longer clear that this procedure of considering all the possible outcomes, their explanatory consequences and then backtracking to find initial probabilities which will work for all outcomes is either particularly easy for the 'inquirer on the street' or to be recommended as a normative procedure.[8]

Notice however, that this difficulty concerning the priors does not arise for Okasha because the explanatory consideration that he takes to determine the prior is the non-relational notion of plausibility of the hypothesis alone.

2.5.2 The context of discovery

In addition to claiming that explanatory considerations help to assign Bayesian probabilities, Lipton and Okasha also claim that explanatory considerations are able to help the Bayesian by saying something about how scientific theories are generated – the so-called 'context of discovery'. The process of inferring hypotheses is often taken to involve two stages: generating candidate hypotheses to put in a 'short-list' and then selecting from them according to some criteria. Philosophers of science have traditionally been inclined to leave the first stage relatively uncharacterised, regarding it as the domain for scientific creativity, rather than evaluative principles. Bayesianism, for example, has generally been thought of as purely defining the selective criteria in the second stage. Lipton and Okasha suggest that one advantage that IBE has over other methods of inference is that it does have something to say about the first stage as well as the second. Thus, this provides another way in which IBE may allegedly supplement the Bayesian framework. Lipton and Okasha

[8]Lipton suggests that explanatory considerations are 'more accessible ..to the inquirer in the street' than Bayesian probabilities (Lipton (2004), p. 114).

again have slightly different proposals concerning the details.

For Okasha, IBE allows us to explain why new hypotheses are sometimes invented. He says 'In those cases where agents respond to new evidence by inventing new hypotheses, the Bayesian model is silent. But IBE provides a useful, if schematic account of what is going on: the agents are trying to explain the new evidence' (Okasha (2000), p. 707).

Lipton, on the other hand, draws an analogy with biology. There are two processes involved in evolution of particular biological traits: the process of generating variants via mutations and the process of eliminating some of these variants by natural selection. It might seem that these processes of generation and selection are fundamentally different from one another. However, Lipton points out that in the process of generation, the mutations occur in the reproduction of creatures which were already subject to natural selection. Thus, selection may be said to play a role in the generation of variants, as well as in selection amongst them. Similarly, Lipton claims, IBE figures in the process by which scientists determine which hypotheses to put in their short-list of candidates, since the candidates to consider are those which seem plausible according to background beliefs and those background beliefs are themselves selected by IBE. Thus the plausibility judgments that determine the candidates depend implicitly on explanatory considerations – namely, how well the background beliefs explained the past data.

I will now argue that Bayesianism is less silent in the context of discovery than Lipton and Okasha have suggested, and in fact offers an account which is remarkably parallel to what IBE is supposed to provide.

Okasha claims that IBE can give a reason for the invention of new hypotheses, whereas Bayesianism has nothing to say. The explanationist will say that scientists invent a new hypothesis because they think the new one will give a better

explanation. However, a Bayesian could just as well say that scientists invent new hypotheses because they think that there may be a novel hypothesis which would have better support (higher posterior) than the other hypotheses, if the hypothesis space were enlarged to include it.

Lipton's account of the origin of the hypothesis space is also sufficiently general that it is not specific to IBE. As his analogy to natural selection makes clear, the same point can be made for any two stage process of generation and selection where candidates for selection are created by variation on entities which were selected in the past. Then it can be said that the production of candidates for the short-list involves the criteria of selection. But this explanation is not sensitive to specific details concerning the criteria of selection. The selection process could be natural selection, or it could be selection according to IBE or Bayesianism. Lipton takes background beliefs, themselves the result of explanatorily based selection, to serve as 'heuristics' which restrict the range of actual candidates. In other words, candidates must be plausible according to the background beliefs. We have seen a parallel story based on hierarchical Bayesian models in Chapter 1. There the candidates must also be plausible (have reasonably high prior probability) with respect to the higher level theory which generates them. And higher level theories are themselves selected by Bayesian updating. Again the Bayesian has been able to capture the same insights concerning the process of discovery as the explanationist.

The arguments presented by Okasha and Lipton do not succeed in showing that in the context of discovery IBE fills in where Bayesianism has nothing to say. Rather the Bayesian has just as much to say as the explanationist, since the same type of account can be given either in terms of IBE or of Bayesianism.

2.5.3 Summary

Overall, then, the composite view is based on the idea that there are certain 'gaps' in the Bayesian account of inference which IBE may fill. This view is only plausible to the extent that a) there really are gaps to be filled, and b) it is explanatory considerations which should fill them. There have been two proposals about what the gaps are. One is the assignment of probabilities which enter into Bayesian conditionalisation. This is only genuinely a gap if one espouses something close to Min Bayes. More objective Bayesians have their own proposals about how to fill this gap. I have argued that proponents of the composite view have not provided us with independent grounds to fill this gap, if there is one, with constraints on probabilities based on explanatory considerations. I have suggested that explanatory considerations (with the exception of the plausibility of the hypothesis) are actually the wrong kind of thing to constrain the priors.

The composite view also proposed a gap in the context of discovery. Here I argued that there is no gap, and Bayesianism has something to say in this domain which is actually not dissimilar to what the explanationist proposes.

I conclude then that the composite view is not a very convincing account of the relationship between IBE and Bayesianism.

2.6 Seeds of another approach

I now want to suggest that there is more to Okasha's account than I have so far indicated. Some of what he says suggests that he sees himself as adopting quite a different strategy, which is essentially the approach I will recommend in the following sections. That is to provide an independently motivated analysis of the concept of IBE and then show how its different components may be represented in

probabilistic terms.

Okasha's analysis of IBE is quite simple. He factors goodness of explanation into two components. One is the plausibility of the hypothesis providing the explanation, and the other is how expected the explanandum is, given the hypothesis. The first of these components is represented by the Bayesian prior probability of the hypothesis and the second is represented by the likelihood of the hypothesis, given the evidence.

For simple cases such as Okasha's doctor example, this two component model of IBE is quite convincing. It seems right that explanatory thinking involves not just finding a 'lovely' explanation, but must also take into account a relatively independent consideration of plausibility of the hypothesis, which acts as a kind of 'sanity check' on the explanation. A conspiracy theory might, if true, provide a great deal of understanding but be also wildly far-fetched, so that no-one would sensibly assess it to be the best explanation. It also seems right that the theory should make the explanandum expected. Suppose, in the doctor case, that the symptoms were only somewhat, rather than exactly, to be expected, if the ligament hypothesis were true. Then, intuitively, we would regard the ligament hypothesis as providing a less good explanation. The explanatory relation between a certain theory T and evidence D is generally regarded as based on deductive or causal relationships. This is made explicit in various specific models of explanation such as the Deductive-Nomological (DN) model, according to which an explanation is a deductive argument containing laws which leads to the explanandum (Hempel and Oppenheim (1948)), or the Causal-Mechanical (CM) model, according to which an explanation provides information about the causal history of the explanandum (Salmon (1984)). But at the same time, it is recognised in discussions of explanation that such deductive or causal links should help to make D more expected, given T. For example, Hempel says 'a DN explanation answers the question "*Why* did the

explanandum-phenomenon occur?" by showing that the phenomenon resulted from certain particular circumstances ... in accordance with the laws ... By pointing this out, the argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*' (Hempel (1965), p. 337).

The problem with Okasha's proposal is that, in moving beyond simple examples, the two-component analysis of IBE does not seem to capture all its relevant aspects. Okasha attempts to suggest that it does by referring to the second component of explanatory thinking in more general terms as an 'appropriate relation' between the explanans and explanandum, and linking it to Lipton's 'loveliness'. However, there seems to be more to these concepts than just making the explanandum expected. Loveliness involves how intelligible the hypothesis makes the explanandum, not just how expected. In fact, we may feel that the two-component model is too thin to capture the richness of the notion of IBE. I suggest that we should go further in analysing what other explanatory considerations are involved in IBE in the hope of uncovering how they may be represented in probabilistic terms. This is the task that I will address in the next section.

2.7 Further analysis of IBE

What more is there to the concept of IBE? As we saw in section 2.2, IBE is often spelled out in terms of the virtues view of better explanation. That is, an explanation is better than another if it has a better combination of 'explanatory virtues'. The virtues view is somewhat promising for making a connection with Bayesianism, since several of the key virtues have already been analysed in Bayesian terms. Simplicity is associated with the likelihood in Rosenkrantz (1977), Jefferys and

Berger (1992) and MacKay (2003) (see also section 1.4), and a Bayesian account of unification is given by Myrvold (2003).

However, the virtues view is still relatively undeveloped. There has been little work, for example, on how the virtues should be traded off against one another, and it remains unclear whether some of the virtues are more fundamental than others, or even whether some reduce to essentially the same concept. In addition, in the IBE literature, there has been a certain desire to keep the account of IBE independent of the particular model of explanation which is employed.[9] This has led to a tendency to treat what makes one explanation better than another as a completely separate matter from what makes them potential explanations at all. This independence from the basic notion of explanation, though perhaps convenient in allowing one to avoid the thickets of the explanation literature, is in some respects a weakness, since it makes it hard to see how the view is truly to do with explanation per se. It makes it difficult to distinguish IBE from accounts which characterise theory assessment in terms of optimally combining 'inferential virtues' of theories, where these virtues are not regarded as particularly tied to explanatoriness, (Kuhn (1977), McMullin (2008)). The lists of virtues may be quite similar on these accounts. For example, Thomas Kuhn lists 'accuracy, consistency, scope, simplicity and fruitfulness' as inferential virtues (Kuhn (1977), p. 322), which is not unlike a list of 'explanatory' virtues.

I suggest therefore that the notion of better explanation should be more closely connected to the basic notion of explanation itself. In fact, I will argue that a comparative notion of better explanation can be derived from the notion of what it means for a theory to explain at all. It is not at first sight obvious how to connect the

[9]Lipton, for example, distinguishes between specific models of explanation and the 'actual explanation relation itself, whatever its correct description turns out to be' (Lipton (2004), p. 57), suggesting to concentrate on the latter.

notion of better explanation to basic ideas about explanation. There are a number of accounts of what an explanation is. Some of the main candidates are the DN model (Hempel and Oppenheim (1948)), the CM model (Salmon (1984)), and the unificationist model according to which explanation consists of providing a unified account of various phenomena (Friedman (1974), Kitcher (1989)). These accounts were generally intended as criteria for what an explanation consists of, perhaps even to provide necessary and/or sufficient conditions for explanation, rather than being particularly amenable to coming in degrees.[10] However, I will suggest in the next section that there is a way of distinguishing two broad approaches to explanation which is helpful in seeing our way through the multiplicity of accounts.[11] Then in section 2.7.2, I will propose an account of the explanation relation and a derivative comparative concept of better explanation, which attempts to combine insights from these two approaches.

2.7.1 Accounts of explanation

One approach to explanation is based on the idea that what is distinctive about the explanatory relation is that it connects the explanandum to a hypothesis or hypotheses with some kind of special status. There are several versions of this type of view, depending on what the special status is taken to be. One simple version is that explanations relate the explanandum to elements with which we are so familiar that we do not question them. For example, the kinetic theory of gases might derive its explanatory power from relating an unfamiliar explanandum, such

[10] Although it seems natural that the ability to unify phenomena could come in degrees, as Woodward (2003) points out, unificationists have in fact adopted a 'winner-takes-all' approach according to which only the most unified theory is explanatory at all.

[11] Something close to this distinction is made in Friedman (1974).

as the Boyle-Charles law, to a familiar one, such as the movement of billiard balls.[12]

However, the notion of familiarity involved in scientific explanation cannot just be everyday familiarity with ordinary macroscopic objects, since often phenomena are explained in terms of rather arcane principles of theories. Rather, it may be the familiarity of a scientist, trained in a particular paradigm, who attempts to relate novel and unexplained phenomena to certain exemplary and understood cases which are taken to be canonical. Such a view is expressed by Stephen Toulmin, who takes scientific explanation to involve explaining phenomena in terms of certain 'ideals of natural order', which scientists treat as self-explanatory (Toulmin (1963)). He gives the example of Copernicus' 'principle of regularity' that all the heavenly bodies should move uniformly in a circle, because that is thought to be their natural form of motion. Copernicus tried to explain planetary motions in terms of this principle and he treated it as self-explanatory, or absurd to deny (Copernicus (1939), p. 57).

Another version of the special status view is given by Clark Glymour, who suggests that a very common pattern of explanation in science is to explain something by showing that it follows from a necessary truth (Glymour (1980a), Glymour (1985)).[13] There are no further scientific questions to be asked about why a mathematical truth holds, so the explanation is regarded as complete and satisfying.

The Deductive-Nomological model of explanation may also be seen as a species of special status view. There the explanandum is derived, not from mathematical necessities, but from laws. According to this view, laws have a special status in explanation not because they are natural stopping points for explanation. Rather

[12]This example is given by Friedman (1974), p. 9.

[13]For example, the observed relation between synodic and sidereal periods expressed by equation 2.3 is a mathematical truth for any two objects that move in closed orbits about a common centre with the inner object moving faster than the outer. Glymour claims that the Copernican theory provides a good explanation of the observed regularity because it reduces the regularity to a mathematical truth.

if an observation such as the apparent bending of a stick underwater is derived from a general law, such as the law of refraction, one can go on to ask why the law of refraction holds. The answer will be that it follows from an even more general law, such as those of the wave theory of light (Hempel and Oppenheim (1948), p. 136).

All these accounts of the explanatory relation have in common that they characterise explanation as a relationship between the explanandum and some hypothesis which has some special status – psychological self-explanatoriness, mathematical necessity, or lawlike nature.

