Inference to the Best Explanation, Bayesianism and the problem of logical constraints

Leah Henderson

Abstract

Many philosophical accounts of scientific theory comparison take as a starting point competition between mutually exclusive alternative hypotheses. However, in scientific inquiry, it often appears that hypotheses which are in competition with one another are not mutually exclusive. For example, a hypothesis which postulates just one cause of a particular event may compete with a hypothesis which postulates a conjunction of causes. It appears that the conjunctive hypothesis does not exclude the single-cause hypothesis, but rather entails it, since the single-cause hypothesis may be seen as a special case of the conjunctive hypothesis. The apparent existence of logical relations between competing hypotheses then presents a problem for models of scientific inference which assume that competing theories are mutually exclusive. I call this the 'problem of logical constraints'. The problem has been raised in slightly different guises for both for Inference to the Best Explanation and Bayesianism. In this paper, I show how taking a hierarchical view of theory comparison allows us to resolve this problem. Scientific theory evaluation takes place at multiple levels, with more general theories competing against each other at higher levels and more specific hypotheses competing at lower levels. Higher-level theories can be seen as mutually exclusive alternatives, even while logical relations are respected at lower levels.

This is the penultimate version to appear in Conjunctive Explanations: New Essays on the Nature, Epistemology, and Psychology of Explanatory Multiplicity, *ed. Jonah Schupbach and David Glass, Routledge.*

1 Introduction

Many philosophical accounts of scientific theory comparison take as a starting point competition between mutually exclusive alternative hypotheses. However, in scientific inquiry, it often appears that hypotheses which are in competition with one another are not mutually exclusive. For example, a hypothesis which postulates just one cause of a particular event may compete with a hypothesis which postulates a conjunction of causes. It appears that the conjunctive hypothesis does not exclude the single-cause hypothesis, but rather entails it, since the single-cause hypothesis may be seen as a special case of the conjunctive hypothesis. The apparent existence of logical relations between competing hypotheses then presents a problem for models of scientific inference which assume that competing theories are mutually exclusive. I call this the 'problem of logical constraints'. The problem has been raised in slightly different guises for both for Inference to the Best Explanation (Schupbach and Glass, 2017) and for Bayesianism (Popper, 1959; Forster and Sober, 1994; Elliott Sober, 2015). Broadly speaking, to resolve the tension, there are two approaches we can take:

1. We can accept that competing theories can be logically compatible with each other, and either abandon existing models of scientific inference or extend or reframe them to account for competition between non-mutually exclusive hypotheses.

2. We can argue that, despite appearances, the competing theories in scientific practice really can be seen as mutually exclusive. In this case, existing models of scientific inference are adequate as they stand.

In this paper, I will argue for the latter approach. I will not argue directly against the first approach, but if my argument that the competing theories really are mutually exclusive succeeds, then the first approach becomes unnecessary. My approach will be based on the recognition that scientific theory evaluation takes place at multiple levels, with more general theories competing against each other at higher levels and more specific hypotheses competing at lower levels. I argue that according to a reasonable conception of higher level theories, they can be seen as mutually exclusive alternatives, even while logical relations are respected at the lower level.

The plan for the paper is the following. I will first explain the problem of logical constraints as it has been raised for both IBE and for Bayesianism (section 2). In section 3, I briefly outline solutions which take the first approach of accepting that competing theories can be logically compatible. In 4, I outline the solution that I favour, based on the second approach above. I explain how this approach makes use of the hierarchical picture of theory comparison. In sections 5 and 6, I show how this approach solves the problem of logical constraints for IBE and for Bayesianism respectively.

2 The problem of logical constraints

In scientific inquiry, we often see cases in which a hypothesis postulating one cause of a particular event competes with a hypothesis which postulates a conjunction of causes. Schupbach and Glass give a helpful example from paleontology (Schupbach and Glass, 2017). Scientists have considered different hypotheses about the cause of the mass extinction at the Cretaceous-Paleogene boundary about 65 million years ago that wiped out the dinosaurs. One influential hypothesis is that the extinction was caused by a bolide impact. Evidence for this ‘impact hypothesis’ includes an unusual layer of clay at that boundary with an anomalously high level of iridium, an element which is not usually so common on Earth, but which is abundant in meteorites and other bolides (Alvarez, 1983). Other scientists have proposed conjunctive explanations invoking multiple causes. For example it has been suggested that the extinction event was caused by climate changes resulting from massive volcanic activity, in combination with and perhaps exacerbated by the bolide impact (Keller, 2014). Thus, part of the scientific inquiry has involved considering one-cause explanations as rivals to hypotheses involving conjunctions of causes.

There may also be cases where one hypothesis competes against a disjunction of hypotheses. For example, according to Aristotelian theories, living organisms could arise either as the result of generation from parent organism(s), or by spontaneous generation from inanimate matter such as earth and water (Lehoux