A second approach to characterising the explanatory relation is to focus on finding some essential property of the relation itself. The main suggestion here has been that what is essential to the explanatory relation is that it unifies different phenomena. There are several attempts to characterise explanatory unification, due to Friedman (1974) and Kitcher (1989). Friedman says it is essentially a reduction in the 'total number of independent phenomena that we have to accept as ultimate or given' (Friedman (1974), p. 15). For example, he says, kinetic theory is an explanation of phenomena involving the behaviour of gases, such as the Boyle-Charles law because it 'also permits us to derive other phenomena involving the behaviour of gases, such as the fact that they obey Graham's law of diffusion and (within certain limits) that they have the specific heat-capacities that they do have, from the same laws of mechanics' (p. 14). Where previously there were three 'independent brute facts', now there are only the laws of mechanics to accept. Kitcher has raised difficulties concerning how such independent facts can really be counted (Kitcher (1976)). He stresses rather the idea that unification is achieved by using similar arguments in deriving different phenomena (Kitcher (1989)). Arguments are taken to be similar insofar as they instantiate the same 'argument pattern', which is a kind of schematic form or type of argument. Thus unification, on Kitcher's view,

is based on repeatedly using a relatively small number of types of argument.

On these unificationist views, there does not need to be any special epistemic status, in terms of familiarity, paradigm, necessity or lawlikeness to the types of argument or basic laws to which the explanandum is related.[14] They just have to be unifying, and the underlying intuition is that they have the power to generate more from less. They explain a variety of phenomena in terms of relatively few core laws or patterns of argument.[15]

2.7.2 Explanation from the core

I now propose an account which may be seen as combining aspects of the special status and unificationist approaches to explanation. On this account, as in the special status approach, explanation does involve connecting the explanandum to some particular kind of hypotheses. These hypotheses comprise the part of the explaining theory which are most central to the theory. They are its 'core'.

The identification of core aspects of theories should be seen as a fundamental component of scientific theorising. The typical presentation of a scientific theory involves setting forth the elements that the author regards as its basic components. Consider for example, Newton's 'axioms or laws of motion' at the outset of the *Principia* (Newton (1687)), the principle of relativity and the light postulate in Einstein's theory of special relativity (Einstein (1905)), or Copernicus' seven assumptions about the planets (Copernicus (1939)). The identification of a theory's core principles should not be seen as solely the responsibility of the original author of the theory. Once a theory is proposed, certain scientists, usually regarded as

[14] Friedman (1974) makes this point, p. 18.

[15] This is a key intuition behind the idea of explanation as information compression (Rosenkrantz (1977), Chapter 8).

towards the more theoretical or foundational end of the spectrum, tend to concentrate their efforts on producing reformulations of theories, and 'experimenting' with new reaxiomatisations.

The core principles of a theory are not generally put forward as such because they are regarded as particularly familiar or self-explanatory from the outset, though that would perhaps be thought an advantage if it were true. Copernicus' assumptions include the claim that 'All the spheres revolve about the sun as their mid-point, and therefore the sun is the centre of the universe' (Copernicus (1939), p. 58), which was certainly not a familiar thought at the time. Nor need the core aspects of a theory be seen as in no need of explanation themselves. For example, towards the end of Principia, Newton says 'Thus far I have explained the phenomena of the heavens and of our sea by the force of gravity, but I have not yet assigned a cause to gravity' (Newton (1687), p. 943). Thus he acknowledges the possibility of an explanation of one of his core postulates, though admitting he does not currently have one.

If a theory becomes accepted and entrenched due to its confirmatory and explanatory success, its core principles may come to seem natural to scientists accustomed to working with the theory. They may come to seem familiar, through habituation, and a tendency may develop to treat the core of the dominant theory as an explanatory stopping point. Explanations are often regarded as providing understanding by relating phenomena to things which are familiar, and as satisfying if they connect the explanandum to things which are taken to require no further explanation. Once a theory becomes entrenched, perhaps as a part of a reigning paradigm, its core components may come to take on these psychological roles.

Components of the core are very often given the title of 'law'. However, I wish to remain non-committal about whether elements of the core have any necessity of a

metaphysical nature. Thus, although like special status views, I characterise explanation as a relation between the explanandum and a distinctive type of hypothesis, I do not regard special familiarity, or necessity, or lawfulness, as the distinguishing feature of that type.

This account also shares some insights with the unificationist view. If one connects various phenomena to certain core principles of a theory, this results in unification of the phenomena, which are all now explained in terms of a small set of core elements. Is the core then just that which produces unification? Certainly, scientists often justify their choice of core principles according to how well they bring together or explain different phenomena. Copernicus presents his assumptions as reasonable because they enable him to solve the problem of the planets 'with fewer and much simpler constructions than were formerly used' (Copernicus (1939), p. 58). Newton justifies his use of the force of gravity by its explanatory effectiveness – the fact that it is 'sufficient to explain all the motions of the heavenly bodies and of our sea' (ibid, p. 943). Part of the point of reformulating theories by giving them different foundations is to find ways of casting the theory which will be more explanatory in the sense of unifying more phenomena.

However, the ultimate justification for any identification of the core of a theory is how well it works, not just in terms of how well it facilitates explanation of different phenomena, but also in terms of how well it is confirmed by the evidence. As we will see in section 2.8, this is the basis for our connection with Bayesianism.

So far, I have talked about the nature of the explanatory relation and suggested that it involves connecting the explanandum to the core of a theory. How can we turn this into a comparative concept of one theory explaining phenomena *better* than another theory? My proposal is this: a theory provides a *better explanation* of a certain phenomenon than another theory, the more it relates that phenomenon

to its core. This means that the theory does not accomplish explanation of the phenomenon by utilising 'auxiliary hypotheses' which are brought in just for the purpose of aiding the explanation of particular phenomena.

I will attempt to illustrate and justify this proposal with two case studies. The first is the comparison of how well the Copernican and Ptolemaic theories, which were rivals in the 16th century, explained observations of planetary positions in the sky. The second is the biological question of whether altruistic traits are better explained by invoking natural selection at the level of groups, as opposed to natural selection on individual organisms alone. These case studies are more complex than those typically discussed in the literature on IBE and its relation to Bayesianism. This will help us to get a richer picture of what is involved in IBE.

2.7.3 Copernicus vs Ptolemy

Both the Copernican and the Ptolemaic theories were capable of predicting and explaining to some extent observations of planetary positions in the sky. The Copernican system did this by placing the sun at the centre of the planetary orbits, whereas the Ptolemaic system placed the earth at the centre. Since the predictive advantage of the Copernican theory was not obvious at the time, Copernicus and his followers stressed what we might now regard as the explanatory advantages of his theory. They emphasised the 'harmony' and 'elegance' of the Copernican explanation, and its ability to unify all the phenomena under a common framework. For example, Rheticus says:

> With regard to the apparent motions of the sun and moon, it is perhaps possible to deny what is said about the motion of the earth, although I do not see how the explanation of precession is to be trans-

> ferred to the sphere of the stars. But if anyone desires to look either to the principle end of astronomy and the order and harmony of the system of the spheres or to ease and elegance and a complete explanation of the causes of the phenomena, by the assumption of no other hypotheses will he demonstrate the apparent motions of the remaining planets more neatly and correctly. For all these phenomena appear to be linked most nobly together, as by a golden chain; and each of the planets, by its position and order and every inequality of its motion, bears witness that the earth moves... (Rheticus (1539), pp.164-165)[16]

The explanatory merits of the Copernican and Ptolemaic theories can be illustrated with respect to the following phenomena. Every night a visible planet appears in a slightly different position in the sky than it did the night before. Relative to the stars, the planet generally moves in an eastward direction, though every now and again it reverses direction and moves westward for a while before moving eastwards again. This phenomenon of doubling back is called 'retrograde motion'. In quantitative terms, the motion of a planet can be measured either with respect to the stars, or with respect to the sun. The time it takes for a planet to return to the same longitude with respect to the stars is called its 'sidereal period'. The time it takes to return to the same position with respect to the sun is called its 'synodic period'. For example, if it is a 'superior' planet, that is, one which is further from the sun than the earth, it will at times be observed to be at an angle of 180^0 from the sun. Then it is said to be 'in opposition' to the sun. The synodic period is the time between successive oppositions. 'Inferior' planets, which are closer to the sun than the earth, are never observed in opposition to the sun, but they can be 'in

[16]I thank Jonah Schupbach for drawing my attention to this quotation in the context of the relationship between the virtues of simplicity and unification.

conjunction' with it – that is, appearing at the same angle as the sun, either in front of it (inferior conjunction), or behind it (superior conjunction). The synodic period for inferior planets can be measured as the time between successive conjunctions of the same type.

Consider first how retrograde motion is explained by the Ptolemaic theory. In this context, retrograde motion is explained by having the planet move in a small circle, or 'epicycle', whose centre moves in a larger circle (the 'deferent'), in the same direction around the earth. When traversing the epicycle, at some points the planet is moving backwards with respect to the deferent - when this happens, according to the theory, retrograde motion occurs (see Figure 2-1).

Retrograde motion is explained quite differently by the Copernican theory. At certain times planets may overtake each other in their orbits around the sun. When, for example, the earth overtakes a planet which is further from the sun, the planet will appear to be moving 'backward' for a certain amount of time: this is the retrograde motion (see Figure 2-2).

According to the Copernican theory, there is a close relationship between the sidereal and synodic periods of each planet. Consider the motion of a superior planet, which we assume moves more slowly around the sun than the earth. Let the time between successive oppositions (the synodic period) be S, and let the orbital period of the earth around the sun be D and the orbital period of the planet be P. At opposition, the earth, sun and planet all lie on a line. The planet will again be in opposition when it is 'overtaken' by the earth at a time when the earth has completed one more orbit around the sun than the planet has. The planet is moving at an angular velocity of $\frac{360}{P}$ and the earth is moving at angular velocity $\frac{360}{E}$. When opposition occurs again, the planet has moved through an angle $\frac{360}{P}S$, and this is equal to the angle through which the earth has moved $\frac{360}{E}S$ minus 360^0 (see Figure

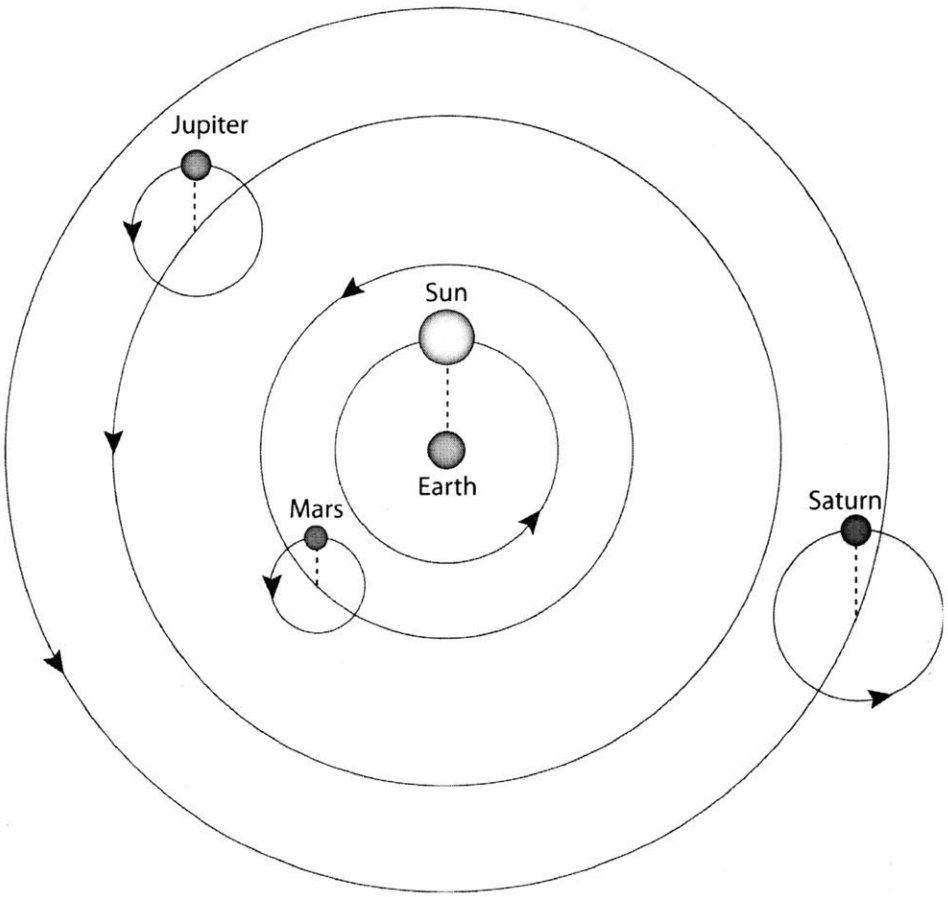

Figure 2-1: In the Ptolemaic theory, retrograde motion of the planets is explained by epicycles.

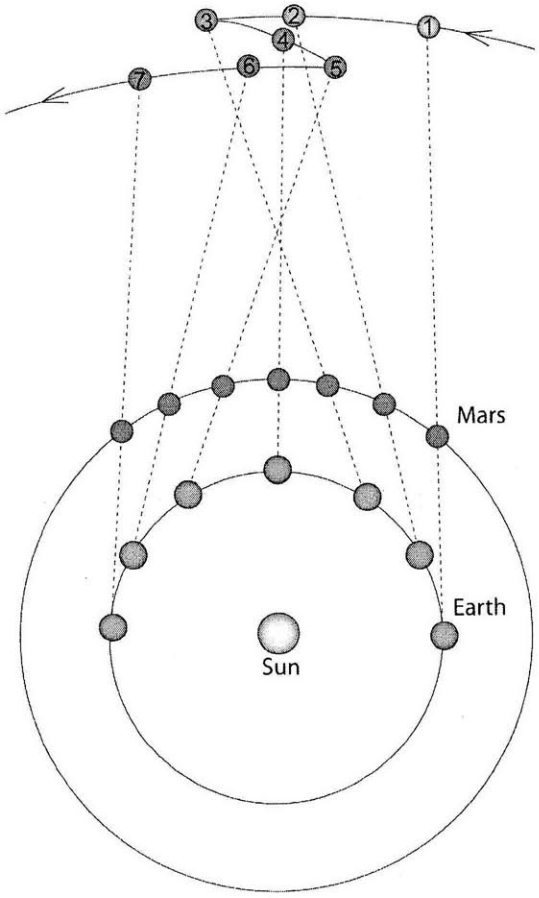

Figure 2-2: In the Copernican theory, the explanation of the observation of retrograde motion is that the outer planet is overtaken by the faster moving inner planet.

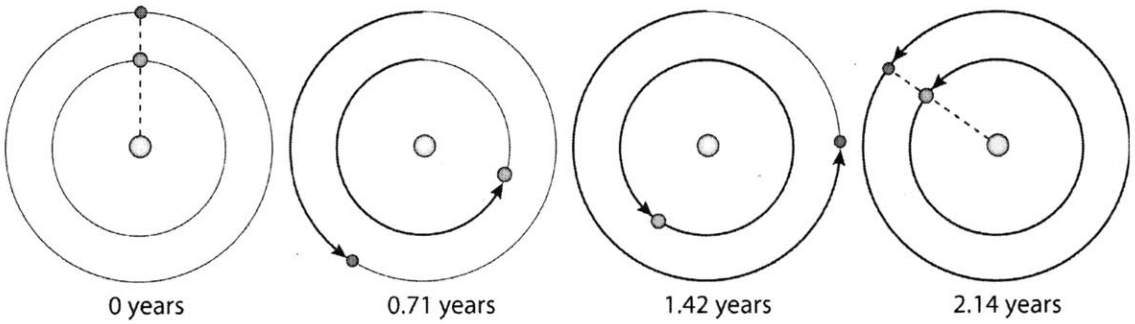

Figure 2-3: When 2.14 years have elapsed since the last opposition, Earth and Mars will be in opposition again. In the meantime, Earth has completed one more complete orbit around the sun than Mars.

2-3). Equating these angles yields the following relation:

$$\frac{1}{P} = \frac{1}{E} - \frac{1}{S} \tag{2.3}$$

(For inferior planets, similar considerations yield the relation: $\frac{1}{P} = \frac{1}{E} + \frac{1}{S}$).

What this means is that in the Copernican framework, once the synodic periods of the planets are given one can calculate the sidereal periods, or vice versa. We do not need independently adjustable parameters for sidereal and synodic periods. In the Ptolemaic system, on the other hand, the synodic period is given by the period of the epicycle. This can be adjusted independently of the period of the deferent, which corresponds to the sidereal period. Also, evidence from one planet has implications for other planets, which it does not on the Ptolemaic view.

It is essential to the Copernican theory, part of its core, that the planets orbit around the sun, though there may be many ways to build Copernican models by giving planets different sizes of orbits or different speeds. The Ptolemaic theory makes a different core assumption that the earth is at the centre of the planetary

orbits. Epicycles are not part of the fundamentals of either theory. Hypotheses about epicycles are auxiliary assumptions which allow the theories to provide models which fit the observations well.

In the Copernican explanation, the observation of retrograde motion appears as a consequence of placing the sun at the centre. Then the earth and planets are both orbiting the sun and it is inevitable that there will be overtaking of one planet by another as long as the planets are not moving at exactly the same angular velocity. In fact, Copernicus makes the auxiliary assumption that planets closer to the sun move more rapidly than planets further away.

In the Ptolemaic case, on the other hand, retrograde motion is not accounted for by the core of the theory. Rather the explanation relies on auxiliary hypotheses about epicycles. We can say then that the Copernican theory provides a better explanation of retrograde motion because it explains it in terms of its core principles, whereas the Ptolemaic explanation rests heavily on hypotheses which are not part of its core.

2.7.4 Group vs individual selection

Another example concerns how to give the best explanation of the existence or prevalence of altruistic traits in biology. Altruistic traits are those which are good for the group to which the organism belongs, but detrimental to the organism possessing them. There are many cases in biology, including signalling to warn others of threats, exercising restraint on one's own reproduction, caring for others' young and self sacrifice in defence of the group. Generally, the prevalence of certain traits in biology is explained by invoking natural selection. The trait is prevalent because organisms without it were less fit and were therefore selected against. Natural selection favoured those who possessed the trait.

The problem is that this appeal to the selection of individual organisms does not, at least at first sight, seem to explain the prevalence of altruistic traits. This is because the possession of altruistic traits bestows a fitness disadvantage, relative to other organisms in the community, and therefore altruistic organisms should have been selected against. Some biologists have proposed that altruistic traits persist in populations because natural selection is operative not just at the level of individual organisms, but also at the level of groups of organisms. That is, different groups may have different fitness from one another and may be more fit the more their members cooperate with one another and engage in altruistic behaviour. Groups, like organisms, may reproduce themselves by splitting into 'colonies', and these colonies will tend to inherit traits from the founder group, particularly if these are determined by the levels of altruistic behaviour within the group. However, reproduction of groups, like that of individuals, also produces some variation. Colonies are not necessarily exactly like their founders. Some by chance might get a greater share of altruistically inclined individuals than others, which would make them more fit. Then fitter groups are selected. This will tend to favour altruistic behaviour since groups with greater altruistic cooperation are generally fitter than groups without it.

For example, this type of explanation can be given for a specific altruistic trait such as the defensive behaviour of musk oxen. This is an example given by Williams (1966), (pp. 218-220), and discussed by Sober (1990). When attacked by wolves, musk oxen form a circle with the males on the outside and the females and young on the inside (this is called 'wagon-training'). It appears that stronger members of a group protect the weaker members from attack, even if they are unrelated to them (see Figure 2-4). An explanation in terms of group selection would say that groups of musk oxen where this trait exists are selected over other groups where it

does not. Thus the reason for the existence of the trait in the male musk ox is its benefit to the group rather than to the ox himself.

On the other hand, some biologists have resisted invoking group level selection to explain phenomena such as these, and attempt to explain them with reference only to natural selection of individual organisms. For example, Williams attempts to explain a number of apparent examples of altruism in terms of individual selection alone (Williams (1966)). He offers the following alternative explanation of musk oxen wagon-training. The defensive formation is the result of a 'statistical effect'. What is really going on is that the threat felt by an ox depends on its size relative to a predator. There is some threshold of predator size which determines whether the ox responds with 'counterthreat or with flight'. For predators of a certain size, larger oxen will be more inclined to stand their ground than to flee and the result is that they will end up in a more exposed position, seemingly protecting weaker members of the herd. He suggests that this case is one of a number of others which may also be explained in terms of different adaptive responses of individuals according to their own strengths.

In this case the explanatory advantage of one theory over the other is less obvious than it was in the Copernican example and the case is much more controversial. Nonetheless, we can diagnose what makes the choice difficult in terms of the account of better explanation delineated above. We need to compare the way in which each explanation relies on core aspects of the theory as opposed to auxiliary hypotheses. Where both explanations invoke auxiliary hypotheses of quite different natures, it can be hard to assess which has the greater reliance on auxiliary hypotheses.

Williams gives a parsimony argument against group selection which may be seen as stemming from his concern that the group selection hypothesis depends on too many auxiliary hypotheses in its explanation of altruistic behaviour. Williams

Figure 2-4: Musk oxen in defensive formation. (Credit: U.S. Fish and Wildlife Service, Adams)

sees group selection as a possible evolutionary process, but one which is rather 'onerous', and which should only be invoked when 'simpler' (genic or individual) forms of natural selection failed to give an explanation. Part of the reason for the perceived onerousness is that group selection can only account for altruistic behaviour under rather particular circumstances. Selection of one group over another is always opposed by selection of individuals within each group. Within the group, selfish individuals will generally have an advantage over altruists. With no other constraints, they would eventually take over, and eliminate altruists from the population. The process of group selection therefore has to be fast enough to occur before individual selection within groups gets a chance to dominate. This imposes a number of constraints on acceptable rates of group reproduction and other parameters. More detailed study of mathematical models of group selection has led to the conclusion that 'although group selection is possible, it cannot override the effects of individual selection within populations except for a highly restricted set of parameter values' (Wade (1978), p. 101). Thus the fact that the explanation of altruistic traits in terms of group selection depends heavily on the particular auxiliary hypotheses about these parameters forms the basis for regarding it as a less good explanation.

On the other hand, the explanation in terms of individual selection offered by Williams is subject to the same kinds of objections. At first sight, it would appear that there could be ways that a male musk ox could protect himself from the wolves at the expense of the group. For example, he could make a run for it, or he could actually stand behind the young, so making it more likely that the wolves will go for them rather than himself. For the individual selection theory to provide an adequate explanation, we have to make auxiliary hypotheses about what it is in the individual musk ox's best interest to do. In particular we have to assume that it is

better for a musk ox to stand in defensive formation than to adopt any of the other possible group-sacrificing strategies.

Determining which of the two explanations is better is not so easy in this case, but it seems clear that intuitively the factors to be taken into account involve assessing the way in which the explanation depends on auxiliary assumptions.

One reason why an explanation is better if it explains the explanandum more in terms of the core of the theory rather than auxiliary hypotheses is that it promotes a certain desirable property of explanations, namely robustness or modal stability. An explanation is more stable, and hence better, if slight changes to the conditions it asserts would not have destroyed its explanatory qualities. If the explanation of a fact depends on the conjunction of a large number of auxiliary assumptions, then it is too finely tuned to fit the particular facts it aims to explain. For example, in Copernicus' theory, it is to be expected that retrograde motion of a superior planet occurs when the planet appears in opposition to the sun; it is because overtaking happens when the planet, sun and earth lie on the same line (see Figure 2-3). The observation that retrograde motion coincides with opposition can be reproduced in Ptolemy's theory, by adjusting the epicycle rates in such a way that a line from the planet to the centre of its epicycle is always parallel to the line from the earth to the sun (see Figure 2-1). However, changes to the epicycle rates would mean that retrogression and opposition no longer coincided. Thus, Ptolemy's theory has been finely tuned to fit phenomena which arise as a more natural consequence of Copernicus' theory.

2.7.5 Simplicity and unification

I have suggested that explanation should connect the explanandum to the core of the theory, and a theory provides a better explanation to the extent that it is possible to

give an explanation in terms of its core, rather than relying on auxiliary hypotheses. This feature of minimal dependence on auxiliaries appears to be what is often meant by 'simplicity' in science. Thagard (1978) proposes this explicitly. He says 'the explanation of facts F by a theory T requires a set of given conditions C and also a set of auxiliary hypotheses A'. Conditions C are accepted independently of T or F, whereas 'an *auxiliary hypothesis* is a statement, not part of the original theory, which is assumed in order to help explain one element of F, or a small fraction of the elements of F'. Thagard proposes that 'simplicity is a function of the size and nature of the set A needed by a theory T to explain facts F'. He gives several scientific cases which seem to support this analysis. One is Lavoisier's oxygen theory of combustion which replaced the phlogiston theory. Lavoisier explains various phenomena such as combustion and calcination of metals, saying

> I have deduced all the explanations from a simple principle, that pure or vital air is composed of a principle particular to it, which forms its base, and which I have named the *oxygen principle*, combined with the matter of fire and heat. Once this principle was admitted, the main difficulties of chemistry appeared to dissipate and vanish, and all the phenomena were explained with an astonishing simplicity.[17]

Lavoisier claims that his theory provides simpler explanations of the different phenomena because they are all made in terms of his basic oxygen principle, and do not need to invoke particular assumptions, such as that the phlogiston given off in combustion has 'negative weight' (see Thagard (1978), pp. 77-78).

[17]Thagard's translation. French original is: 'J'ai déduit toutes les explications d'un principe simple, c'est que l'air pur, l'air vital, est composé d'un principe particulier qui lui est propre, qui en forme la base, et que j'ai nommé principe oxygine, combiné avec la matière du feu et de la chaleur. Ce principe une fois admis, les principales difficultés de la chimie ont paru s'évanouir et se dissiper, et tous les phénomènes se sont expliqués avec une étonnante simplicité' (Lavoisier (1783)).

Perhaps the most dominant intuition in the philosophy of science literature regarding the simplicity of a scientific theory has been that adding adjustable parameters makes a theory less simple (Popper (1959), Rosenkrantz (1977), Howson and Urbach (2006)). This does not necessarily mean that one can compare the simplicity of theories by simply counting parameters. But if one adds an additional adjustable parameter to an existing theory, it will make it more complex. For example, let theory T_1 be the theory that the relation between two quantities is linear. A line is described by $y = \theta_1 x + \theta_0$, which has two parameters. We could add another adjustable parameter, θ_2, to get quadratic curves. Let T_2 be the theory that the relation between the quantities is quadratic. Intuitively, the linear theory T_1 is simpler than the quadratic theory T_2, or any higher order polynomial theory. As another example, suppose that T_1 is the theory that planets move in circular orbits, and T_2 is the theory that planets move in elliptical orbits. Circular orbits can be described by an equation with three parameters, namely $\frac{(x-x_0)^2+(y-y_0)^2}{r^2} = 1$, whereas the ellipse $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1$ has four parameters. Again, the idea is that intuitively T_1 is simpler. Specifying the values of adjustable parameters in a theory may be seen as a special case of auxiliary hypotheses. If a theory relies on setting these parameters in particular ways in order to explain phenomena, then the explanations are not based on the core of the theory, but on auxiliary hypotheses.

If the phenomena to be explained belong to various domains, then the explanation is better – and may be called 'simpler' – the less it relies on particular auxiliaries for each domain. In this case, the relevant explanatory virtue may also be called 'unification'. For example, the Newtonian theory of light treated it as a stream of corpuscles. It could explain diffraction phenomena, but it invoked different auxiliary hypotheses for different types of diffraction. The pattern of light observed on a screen as the result of shining light on a hair was explained in terms of an ether

surrounding the hair whose variation in density produced the pattern on the screen. The observation of rings between two pieces of glass was explained by postulating that by passing through a medium such as glass, corpuscles of light acquire a disposition to either reflect or transmit through a subsequent surface. Newton called these dispositions 'fits of easy transmission and easy reflection'. The corpuscular explanation of thin slit diffraction relied on yet further auxiliary hypotheses. By contrast, the wave theory of light could use the same basic principles to explain all these different phenomena of diffraction. Therefore it was a more unifying theory with respect to these observations than the corpuscular theory.

The Copernican explanation of retrograde motion may also be seen as more unifying, due to its greater dependence on core principles, than the Ptolemaic explanation. The Copernican theory relates different phenomena, such as the sidereal and synodic periods, which are unrelated in Ptolemy's theory. It also relates the different motions of the planets in that they all depend on the common factor of the earth's motion with respect to the sun.

When it comes to explaining just one phenomenon, rather than phenomena from various domains, the notion of unification does not apply, but a theory can still possess the explanatory virtue of simplicity in relation to this phenomenon if it explains it in terms of core principles, rather than auxiliary hypotheses. The fact that, when it comes to explaining diverse phenomena, simplicity and unification essentially amount to the same consideration explains why it can sometimes be hard to distinguish which virtue people are invoking. For example, the quotation from Rheticus in section 2.7.3 could be seen as an appeal to either simplicity or unification.

2.7.6 Other explanatory virtues

So far we have considered comparing theories which provide explanations of the same set of phenomena, either in a single domain, or in various domains. One theory can be better than another at explaining these given phenomena. However, another aspect of providing a better explanation is providing an explanation of more phenomena or more various phenomena. This virtue, which is sometimes called 'scope' or 'consilience' typically trades off against the virtue of simplicity or unification, since it may be necessary to bring in more specific auxiliary hypotheses in order to allow the theory to provide explanations across a wider range of domains.

The explanatory performance of a theory may also be judged over time. Thus some theories have the virtue of 'fruitfulness', meaning that they are able to provide simple explanations of an increasingly wide range of phenomena as more evidence is collected.

2.7.7 Summary

I suggest then that there is a certain conceptual 'skeleton' of IBE which governs how it assigns degrees of belief. The considerations which go into making a theory a better explanation of certain phenomena than another are first that it is more plausible, second, that it makes the explanandum more expected, and third that it does so in terms of its core principles. I have illustrated the third consideration in a couple of complex examples. In simple cases, where the theories in question are just simple hypotheses with no structure, the third consideration does not play a role. That is why the two-component picture proposed by Okasha was quite convincing in simple cases like the doctor's diagnosis.

2.8 Bayesian representation of IBE

This skeleton of IBE can readily be represented in Bayesian terms. The Bayesian represents the component of IBE which involves plausibility of the theory by the prior probability for the theory. And the component of how expected the hypothesis makes the evidence is represented by the likelihood of the hypothesis, given the evidence.

What about the concern with explaining in terms of the core? How is this represented in Bayesian terms? In Chapter 1, I gave an account of Bayesian confirmation of theories which are structured into framework theories consisting of more abstract or general claims, and more specific models. The framework theories compete in a 'higher level' hypothesis space, whereas the more specific parts compete in a 'lower level' hypothesis space. I now suggest that the scientific practice of identifying core elements of a theory to serve in explanation and identifying framework theories to serve in confirmation is one and the same. Thus, the core of one theory forms a higher level hypothesis that competes against the core of another theory. For instance, we may consider the higher level hypothesis space consisting of T_{cop}, the core of the Copernican theory, and T_{ptol}, the core of the Ptolemaic theory.

Now consider how the Bayesian assesses the support for high level theories such as T_{cop} and T_{ptol}. The Bayesian computes the posterior probability, which depends on the priors $p(T_{cop})$ and $p(T_{ptol})$ and the likelihoods $p(D|T_{cop})$ and $p(D|T_{ptol})$. The likelihood, say of T_{cop}, is an average of the the likelihoods for specific Copernican models generated by T_{cop}. Suppose these models are parametrised by θ, which may represent variables like the number of planets, the periods of their orbits and any epicycles, the ordering of the planets and so on. Then

$$p(D|T_{cop}) = \int p(D|\theta, T_{cop})p(\theta|T_{cop})d\theta \tag{2.4}$$

T_{cop} will not necessarily have a high likelihood just because it is able to generate a specific model which fits the data extremely well. It also matters what the proportion of well-fitting specific cases to ill-fitting ones is. If a higher level theory generates a relatively large number of specific models which don't fit well at all, then the ill-fitting models bring the average down. It will generally do better if a good fit can be achieved across a wider range of specific models. This means that the good fit is due to T_{cop} itself, and not just to specific values of parameter settings.

The fact that the Bayesian likelihood for T_{cop} is an average of lower level likelihoods represents the explanatory concern that the theory should explain in terms of its core. There is no need for the assignment of priors and likelihoods to be directly governed by explanatory considerations such as simplicity. Rather, the priors are regarded, in the usual Bayesian way, as representing plausibility judgments.

For example, consider how to represent Williams' parsimony argument against group selection in Bayesian terms. Sober has suggested a representation based on taking the prior for the group selection theory to be lower based on its lack of simplicity. This is a representation, such as the composite view advocates, where explanatory considerations are put in charge of assigning Bayesian probabilities.

I suggest an alternative way to represent this example in which priors are still assigned in the usual way to represent plausibility of the theory. The group selection core theory T_{GS} and the individual selection core theory T_{IS} may initially be regarded as equally plausible. But we saw that scientists were concerned that the range of parameter values for which group selection can override individual selection is very small. This may be thought of as the claim that the likelihood of altruistic traits is only high for a small region of parameter values. If we don't know before-

hand that the natural system falls in this parameter range, we will not concentrate all the prior probability $p(\theta|T_{GS})$ in that range. Thus, this will mean a low average likelihood for the theory.[18] This gives us a Bayesian account of Williams' parsimony argument which reflects the concern that the group selection theory needs to be overly fine-tuned in order to accommodate the data. It says that the group selection theory is penalised by the likelihood, even if it is initially given the same prior probability as the individual selection theory.[19]

2.8.1 A concern over objectivity

Some may be concerned about taking IBE to depend in part on a judgment of the plausibility of the theory in question. IBE is generally regarded as a fairly 'objective' rule, in the sense that it has a limited range of permissible assessments of certain hypotheses in the light of given evidence. However, plausibility judgments could be subject to a great deal of intersubjective disagreement. Doesn't this undermine the objectivity of IBE?

My answer is that insofar as IBE shares this component of a plausibility judgment with Bayesianism, it shares the same issues over its objectivity. If one regards IBE as an objective rule, it may be best married with more objective forms of Bayesianism, according to which more constraints are placed on which judgements of plausibility are rational. Then, to be admitted, similar objections to the constraints imposed by objective Bayesianism would apply. In particular, there are

[18] If we had independent reason to believe that the natural system is likely to have parameters in the region which gives a high likelihood for the selection of altruistic traits, then the likelihood $p(E|T_{GS})$ would be much higher.

[19] As I mentioned in section 2.7.4, the explanation provided by the individual selection theory also depends on auxiliary hypotheses. Once these are taken into account, one would have to investigate in more detail which of the two theories has a lower likelihood, given the evidence for altruism.

certain classic objections to the Principle of Indifference. These objections are based on the idea that there is no principled way to decide what we should be indifferent with respect to.

For example, consider someone who takes a trip of 100 miles in between 1 and 2 hours.[20] What is the probability that the trip took between 1 hour and $1\frac{1}{2}$ hours? If we assign a uniform probability density over times, we might say $\frac{1}{2}$, since the interval between 1 hour and $1\frac{1}{2}$ hours is half of the interval from 1 to 2 hours. But equally, we know that the average speed on the trip was between 50 and 100 miles per hour, and and we have no evidence to favour any one speed over any other within that interval. The trip took between 1 hour and $1\frac{1}{2}$ hours if the average speed was between $66\frac{2}{3}$ miles per hour and 100 miles per hour. The interval between $66\frac{2}{3}$ miles per hour and 100 miles per hour is $\frac{2}{3}$ of the interval of possible speeds. Therefore, if we employ the Principle of Indifference with respect to speeds, we get a probability of $\frac{2}{3}$ which is different from the $\frac{1}{2}$ obtained by applying the Principle of Indifference to times.

What this problem indicates is that the tempting first thought that a non-informative prior should be uniform over the parameter space does not work because it is no longer uniform on reparametrisation. This can be seen in general terms in a simple case where there is just one parameter θ, and we reparametrise in terms of another parameter $\phi = f(\theta)$. To convert the measure between the two coordinate systems, we must multiply by the Jacobian, which in this case is $\left|\frac{d\theta}{d\phi}\right|$. That is, $p(\theta)$ is equivalent to $p(\phi)\left|\frac{d\theta}{d\phi}\right|$ in the new parametrisation. Thus, a prior which is uniform in θ is not uniform in ϕ.

There is a large literature on better ways to construct 'informationless priors'.[21]

[20] This example is discussed in Huemer (2009), pp. 349-350.

[21] See for example Bernardo and Smith (1994), Kass and Wasserman (1996) and references therein.

One may, for example, use a construction such as Jeffreys'. Jeffreys proposed that a reparametrisation-invariant prior can be defined as $p(\theta) \propto \sqrt{J(\theta)}$ where $J(\theta)$ is the 'Fisher information', defined as $J(\theta) = -E[\frac{d^2 \log p(y|\theta)}{d\theta^2}]$ (Jeffreys (1961)). This is invariant to reparametrisation in the sense that $p(\phi) = \sqrt{J(\phi)} = \sqrt{J(\theta)} \left|\frac{d\theta}{d\phi}\right|$.[22]

I do not want to enter into a discussion of this type of approach here. Indeed the extension of Jeffreys' priors to multi-variable situations may be quite non-trivial. I merely bring it up in order to indicate that there are potentially options available to both objective Bayesianism and objective IBE, despite the well-known objections to the Principle of Indifference. My concern here is to show how the key concepts involved in IBE can be represented in Bayesian terms. I have suggested that IBE and Bayesianism share a dependence on plausibility judgements, and may thus also share both concerns over whether these judgements should be further constrained, and possible strategies for constraining them.

2.8.2 Bayesian simplicity and unification

I have suggested that a theory provides a better explanation of certain phenomena if it explains more in terms of its core, and that this is essentially what the explanatory virtues of simplicity and unification amount to. Thus, the Bayesian account of better explanation I have just given is what has been offered as a Bayesian account of simplicity by various authors (e.g. Rosenkrantz (1977), Jefferys and Berger (1992), MacKay (2003)). If, as I have claimed, the virtues of unification and simplicity essentially coincide for phenomena which extend across multiple domains, we would expect the Bayesian account of unification to have the same basis as the Bayesian

[22] Kass (1989) and Myung et al. (2000) have given an interpretation of the Jeffreys' prior as the measure associated with the metric on the space of probability distributions given by the Fisher information. There are various arguments in favour of taking Fisher information as the natural metric on this space.

account of simplicity. However, the Bayesian account of unification by Myrvold (1996) has been presented as something quite different. Myrvold claims that it is unification which has epistemic significance, not simplicity – according to him, Ockham's razor is disposable.[23]

However, once one takes into account the various ceteris paribus conditions that Myrvold imposes, it turns out that his account really does rely on the more unifying theory having a higher likelihood than the less unifying one, and hence it does have the same Bayesian underpinning as the preference for simpler theories.[24] I will now show how this works in more detail.

Consider an explanandum consisting of two different phenomena, D_1 and D_2, which may at first sight appear to have little to do with each other. Myrvold aims to capture in Bayesian terms, the key intuition that a theory possesses the virtue of unification with respect to certain evidence if it makes one phenomenon yield information about some other phenomenon. In the case of the Copernican theory, for example, the synodic period of a planet yielded information about its sidereal period, and also information about one planet yielded information about another planet.

Myrvold measures how unifying a theory is by the extent to which it increases the probabilistic correlation between different pieces of evidence (Myrvold (1996)). The pieces of evidence D_1 and D_2 are positively correlated if learning D_1 raises the probability of D_2, $p(D_2|D_1) > p(D_2)$. They are negatively correlated if $p(D_2|D_1) < p(D_2)$, and probabilistically independent if $p(D_2|D_1) = p(D_2)$. Given certain reasonable

[23]This claim was made in a talk entitled 'Simplicity and Theory Choice: Ockham's disposable razor', presented by Myrvold at Pittsburgh-CMU graduate student conference, Pittsburgh, 2010. Myrvold also distances his account of unification from Forster and Sober's account of simplicity in terms of the Akaike Information Criterion (Forster and Sober (1994)), which is closely related to the Bayesian account of simplicity (Myrvold (1996), pp. 664-665).

[24]Schupbach (2005) makes a closely related point.

requirements on a measure of degree of informational relevance of D_1 to D_2, the following measure can be derived (Myrvold (1996)):

$$I(D_1, D_2) = \log \frac{p(D_2|D_1)}{p(D_2)} \tag{2.5}$$

$$= \log \frac{p(D_1, D_2)}{p(D_1)p(D_2)} \tag{2.6}$$

which is > 0 if D_1 and D_2 are positively correlated, $= 0$, if they are independent, and < 0 if they are negatively correlated. The informational relevance may be different when conditionalised on the theory T, and the conditional informational relevanceis given by

$$I(D_1, D_2|T) = \log \frac{p(D_2|D_1, T)}{p(D_2|T)}$$

$$= \log \frac{p(D_1, D_2|T)}{p(D_1|T)p(D_2|T)}$$

Myrvold proposes that the unifying power of theory T is the extent to which theory T increases the informational relevance of D_1 to D_2. Thus it may be measured by

$$U(D_1, D_2; T) = I(D_1, D_2|T) - I(D_1, D_2)$$

Myrvold argues that one can use this measure to demonstrate that the Copernican theory has greater unifying power than the Ptolemaic theory with respect to the phenomena given by the sidereal period and the synodic period of a particular planet. He takes p_m to be the statement that Mars traverses the sphere of fixed stars with a period that, within small observational error, is 1.88 years, and r_m as the statement that Mars retrogresses within a period equal to 2.14 years (within observational error). On the Copernican theory r_m and p_m are related by the equa-

tion 2.3 above, so p_m holds just when r_m does. This means that $p(p_m|r_m, T_{cop}) = 1$, and hence $I(p_m, r_m|T_{cop}) = -\log(\ p(p_m|T_{cop}))$. Since the probability for any particular value of the synodic period is quite small, $p(p_m|T_{cop})$ is small, and hence $I(p_m, r_m|T_{cop})$ is reasonably large. By contrast, on the Ptolemaic theory, r_m yields no information about p_m and so $I(p_m, r_m|T_{ptol}) \approx 0$. Thus, the unifying power of the Copernican theory with respect to these phenomena is greater than that of the Ptolemaic theory $U(p_m, r_m; T_{cop}) > U(p_m, r_m; T_{ptol})$.

The claim that the unifying power of the Copernican theory is greater than that of the Ptolemaic theory is based on $I(p_m, r_m|T_{cop}) > I(p_m, r_m|T_{ptol})$ since $I(p_m, r_m)$ is the same quantity subtracted from each. The basic reason why this inequality holds is that the Copernican theory makes the phenomena more correlated

$$p(p_m, r_m|T_{cop}) > p(p_m|T_{cop})p(r_m|T_{cop})$$

whereas the Ptolemaic theory does not

$$p(p_m, r_m|T_{ptol}) = p(p_m|T_{ptol})p(r_m|T_{ptol})$$

Myrvold suggests (p. 414), that the evidential support lent to T_{cop} by p_m is approximately the same as the evidential support lent to T_{ptol} by p_m (and similarly for r_m). Presumably just learning the synodic period of one planet, or the sidereal period of that planet, doesn't give much, if any, support to either hypothesis. That is, we take $p(p_m|T_{cop}) \approx p(p_m|T_{ptol})$ and $p(r_m|T_{cop}) \approx p(r_m|T_{ptol})$. Then the greater unifying power of the Copernican theory rests on the Copernican theory having a greater likelihood for producing these two pieces of evidence than the Ptolemaic theory

$$p(p_m, r_m|T_{cop}) > p(p_m, r_m|T_{ptol})$$

Thus, Myrvold's Bayesian account of unification and the Bayesian account of simplicity both have a common basis. Given the data, the likelihood of the simpler or more unifying theory is higher than that of the theory which is less simple or unifying.

2.8.3 Summary

I have argued that there is more to the concept of IBE than in Okasha's simple breakdown into two components corresponding to prior and likelihood. I have proposed that the further considerations in IBE arise from the desire that the theory explain the phenomena in terms of its core principles. That is, ideally the theory should make the explanandum expected no matter how the details of the theory are filled in. This preference is naturally represented by the Bayesian who computes the likelihood for the core of the theory by averaging over the specific models that it generates.

2.9 Implications for IBE

The significant idea behind IBE as an inductive method is that explanation is the driving force in inference. In this chapter, I have attempted to extract the key features of IBE – its conceptual 'skeleton' – by thinking about the motivating concept of explanation. I then showed how this conceptual skeleton maps onto the probabilities involved in inference according to Bayesian principles. What is the upshot of this for IBE? Should the conceptual skeleton be simply identified with its Bayesian representation? Is what I have provided an eliminative reduction of IBE in terms of Bayesianism?

I am not claiming that the concept of IBE should be identified with the Bayesian

representation. Thinking about inference in terms of explanation still remains a distinct way of conceptualising it than thinking about it in terms of Bayesian probabilities. In particular, as we have seen, thinking about explanation may involve more of a focus on the details of deductive or causal relationships than is captured by conditional probabilities alone. IBE and Bayesianism might be regarded as different levels of description of the inference process, where IBE describes the gross phenomena of inferential thinking captured by IBE, whereas the quantitative detail which gives rise to these explanatory features is captured by the Bayesian model. As a suggestive analogy, consider the different levels of description provided by thermodynamics and statistical mechanics, where thermodynamics describes thermal phenomena at the macroscopic level and statistical mechanics provides a lower-level description of the molecular underpinnings. Thermodynamic concepts like heat, entropy and energy are not made redundant by statistical mechanics. Rather they are taken to refer to certain properties of collections of molecules, properties which may themselves be represented in terms of probabilities. Similarly, explanatory concepts like simplicity, scope and unification are not made redundant by the Bayesian account. They refer to certain properties of the relationship between a theory and the evidence, properties which may in turn be represented in terms of probabilities.

Although I admit that there may be more to the concept of IBE, I want to suggest that the skeleton of IBE I have identified may be sufficiently inclusive to capture all that is epistemically relevant about IBE. That is, there are no other explanatory considerations which affect the degrees of belief that an explanationist should assign to the various hypotheses. This does not rule out the possibility of fleshing out the skeleton with further considerations of a pragmatic nature. These considerations could be represented by a Bayesian using a full decision theoretic

model with the pragmatic considerations represented by utilities.

2.10 Conclusion

In this chapter I have considered two existing pictures of the relationship between IBE and Bayesianism. The first is the incompatibilist view that IBE and Bayesianism are mutually exclusive alternatives. The second is the composite view, which says that IBE can be represented in Bayesian terms by virtue of explanatory considerations playing the role of determining the priors and likelihoods. I argued that neither of these pictures is particularly appealing.

I proposed a third alternative, which is that IBE can be represented in Bayesian terms without explanatory considerations providing any special constraints on the priors and likelihoods. I attempted to isolate within the concept of IBE a skeleton which can be represented in Bayesian terms. The skeleton is the following: a theory provides a better explanation of certain evidence to the extent that a) it is initially plausible, b) it provides a connection to the evidence which makes that evidence expected, and c) it does so without that connection relying too heavily upon auxiliary hypotheses. This skeleton is then readily represented in Bayesian terms. Plausibility is represented by the prior, expectedness of the evidence by the likelihoods, and the concern about reliance on auxiliary hypotheses is reflected in the Bayesian practice of averaging over specific hypotheses generated by the theory to get the likelihood of the theory. To the extent that the conceptual skeleton I have identified captures all the considerations which affect the degrees of belief assigned by an explanationist, IBE will be compatible with Bayesianism. But it is a different picture of compatibility than that proposed by the composite view, since there is no need for explanatory considerations to play a role in assigning priors and

likelihoods. Rather explanatory virtues emerge as considerations naturally taken into account in Bayesian updating. If the two methods are compatible in this sense, they may be mutually illuminating. IBE provides a way to describe in qualitative terms what is going on in Bayesian inference, and the Bayesian account provides a way to explain why explanatory virtues are preferred in IBE.

Chapter 3

No Miracles, No Fallacies

3.1 Introduction

Is it reasonable to think that theories in mature science are true, or at least approximately true? That is, should one be a scientific realist? Or is it better to take the anti-realist view that scientific theories just 'save the phenomena'? On this view, scientific theories are roughly right about what could be observed, but we need not believe their more theoretical claims about the world behind the phenomena.

If a theory makes a successful prediction, perhaps this is a reason to think it is approximately true. After all, wouldn't it be a miracle if a theory that was nowhere close to the truth nonetheless managed to hit upon a successful prediction? We have no need to accept such miracles, so we are better to conclude that a predictively successful theory is close to the truth. This is the thought behind one of the most compelling arguments for scientific realism: the 'No Miracles' argument (NMA) (Putnam (1975), Boyd (1983)).

On the other hand, in the history of science, there are many theories which were predictively successful at the time, yet which, by the standards of today's theories,

made theoretical claims which were nowhere close to the truth. Absent any reason to think that there is a salient difference between the past and the present, these counterexamples in the past should make us doubt that the success of a currently held theory gives us any indication of its approximate truth. This, roughly speaking, is the argument against realism which has been called the 'Pessimistic Induction' (PI) (Laudan (1981)).

The NMA and the PI are often seen as articulating the central considerations for and against scientific realism. However, recently there have been attempts to dismiss either or both arguments on the grounds that they involve the common fallacy of neglecting the 'base-rate' (Howson (2000), Lipton (2004), Lewis (2001), Magnus and Callender (2004)). Allegedly, the realist making the NMA ignores the fact that before taking its success into consideration, a scientific theory is very unlikely to be approximately true (Howson (2000), Lipton (2004)). And the proponent of the PI stands independently accused of this very same neglect (Lewis (2001)).

In this paper, I will argue that, despite these allegations, both the NMA and the PI can be reconstructed in a non-fallacious manner. In the NMA, the realist does not neglect the low base-rate, but instead denies that it is very low. This denial can be expressed as an explicit premise of the argument. I will sketch a possible defence of this extra premise and indicate its weaknesses. The claim that the PI involves the base-rate fallacy is based on misconstruing what the target of the PI is. When correctly construed, it becomes clear that no fallacy is involved in the PI either.

3.2 The base-rate fallacy

Suppose that doctors conduct a test for a rare disease that occurs in 1 in a thousand patients. There is virtually no chance that a person with the disease will test negative (i.e. the false negative rate can be regarded as zero). On the other hand, the chance of a person without the disease testing positive is about 5% (i.e. the false positive rate is 5%). A patient takes the test and his result is positive. What is the probability that the patient has the disease?[1]

It is a surprise to many people that the correct answer to this is just under 2%. Although we should judge the probability that the patient has the disease as higher now than before he took the test, it is still quite unlikely that he actually has the disease. He is more likely to be one of the false positives. Out of 1000 patients, say, there will be approximately 50 false positives (5% of 1000), but only one true case of the disease. Therefore the fraction of those with a positive result who actually have the disease is $\frac{1}{51}$ or just under 2%.

The result can be derived more formally using Bayesian analysis. Let P represent 'the patient tests positive for the disease' and D represent 'the patient has the disease'. The information that was given in the problem is:

(i) $p(P|D) = 1$ (the false negative rate is zero)

(ii) $p(P|\neg D) = 0.05$ (the false positive rate is 5%)

(iii) $p(D) = \frac{1}{1000}$

The question is what is $p(D|P)$, that is, the probability that the patient has the

[1]This example is cited by Howson (2000) (p. 52), as a problem which was actually given to students and staff at Harvard Medical school.

disease given that he tests positive for it. This can be calculated using Bayes' rule

$$p(D|P) = \frac{p(P|D)p(D)}{p(P|D)p(D) + p(P|\neg D)p(\neg D)}$$
$$= \frac{p(D)}{p(D) + BF.(1 - p(D))}$$

where BF is the 'Bayes factor' $BF = \frac{p(P|\neg D)}{p(P|D)}$ which in this case is 0.05. Combining this with $p(D) = \frac{1}{1000}$ gives the correct conclusion that $p(D|P) = \frac{1}{51}$.

There is a tendency when presented with a problem like this for people to overestimate the probability that the patient with the positive test has the disease. Psychologists have demonstrated people's tendency to overestimate the probabilities in a range of similar problem cases. These problems all require a subject to integrate two types of probabilistic information: generic information about base rates in a population – such as the rate of the disease – with specific information about which scenario would make the particular result found more likely – such as the probability that someone with the disease will test positive. One reason that people may overestimate the probabilities is because they neglect the generic base-rates in favour of the more specific information – this inappropriate neglect of base rates is called the 'base-rate fallacy'. So, in the disease case, for instance, people ignore the given information that the overall rate of disease in the population is low, and focus on the information that a patient with the disease is very likely to test positive.

There are a number of psychological theories about the underlying reasons for the base-rate neglect. The original idea, due to Kahneman and Tversky, was that people failed to follow the Bayesian norms, because they instead followed heuristics, such as judging 'representativeness'. The idea roughly is that people judge that a positive test is more characteristic or representative of someone who has the disease and hence the person is more likely to have the disease (Kahneman and

Tversky (1972), Kahneman and Tversky (1973)). Other psychologists have argued that people do perform Bayesian inference, but the manner in which the probabilistic information is presented to subjects in the standard experiments on base-rate neglect obscures this.[2]

3.3 The scientific realism debate

Before explaining how the NMA and the PI might be regarded as instances of the base-rate fallacy, I will first attempt to clarify what is at stake in the issue of scientific realism. Scientific theories describe many things, some of which can be relatively easily observed, such as the times of tides or the behaviours of animals, and others which are so theoretical that they are impossible to observe at the current time and quite possibly also in principle. Examples in the latter category might include the curvature of space-time or the forces between fundamental particles. Nonetheless, it is common to think that the unobservable parts of reality do exist and furthermore that scientific theories attempt to describe them and can be right or wrong in what they say about them. This much is common ground in the current debate over scientific realism, at least that part of the dispute in which the NMA and PI figure.[3] There are several different ways to characterise the difference between the realist and the anti-realist, and as we shall see in the next section, these will result in slightly different formulations of the NMA.

One way to characterise the difference between a realist and an anti-realist is in terms of what they say about the aim of science. For example, on Bas van Fraassen's

[2]See, for instance, Gigerenzer and Hoffrage (1995), Krynski and Tenenbaum (2007).

[3]That is, the debate involving the NMA and the PI takes for granted both 'metaphysical realism', the thesis that there is a mind-independent world, and 'semantic realism', the view that scientific theories may be read 'literally', so that theoretical claims may be true or false, and not just via a reduction to an observation language.

view, realism makes the claim that 'Science aims to give us, in its theories, a literally true story of what the world is like' (van Fraassen (1980), p. 8), whereas his brand of anti-realism, constructive empiricism, claims that 'Science aims to give us theories which are empirically adequate' (ibid, p. 12). A theory is empirically adequate if 'what it says about the observable things and events in this world is true – exactly if it "saves the phenomena"' (ibid, p. 12). That is, it is a weaker claim to say that a theory is empirically adequate than to say that it is true, since the empirically adequate theory is only right about the observable aspects of the world, but does not need to also be right about unobservable matters, as the true theory does. Indeed an empirically adequate theory could, in principle, be radically false in what it says about the unobservable. Thus, according to the realist, science concerns itself with getting the right story about the unobservable as well as the observable, whereas according to the anti-realist, it concerns itself only with the observable.

It can be difficult to pin down what is meant by the 'aim' of science (Rosen (1994)), so a number of authors prefer to pose the realist view as a claim about what the scientific method actually achieves. The realist may claim, for example, that the 'theories accepted in mature science are typically approximately true' (Putnam (1975), p. 73). This may be thought of as a 'global' claim that the scientific method is generally reliable in finding approximately true theories. The qualification 'approximately' is included, because the realist realises that it would be too strong to claim that theories are typically true, given the extent of idealisation, approximation and indeed error, in science. There have been attempts to give formal theories of what approximate truth amounts to (Niiniluoto (1998)), but some philosophers suggest that an intuitive notion is sufficient (Psillos (1999), Chapter 11). In contrast to the realist, the anti-realist claims that 'theories accepted in mature science are typically approximately empirically adequate'. A theory is ap-

proximately empirically adequate just if what it says about the observable things and events in the world is approximately true. Thus the anti-realist claims that the achievements of science are more modest than the realist does.

Another way to distinguish between realism and anti-realism is to see them as making different 'local' claims about how to regard a particular theory in a mature science – this may be a theory which one 'accepts'. The realist will say that such a theory is approximately true, whereas the anti-realist will say it is approximately empirically adequate. Some may prefer to claim more cautiously that the theory is probably approximately true, or probably approximately empirically adequate. Whilst the realist may recommend an attitude of belief in the theory, there are two possible attitudes that the anti-realist could adopt. One is that the theory is unlikely to be approximately true, since there can be so many ways to be approximately empirically adequate without being approximately true. However, in practice, anti-realists do not commonly recommend low credence or disbelief in the theory, but rather suspension of belief in the unobservable parts of the theory – an attitude of agnosticism.

3.4 The NMA

The NMA is an argument that predictive success in science provides reason to be a realist. The NMA is not generally presented as an a priori philosophical argument, but is rather put forward as part of a naturalistic epistemology according to which the thesis of realism has a similar status to ordinary scientific hypotheses. The argument basically has the following form. If realism were not true, predictive success would be a great coincidence or miracle. We should rule out such miracles, at least if we have a non-miraculous alternative, and we do, since on the realist

view, predictive success would be just what was expected. Therefore we should infer that realism is true.

3.4.1 Retail and wholesale NMA

In section 3.3, we saw several different ways of formulating the thesis of realism. We distinguished in particular between the local claim that a particular theory is approximately true, and the global claim that the theories accepted in mature science are typically approximately true. The NMA can be formulated as an argument for either one of these realist claims.

According to what has been called the 'retail' version of the NMA,[4] the predictive success of a particular theory is a reason for adopting the local realist claim that the theory is (probably) approximately true. A classic case of the type of success in question is the prediction of the Poisson spot by Augustin-Jean Fresnel's wave theory of light. In the early 19th century, the wave theory emerged as a rival to Newton's theory according to which light consisted of a stream of particles. One task was to explain diffraction patterns, where dark and light bands were observed when light was shone through slits or around obstacles. According to the corpuscular theories, the explanation involved postulating forces exerted by the obstacle on the particles, whereas wave theories allowed one to calculate intensities on the screen by adding up contributions from different points of the wavefront, imagining each point as a little source. In a well-known piece of scientific history, when Fresnel presented his theory, Siméon-Denis Poisson, a proponent of corpuscular theories, pointed out that according to Fresnel's theory, if an opaque disk was placed between a light source and a screen, a bright spot would appear at the centre of the disk's

[4]The 'retail' terminology is due to Magnus and Callender (2004), and is contrasted with 'wholesale'.

shadow on the screen. No-one expected that this could be true, but then the experiment was performed, and what is now known as the 'Poisson spot' was observed. Thus Fresnel's theory experienced a striking predictive success. The intuition behind the retail NMA is that Fresnel's theory could not have been so successful in prediction if it was not somehow latching onto the correct story about the unobservable nature of light. John Worrall expresses the intuition in the following way: 'it seems implausible that such a "nice" theory as the wave theory of light should get such a striking phenomenon as the white spot correct and yet not be somehow "on the right lines" concerning what it says about "deep structure"' (Worrall (2005), p. 24). According to this thinking, the alternatives to local realism – such as that the theory in question is approximately empirically adequate (but not necessarily approximately true), or that the theory is not even empirically adequate – would not explain the success – they would make it a miracle.

The 'wholesale' version of the NMA is an argument for the global claim that the 'theories accepted in mature science are typically approximately true'. The success invoked in this version is that of science as a whole – the fact that so many scientific theories do make successful predictions. The classic formulation of this version of the argument is due to Putnam:

> 'The positive argument for scientific realism is that it is the only philosophy that does not make the success of science a miracle. That terms in a mature science typically refer (this formulation is due to Richard Boyd), that the theories accepted in a mature science are typically approximately true, that the same terms can refer to the same even when they occur in different theories – these statements are viewed not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate description of science and

its relations to its objects' (Putnam (1975), p. 73).

The intuition here is that scientists would not manage to come up with successful theories if their method were not generally reliable in reaching approximately true theories. If, on the other hand, the scientific method tends to produce theories which are only approximately empirically adequate, rather than approximately true, it would be very surprising if those theories were so often successful in prediction.

3.4.2 Probabilistic formulation

It has recently been suggested that the retail version of the NMA can be formulated in terms of probabilities (Howson (2000), Lipton (2004)). The suggestion is that the argument has the form $\mathbf{NMA}_{R1}$:

(i) if T is approximately true, it is quite likely that the theory is predictively successful

(ii) if T is not approximately true, it is extremely improbable that the theory would be predictively successful

Conclusion: given that the theory is successful, we should infer that it is approximately true. Or, in probabilistic terms: the probability that the theory is approximately true is very high, given that it is successful.

Let R be the realist claim that 'theory T is approximately true', and S be the claim that 'theory T is predictively successful'. Then $\mathbf{NMA}_{R1}$ can be written as:

(i) $p(S|R)$ high

(ii) $p(S|\neg R)$ relatively low

C: $p(R|S)$ is high

The first premise is not very controversial.[5] It may even be taken to impose a

[5]Laudan takes issue with the claim that 'if a theory were approximately true, it would de-

constraint on what is meant by approximate truth that any adequate account must satisfy (Howson (2000), p. 56).

There is an interpretation of the second premise which raises difficulties. If one thinks of the probability in the second premise as a 'chance', defined by the ratio of the cases where T is successful and false to the total number of cases where T is false, then, for a number of reasons explained in Howson (2000), it seems impossible to justify. The total number of ways that T could be false is surely infinite, and there is no natural metric structure on the space of possibilities. Therefore, there can be no determinate value for the chance. Not only this, but one would also need to assume something about how probable it is that each possible world is the true world, and there seem to be no grounds for knowing this. Thus trying to establish even that there is a definitive chance, let alone that its value is low, seems like a hopeless task.

However, one may also think of probabilities in Bayesian terms as representing the degrees of belief which it is reasonable to hold. Since they represent reasonable degrees of belief, they are not entirely subjective, but rather any assignment must be supported by an argument. Now it seems possible to give an argument that the probability $p(S|\neg R)$ is low. The realist will say that, in the world, unobservable parts of reality may be causally connected to observable events, just as other observable parts of reality are. Part of the reason that a scientific theory manages to give a successful prediction is that it correctly captures some of the causal structure behind the predicted event, and these causal relations may be unobservable as well as observable. Since unobservable causal structure plays a role in determining what predictions a theory will make, it would be very surprising if the actual unobserv-

---

ductively follow that the theory would be a relative successful predictor and explainer of the observable phenomena' (Laudan (1981), p. 30), but this is a stronger claim than that approximate truth makes success likely.

able causal structure was very different from what the theory says, yet the theory still managed to predict correctly. This is one way that the realist might argue for premise (ii).

3.5 The NMA and the base-rate fallacy

3.5.1 The allegation

According to Colin Howson, all the philosophers who use the No-Miracles argument commit the base-rate fallacy (Howson (2000), p. 54), and the idea that philosophers of science are particularly susceptible to this 'blindspot' is also put forward by Peter Lipton (Lipton (2004), p. 197). The allegation that the argument involves the base-rate fallacy involves, not the justification for premises (i) and (ii), but the move from these to the conclusion. In assessing the base-rate fallacy allegation, we may take premises (i) and (ii) as given.

The argument that this step of the NMA is fallacious is based on drawing a parallel with the disease problem we saw in section 3.2. A theory being approximately true is analogous to the patient having the disease. Being successful in prediction is analogous to testing positive for the disease. Thus, premise (i) is the analogue of the information that the patient is almost certain to test positive if they have the disease (the false negative rate is negligible). Premise (ii) is the analogue of the information that the patient is unlikely to test positive if they don't have the disease (the false positive rate is low). Concluding that the theory is probably approximately true, given that it is successful, is like concluding that the patient is very likely to have the disease, given a positive test result. In the disease case this was a fallacy because, at least allegedly, it involved neglecting the very low base-rate of the disease in the overall population. In the case of the NMA the idea

must also be that the realist neglects the very low base-rate that there are very few approximately true theories amongst all possible theories.

3.5.2 Amending the probabilistic formulation of the NMA

As it stands, argument $\mathbf{NMA}_{R1}$ is not valid. To make it valid, one needs an extra premise that

(iii) $p(R)$ is not small compared to $p(S|\neg R)$

Similarly, in the diseases case, it is only valid to draw the conclusion that $p(D|P)$ is high if one accepts the premise that the prior probability $p(D)$ is not small compared to $p(P|\neg D)$. The problem in the diseases case was that this premise was explicitly denied in the set-up, yet people failed to pick up on the information that $p(D)$ was actually extremely small. Therefore it was reasonable to interpret those people's responses as involving base-rate neglect.

In the NMA case, on the other hand, premise (iii) was not explicitly denied. No-one gave realists the appropriate base-rate. Therefore the realists did not ignore something which was explicitly available to them.

Rather an alternative interpretation seems more natural. That is, the NMA was not adequately represented by $\mathbf{NMA}_{R1}$. It should have been formulated, including premise (iii), as $\mathbf{NMA}_{R2}$:

(i) $p(S|R)$ high

(ii) $p(S|\neg R)$ relatively low

(iii) $p(R)$ is not small compared to $p(S|\neg R)$

Conclusion: $p(R|S)$ is high.

It is not unreasonable to amend the argument in this way. Psychological experiments which find a tendency to commit the base-rate fallacy constitute quite a different context from the philosophical context in which the NMA is put for-

ward. The psychological experiments aim to elicit people's initial judgments, not their considered opinion after reflection and debate. It is of course possible that the initial attraction of the NMA is derived in part from a seductive fallacy. But people are attracted to arguments for many different reasons. The concern in the philosophical context is whether the argument can be posed in a non-fallacious manner. Consider what would be typical advice in the philosophical context on what to do if you discover a fallacy. As Stephen Toulmin puts it in an introductory textbook on reasoning: 'the perpetrator of a fallacy does not have to withdraw from the discussion in disgrace but simply must restate the argument with modifications that eliminate the fallacy' (Toulmin et al. (1984), p. 178). That is what the above formulation has done.

It is in fact reasonable to see premise (iii) as implicit in the No Miracles reasoning because dismissing the possibility of a mere coincidence is something that one should only do if one thinks that the alternative has a reasonable chance of being true. For example, consider a murder investigation. Suppose DNA from Mr X is found on the victim's clothing. The chance of this match if Mr X is the murderer is very high, but it is extremely low if Mr X was not involved in the murder. We conclude that Mr X is probably the murderer, dismissing the possibility that the match was a mere coincidence. Implicitly we are assuming that it is not out of the question that Mr X was involved, even before seeing the DNA evidence. However, if Mr X had come up with an excellent and well-corroborated alibi, this would have meant that we had a much lower prior probability for his involvement, and in this case, it may no longer be reasonable to dismiss the possibility of a merely coincidental match.

Similarly, suppose a child develops spots characteristic of measles. You conclude on this basis that the child probably has measles. It is unsurprising that the child

has spots if she does indeed have measles, but on the other hand, if she doesn't have measles, the spots would be quite a coincidence – a minor miracle for which we have no explanation. We rule out such coincidences or miracles, so we conclude that she probably does have measles. It was only reasonable to rule out the coincidence if there was a reasonable chance that the child had somehow contracted measles. If we happened to know that she had absolutely not been in any situation where she could have caught them, we might be inclined to take the resemblance to spots produced by measles as just a coincidence.

Or suppose that a German octopus named Paul manages to accurately predict a sequence of outcomes of German World Cup matches. If Paul were psychic, his success would be quite likely, but if he were not psychic, it seems implausible that he could have got so many games right by chance. However we are still reluctant to conclude that he is psychic, because we have such a low prior probability for the existence of psychic powers, especially in German octopi.

Thus, it seems to be a prerequisite for ruling out the coincidence or 'miracle' that the alternative hypothesis which would make the observations non-miraculous is not extremely unlikely a priori. Therefore it makes sense to incorporate this prerequisite explicitly into the argument above as premise (iii).

3.5.3 Defence of additional premise

Our attention is now refocused on the premises, rather than the validity of the NMA. We have added an additional premise to the probabilistic formulation of the NMA. What is its status? How would the realist defend it?

One line of thinking which seems to cast doubt on the premise goes as follows. There are enormously many possible theories, and only a very small fraction of them can be approximately true. Admittedly, the probability of a theory being ap-

proximately true depends on what is meant by approximate truth: the more liberal the notion of approximate truth, the more theories would count as approximately true. However, so the thought goes, no reasonable account of approximate truth can be liberal enough that a large proportion of possible theories would count as approximately true. Surely they must always be vastly outnumbered by false theories, since there are so many ways to go wrong and comparatively few to go right.

This argument suffers from the same objections which attended the attempt to spell out premise (ii) in terms of the ratio of successful to all possible cases. If the relevant set of theories is all possible theories, this is an infinite and metrically unstructured space of possibilities in which we will have a hard time pinning down any measure. Also, since many of these possible theories are currently unimagined, we have no grasp of what the space consists of (Howson (2000), Worrall (2005)). It would not be unreasonable for the realist to reject this type of argument.

What alternative can the realist offer? Consider again the diseases case. To get a prior probability for the patient having the disease, you will have to assume something about how the patient was drawn from the population at large. Even if you know that the disease is indeed very rare in the population at large, you might nonetheless think that the patient is more likely to have the disease just because he has presented himself as a patient to be tested for the disease. Probably he experienced some symptoms which led to him being tested. Thus the patient is not randomly drawn from the population. This is a reason that you might legitimately think that the prior probability of the patient having the disease is higher than the base-rate in the general population.

Similarly, the theories which are candidates for novel predictive success are not a random sample of all possible theories. Rather these are the theories which are deemed best-confirmed by scientific methods in their particular domain. The

realist may at least partially base a higher prior probability that a particular theory is approximately true on their assumption that the proportion of best-confirmed theories which are approximately true is high.

In the diseases case, even if the probability that a patient has the disease is different from the probability that a random member of the population has it, there is still room for disagreement about whether the process of selecting someone as a patient tends to produce a high proportion of individuals with the disease or whether it tends to produce a low proportion. Similarly, there can be disagreement over whether the process of selecting theories as well-confirmed tends to produce a high or low proportion of approximately true theories. This is exactly what divides the realist from the anti-realist with respect to the global version of realism. The realist thinks that the theories accepted in mature science are typically approximately true, whereas the anti-realist thinks they are typically only approximately empirically adequate, and hence not so often approximately true, since the vast majority of approximately empirically adequate theories are not approximately true.

How would we decide in the diseases case whether or not a large proportion of the patients have the disease? We could look at the overall proportion of the patients testing positive for the disease. If we get a large proportion testing positive, we might conclude that a large proportion of the patients have the disease.

Similarly, in the case of theories, we might consider how many successful theories there are amongst the theories accepted in mature science. If science is mostly successful, this might be reason to think that most theories are approximately true.

This is essentially an appeal to the wholesale version of the NMA. That is, the fact that scientists so often succeed in coming up with successful theories is a reason to think that the scientific method generally produces approximately true theories. In section 3.5.2, I argued that the retail NMA could be formulated in probabilistic

terms as $\mathbf{NMA}_{R2}$. I now suggest that the wholesale version $\mathbf{NMA}_W$ takes the same form as the retail version:

(i) $p(S_w|R_w)$ high

(ii) $p(S_w|\neg R_w)$ relatively low

(iii) $p(R_w)$ is not small compared to $p(S_w|\neg R_w)$

Conclusion: $p(R_w|S_w)$ is high.

Here R_w stands for the realist claim that theories accepted in mature science are typically approximately true. It may be seen as the claim that the scientific method typically gives rise to approximately true theories – thus, it is a claim that the scientific method is reliable with respect to approximate truth. S_w stands for the claim that science is on the whole successful – that is, scientists often manage to come up with predictively successful theories.

If this argument succeeds, it makes R_w likely, given overall scientific success S_w. R_w in turn can provide a reason to think that the prior probability of any particular accepted theory is approximately true is reasonably high, since it was produced by a method which tends to produce approximately true theories. Thus it supports premise (iii) of the retail version of the NMA.

But doesn't this just push the problem one step back? We now have to justify premise (iii) in the wholesale argument. Granted, but perhaps this is a premise which the realist could admit to simply assuming. What is needed is that the prior probability allotted to the realist alternative is not very small (compared to the small $p(S_w|\neg R_w)$). That is, the reliability of the scientific method with respect to approximate truth is not extremely unlikely to start off with. If someone refuses that starting point, and insists that the scientific method is very likely to be unreliable, they will not be convinced by the wholesale NMA, and hence retail NMAs, to opt for realism.

Although premise (iii) in either version is needed for the validity of the argument, one might wonder if assumptions about the prior probabilities are really so critical in the NMA. If it is repeated and varied successes that drive the NMA conclusion, not just one-off successful predictions, then it doesn't matter much whether the prior probability is high or low. If we update our prior probability for the realist claim R (or R_w) in the light of success of the theory, it will, as long as the theory continues to predict successfully, get higher and eventually we will end up with a high value for $p(R|S)$ (or $p(R_w|S_w)$). Thus, even someone who starts with a very low prior probability for the realist claim, will potentially end up a realist if enough success is gathered.

The realist will not necessarily want to rely on this convergence of priors, since we don't know how long we will have to wait before the realist claims accrues substantial posterior probability. The strategy risks weakening the realist position since we cannot say confidently at any given moment that the theory is probably approximately true. We can only say that it is getting more likely that it is. Thus most realists will still want to make an argument for premise (iii).

Nonetheless, some anti-realists have found the lack of an in-principle barrier to conversion to realism to be sufficiently concerning that they have adopted a specific strategy for dealing with it. They argue that the anti-realist position is not reflected in assigning a low prior probability to R (and R_w), but rather in adopting an agnostic attitude towards R (and R_w). The anti-realist is skeptical about the reliability of scientific method with respect to finding the truth about unobservable matters. Presumably, any claims about approximate truth must fall into the unobservable category. Therefore, the anti-realist will be agnostic about whether R (and also R_w) is true. In probabilistic terms, it has been proposed that this agnostic attitude should be represented by a 'vague' probability – that is, a probability interval of the

form $[0, x]$, rather than a sharp value (Van Fraassen (1989), p.194). This represents the lack of commitment as to whether the probability is zero, or any other value up to x. The very fact that the anti-realist represents her opinions in this way means that she cannot be moved away from agnosticism by conditionalisation on any evidence, including scientific success. This means that there is no need for the anti-realist to engage with the other premises of the NMA, since she has simply ruled out the possibility that success could be a reason for adopting realism by fiat.

In summary, then, we have seen that premise (iii) of the retail version can be supported by appeal to the wholesale version of the NMA. A higher prior that a particular theory is approximately true may be appropriate if that theory is produced by a method which is reliable in producing approximately true theories. However, it is necessary to assume that it is not extremely unlikely that scientific method is reliable in this way. The antirealist could resist this either directly by denying the assumption, or by refusing to assign R sharp-valued probabilities at all.

Neither of these are perhaps the most effective way for the anti-realist to resist the NMA. It has been more common to take issue with premise (ii) – that is, to deny that the success would be very unlikely if the theory were not approximately true. Anti-realists have suggested alternative explanations of success which are allegedly better than the explanation provided by realism. One suggestion along these lines is van Fraassen's idea that the success of theories can be explained by the way in which they are selected, just as the adaptedness of organisms can be explained by their past history of natural selection (van Fraassen (1980), pp. 39-40). Another is that the success of a theory can be explained by what Leplin calls 'surrealism', the view that the world is just *as if* the theory is true.[6]

[6]This view is outlined and criticised in Leplin (1987), who sees it as inspired by Fine (1986).

In the next section we will see that the allegation that the PI is an instance of the base-rate fallacy is based on taking it to be an argument against premise (ii), whereas I claim that it is supposed to be an argument directly against the conclusion of the NMA.

3.6 The PI and the base-rate fallacy

In Laudan (1981), Larry Laudan critiques No-Miracle style arguments to the effect that realism constitutes the best explanation for the success of science. The most widely discussed section of this paper is that entitled 'Approximate truth and success: the upward path' which presents the argument which has subsequently become known as the 'Pessimistic Induction'. Laudan's declared target in this section is the claim he calls 'T2' : 'if a theory is explanatorily successful, then it is probably approximately true' (Laudan (1981), p.33).[7] He combats this claim with a series of historical examples of theories, which he argues are now regarded as not even approximately true, but which were nonetheless successful at some time in the past. Laudan thinks that no realist would want to say a theory was approximately true if its central theoretical terms failed to refer (Laudan (1981), p. 33), and therefore a number, though not all, of his examples are theories now thought to contain non-referring terms. Among the examples he lists are the crystalline spheres of ancient and medieval astronomy, the phlogiston theory of chemistry and the optical ether. After itemising his list of examples, Laudan says 'I daresay that for every highly successful theory in the past of science which we now believe to be a genuinely

A version of it is endorsed in Kukla and Walmsley (2004).

[7]As we saw above, there may be different versions of what constitutes success of a theory. Laudan poses the argument in terms of explanatory, rather than predictive, success. This will not affect the points I am making here.

referring theory, one could find half a dozen once successful theories which we now regard as substantially non-referring' (Laudan (1981), p. 35). Thus his claim is that the proportion of successful theories in the past which were approximately true is very small. Absent any reason to think we are in a better epistemic situation now than in the past, we might assume the proportion of approximately true theories amongst currently successful theories is also small. This should give the realist pause about inferring from the success of a theory to its truth, according to T2.

Peter Lewis's claim that Laudan's PI is an instance of the base-rate fallacy is based on the following reconstruction of the argument, $\mathbf{PI}_1$:

> (1) Assume that the success of a theory is a reliable test for truth.
>
> (2) Most current scientific theories are successful.
>
> (3) So most current scientific theories are true.
>
> (4) Then most past scientific theories are false, since they differ from current theories in significant ways.
>
> (5) Many of these false past theories were successful.
>
> (6) So the success of a theory is not a reliable test for its truth.
>
> (Lewis (2001), p. 373)

Lewis understands a test to be reliable if it has low error rates, in other words if the rates of false positives and false negatives are both low. In particular, he takes Laudan to be trying to show that the test is in fact unreliable because the false positive rate is high rather than low. This would mean that the proportion of false theories which are successful is high.

Lewis then complains that Laudan's historical evidence does not establish this. Rather, he claims, Laudan's evidence tells us about the proportion of successful theories which are false. Lewis takes Laudan's claim to be that in the past, successful

false theories outnumber successful theories which are approximately true. However, Lewis claims, even if this were established, it is quite consistent with a low error rate. Just as in the disease case, even if the test has a low false positive rate, there could still be more people without the disease than with it amongst those who tested positive. This can happen if the disease is sufficiently rare. Similarly, even if relatively few of the false theories are successful, there could still be more successful false theories than successful approximately true ones. This could happen if the truth is sufficiently rare amongst theories. Thus, according to Lewis, to draw the conclusion that the test has a high error rate and hence is unreliable would be to commit the base-rate fallacy.

To put this in terms of probabilities, for reliability one needs $p(S|\neg T)$ (the false positive rate) and $p(\neg S|T)$ (the false negative rate) to be low. Showing that $p(T|S)$ is low does not mean that $p(S|\neg T)$ is high. Rather it could be that $p(T)$ is particularly low.

In my view, Lewis has not established that the PI involves the base-rate fallacy, because he has incorrectly reconstructed what is going on in Laudan's argument. As we saw, Laudan's ostensible concern is with the success-to-truth inference expressed in T2, not with reliability in Lewis's sense. In fact Lewis accepts that 'Laudan's evidence does indeed undermine (T2) for certain times in the past' (Lewis (2001), p. 377).

A better way to reconstruct the argument is as a reductio not of the reliability of success as a test for truth, as in Lewis's version, but of T2. It may be written as $\mathbf{PI}_2$:

(1) Assume that if a theory is successful, then it is probably approximately true.

(2) Then, there is a high proportion of currently successful theories

which are approximately true.

(3) By the lights of these approximately true current theories, the proportion of past successful theories which are approximately true is low (because in many cases, current successful theories postulate radically different ontology than their predecessors).

(4) By induction from the past to the present, there is a low proportion of current successful theories which are approximately true.

Since (4) contradicts (2), we have a reductio of the assumption that if a theory is successful, then it is probably approximately true. (3) is established by a subargument, which involves giving a series of counterexamples to retail versions of the NMA. For example, the retail NMA for Fresnel's theory concludes that since Fresnel's theory is successful, it is probably approximately true. Laudan argues that if we assume that the current theory of light is approximately true, then there is no mechanical ether. Since this key part of Fresnel's theory was non-referring, Laudan argues that the theory cannot be approximately true, despite its success. His other counterexamples follow a similar pattern. The overall observation is that supposedly there are more examples where the past successful theory is false, than there are where it is approximately true.

This reconstruction is preferable to Lewis's not just because it is more faithful to Laudan's explicitly declared target. It also means that the assumption up for reductio is not a premise of the NMA, but its conclusion. The historical evidence is supposed to tell us about the proportion of successful theories which are approximately true. In probabilistic terms, this bears on $p(T|S)$, the probability that a theory is approximately true, given that it is successful. Thus, the historical examples Laudan presents are supposed to undermine the conclusion of the NMA, which we presented above, by indicating that $p(T|S)$ is low, rather than high. If

Laudan's historical evidence has indeed undermined the conclusion of the NMA, it is unclear why there would be any need to argue backwards from problems with the conclusion to problems with one particular premise.

Realists have resisted the Pessimistic Induction using several strategies. One is to question whether past theories are really false in the light of current theories, or whether common elements between past and current theories can be found. Such common elements can then be regarded as the part of the theory which got things approximately right. There have been different suggestions as to what the common elements are. For some, what is preserved through theory change is 'structure' (Worrall (1989)). For others, it is those theoretical constituents which 'contribute essentially to, or "fuel"' the successes (Psillos (2007), p.110). Either way, this strategy aims to undermine (3) of $\mathbf{PI}_2$. (3) can also be undermined by denying that some of Laudan's examples do consist of genuinely 'successful' past theories.

An alternative type of realist response has been to resist the inductive step from the past to the present (i.e. resist (4) in $\mathbf{PI}_2$). Some realists have suggested that perhaps the inductive base is not big enough, or should be restricted in some way, for instance by confining examples to those of 'mature' sciences. One might also try to find reasons to think that we are in a better epistemic position now than we were in the past.

Other comments that Lewis makes suggest that he is actually advocating some version of this second familar realist strategy. He accepts that Laudan's historical evidence shows that the proportion of successful theories which are approximately true, $p(T|S)$, is low for theories in the past. He then argues that the realist could claim that the difference between the past and present is that in the past false theories vastly outnumber true theories, whereas this is no longer true. In probabilistic terms, this would mean that the proportion of theories $p(T)$ which are approxi-

mately true is very low in the past, meaning that $p(T|S)$ is also low, whereas in the present $p(T)$ is not low and so neither is $p(T|S)$ currently. The real question of course is to address why there is such a change in what proportion of theories are approximately true and Lewis does not offer any explanations here.

Lewis's argument that the PI commits the base-rate fallacy is based on a misrepresentation of the argument. The inference that he claims to be fallacious is in fact not part of the PI. Furthermore, Lewis's recommendations to the realist suggest that he is not so much uncovering a new fault in the logic of the PI, as advocating a version of a familiar realist rejection of one of the premises.

3.7 Conclusion

In this chapter I have argued that both the NMA and the PI may be stated in a way which does not involve the base-rate fallacy. The discussion of the case of the NMA has helped to clarify the relationship between the retail version of the argument, which concerns the status of a particular theory, and the wholesale version, which concerns the reliability of the scientific method overall. The allegation that the NMA is an instance of the base-rate fallacy has been made for the retail version of the argument. The worry is that the realist neglects the very low probability that the theory in question is approximately true. I have argued that the realist does not neglect this low probability, but rather denies that it is particularly low. The reason for the denial is that the scientific method is fairly reliable in finding approximately true theories. The scientific method is taken to be reliable because of the general success of science as a whole. Thus the justification for taking a higher prior in the retail NMA rests on the wholesale NMA. I have suggested that the wholesale NMA has the same form as the retail NMA. Thus, it also needs to be the case that the

prior probability in the wholesale argument is not too low. The wholesale NMA, and hence the retail NMAs also, rest on the assumption that it is not extremely unlikely to start out with that the scientific method reliably finds approximately true theories. Denying this assumption is one way that an anti-realist could resist the NMA. Another is to deny that the prior probabilities are sharp-valued at all.

However, rather than taking issue with the priors in either of these ways, the anti-realist could deny that success is unlikely when the realist claim is false (this is to deny premise (ii) in the NMA). The usual way to do this is to offer alternative explanations for success which don't rely on realist claims. Peter Lewis has been tempted to think that the PI is also an argument against premise (ii) of the NMA. The historical evidence presented in the PI ostensibly provides grounds to reject the conclusion of the NMA. Lewis locates the base-rate fallacy in an inference from the denial of the conclusion of the NMA to the denial of premise (ii). I claim that the PI should simply be seen as rejection of the conclusion of the NMA based on counterexamples. There is no need to see the PI as providing a specific diagnosis of which premise is at fault.

The possibility that the NMA and the PI were based on a common fallacy seemed to make it possible to reject them relatively easily, in a way which would not involve us in the usual travails of engaging with the premises or analysing concepts like approximate truth and success. I have shown that these arguments cannot be dismissed on the grounds of faulty form alone. This should return our focus to the task of deciding whether the premises on which they rest are true.

Bibliography

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

Boyd, R. N. (1983). On the current status of the issue of scientific realism. *Erkenntnis 19*, 45–90.

Carey, S. and E. Spelke (1996). Science and core knowledge. *Phil. Sci. 63*, 515–533.

Copernicus, N. (1878 (1939)). Commentariolus. In E. Rosen (Ed.), *Three Copernican Treatises*, pp. 55–90. New York: Columbia University Press.

Crick, F. H. C. (1958). On protein synthesis. *Symp. Soc. Exp. Biol. XII*, 139–163.

Darwin, C. (1859 (1962)). *The Origin of Species by Means of Natural Selection* (6th ed.). New York: Collier.

Davidson, E. (2006). *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution.* Burlington, MA: Academic Press.

Day, T. and H. Kincaid (1994). Putting inference to the best explanation in its place. *Synthese 98*, 271–295.

Dowe, D. L., S. Gardner, and G. Oppy (2007). Bayes not bust! Why simplicity is no problem for Bayesians. *Brit. J. Phil. Sci. 58*, 709–754.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* Cambridge MA: MIT Press.

Einstein, A. (1905). On the electrodynamics of moving bodies. *Ann. der Physik XVII*, 891–921.

Fine, A. (1986). Unnatural attitudes: Realist and instrumentalist attachments to science. *Mind 95*, 149–179.

Forster, M. (1995). Bayes and bust: Simplicity as a problem for a probabilist's approach to confirmation. *Brit. J. Phil. Sci. 46*, 399–424.

Forster, M. R. and E. Sober (1994). How to tell when simpler, more unified, or less ad hoc theories provide more accurate predictions. *BJPS 45*, 1–35.

Friedman, M. (1974). Explanation and scientific understanding. *J. Phil. 71*, 5–19.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis.* Chapman and Hall.

Giere, R. N. (1996). The scientist as adult. *Phil. Sci. 63*, 538–541.

Gigerenzer, G. and U. Hoffrage (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psych. Rev. 102*, 684–704.

Gillies, D. (2001). Bayesianism and the fixity of the theoretical framework. In D. Corfield and J. Williamson (Eds.), *Foundations of Bayesianism*, pp. 363–380. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Glymour, C. (1980a). Explanations, tests, unity and necessity. *Nous 14*, 31–50.

Glymour, C. (1980b). *Theory and Evidence.* New Jersey: Princeton University Press.

Glymour, C. (1985). Explanation and realism. In P. Churchland (Ed.), *Images of Science*, pp. 173–192. Chicago: University of Chicago Press.

Godfrey-Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science.* University of Chicago Press.

Gopnik, A. (1996). The scientist as child. *Phil. Sci. 63*, 485–514.

Griffiths, T. and J. B. Tenenbaum (2007). Two proposals for causal grammars. In A. Gopnik and L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy and Computation.* Oxford University Press.

Harman, G. (1965). The inference to the best explanation. *Phil. Rev. 74*, 88–95.

Hempel, C. G. (1958). The theoretician's dilemma. In H. Feigl, M. Scriven, and G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science II: Concepts, Theories and the Mind-Body Problem*, pp. 37–99. Minneapolis: University of Minnesota Press.

Hempel, C. G. (1965). *Aspects of Scientific Explanation.* New York: Free Press.

Hempel, C. G. and P. Oppenheim (1948). Studies in the logic of explanation. *Phil. Sci. 15*, 135–175.

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

Howson, C. and P. Urbach (2006). *Scientific Reasoning: The Bayesian Approach.* Chicago and La Salle, Illinois: Open Court.

Huemer, M. (2009). Explanationist aid for the theory of inductive logic. *Brit. J. Phil. Sci. 60*, 345–375.

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

Jefferys, W. H. and J. O. Berger (1992). Ockham's razor and Bayesian analysis. *Amer. Scientist 80*, 64–72.

Jeffreys, H. (1961). *Theory of Probability.* London: Oxford University Press.

Jeffreys, H. (1998). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press. First edition 1939.

Kahneman, D. and A. Tversky (1972). On prediction and judgment. *Oregon Research Bulletin 12.*

Kahneman, D. and A. Tversky (1973). On the psychology of prediction. *Psych. Rev. 80*, 237–251.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science 4*, 188–234.

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. *J. Am. Stat. Ass. 91*, 1343–1370.

Kemp, C. (2007). *The Acquisition of Inductive Constraints.* Ph. D. thesis, MIT.

Kemp, C., T. L. Griffiths, and J. B. Tenenbaum (2004). Discovering latent classes in relational data. *MIT AI Memo 2004-019.*

Kemp, C. and J. B. Tenenbaum (2008). The discovery of structural form. *Proc. Nat. Acad. Sci. 105*, 10687–10692.

Kitcher, P. (1976). Explanation, conjunction and unification. *J. Phil. 73*, 207–212.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher and W. Salmon (Eds.), *Minnesota Studies in the Philosophy of Science: Scientific Explanation*, Volume 13, pp. 410–505. University of Minnesota Press.

Krynski, T. R. and J. B. Tenenbaum (2007). The role of causality in judgment under uncertainty. *J. Expt. Psych. 136*, 430–450.

Kuhn, T. S. (1962/1996). *The Structure of Scientific Revolutions* (3rd ed.). Chicago/London: University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, value judgment and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, pp. 320–339. Chicago: University of Chicago Press.

Kukla, A. and J. Walmsley (2004). Predictive success does not warrant belief in unobservables. In C. Hitchcock (Ed.), *Contemporary Debates in Philosophy of Science*, pp. 133–148. Oxford: Blackwell.

Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worrall and G. Currie (Eds.), *The Methodology of Scientific Research Programmes*, pp. 8–101. Cambridge, UK: Cambridge University Press.

Laudan, L. (1978). *Progress and its Problems.* Berkeley: University of California Press.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science 49*, 19–49.

Lavoisier, A. (1783). Réflexions sur le phlogistique. *Mémoires de l'Académie des Sciences*, 505–538.

Leplin, J. (1987). Surrealism. *Mind 96*, 519–524.

Lewis, D. (1986). A subjectivist's guide to objective chance. In *Philosophical Papers Volume II*, Chapter 19, pp. 83–113. Oxford UK: Oxford University Press.

Lewis, P. J. (2001). Why the pessimistic induction is a fallacy. *Synthese 129*, 371–380.

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). London, UK.: Routledge.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms.* Cambridge: Cambridge University Press.

Magnus, P. D. and C. Callender (2004). Realist ennui and the base-rate fallacy. *Phil. Sci. 71*, 320–338.

Mansinghka, V. K., C. Kemp, J. B. Tenenbaum, and T. Griffiths (2006). Structured priors for structure learning. *Proceedings of the 22nd conference on uncertainty in Artificial Intelligence.*

McMullin, E. (2008). The virtues of a good theory. In S. Psillos and M. Curd (Eds.), *The Routledge Companion to the Philosophy of Science*, pp. 498–508. New York: Routledge.

Myrvold, W. (1996). Bayesianism and diverse evidence: A reply to Andrew Wayne. *Phil. Sci. 63*, 661–665.

Myrvold, W. (2003). A Bayesian account of the virtue of unification. *Phil. Sci. 70*, 399–423.

Myung, I. J., V. Balasubramanian, and M. A. Pitt (2000). Counting probability distributions: Differential geometry and model selection. *PNAS 97*, 11170–11175.

Newton, I. (1687). *The Principia: Mathematical Principles of Natural Philosophy.* Berkeley: University of California Press.

Niiniluoto, I. (1998). Truthlikeness: The third phase. *Brit. Jour. Phil. Sci. 49*, 1–31.

Niiniluoto, I. (1999). Defending abduction. *Phil. Sci. 66*, S436–S451.

Okasha, S. (2000). Van Fraassen's critique of inference to the best explanation. *Stud. Hist. Phil. Sci.*, 691–710.

Popper, K. R. (1934/1959). *The Logic of Scientific Discovery.* London, UK: Hutchinson and co. First edition 1934.

Popper, K. R. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* London, UK: Routledge and Kegan Paul.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth.* New York: Routledge.

Psillos, S. (2007). The fine structure of inference to the best explanation. *Phil. and Phen. Res. LXXIV*, 441–448.

Putnam, H. (1975). *Mathematics, Matter and Method.* Cambridge: Cambridge University Press.

Quine, W. V. (1970/1986). *Philosophy of Logic.* Cambridge, MA: Harvard University Press.

Rheticus, G. J. (1539). Narratio prima. In E. Rosen (Ed.), *Three Copernican Treatises*, pp. 107–196. New York: Columbia University Press.

Rosen, G. (1994). What is constructive empiricism? *Phil. Stud. 74*, 143–178.

Rosenkrantz, R. D. (1977). *Inference, Method and Decision: Towards a Bayesian Philosophy of Science.* Dordrecht, Holland / Boston, USA: Synthese Library.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Schupbach, J. N. (2005). On a Bayesian analysis of the virtue of unification. *Phil. Sci. 72*, 594–607.

Seidenfeld, T. (1979). Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz. *Theory and Decision 11*, 413–440.

Sober, E. (1990). Let's razor Ockham's razor. In D. Knowles (Ed.), *Explanation and its Limits*, pp. 73–93. UK: Cambridge University Press.

Strevens, M. (2004). Bayesian confirmation theory: Inductive logic, or mere inductive framework? *Synthese 141*, 365–379.

Tenenbaum, J. B., T. L. Griffiths, and S. Nigoyi (2007). Intuitive theories as grammars for causal inference. In A. Gopnik and L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy and Computation.* Oxford University Press.

Tenenbaum, J. B. and S. Nigoyi (2003). Learning causal laws. *Proc. 25th Annual Conf. of Cog. Sci. Soc.*, 1152–1157.

Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *J. Phil. 75*, 76–92.

Toulmin, S. E. (1963). *Foresight and Understanding.* New York: Harper and Row.

Toulmin, S. E., R. Rieke, and A. Janik (1984). *An Introduction to Reasoning.* New York: Prentice-Hall.

van Fraassen, B. C. (1980). *The Scientific Image.* Oxford: Oxford University Press.

Van Fraassen, B. C. (1989). *Laws and Symmetry.* Oxford, UK: Clarendon Press.

Wade, M. J. (1978). A critical review of the models of group selection. *Quarterly Rev. Biol. 53*, 101–114.

Weisberg, J. (2009). Locating IBE in the Bayesian framework. *Synthese 167*, 125–144.

Wellman, H. M. and S. A. Gelman (1992). Cognitive development: Foundational theories of core domains. *Annu. Rev. Psychol. 43*, 337–375.

White, R. (2005). Why favour simplicity? *Analysis 65.3*, 205–210.

Williams, G. C. (1966). *Adaptation and Natural Selection: A Critique of some Current Evolutionary Thought.* New Jersey: Princeton University Press.

Woodward, J. (2003). Scientific explanation. *Stanford encyclopedia of philosophy.*

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica 43*, 99–124.

Worrall, J. (2005). Miracles, pessimism and scientific realism. *LSE webpage.*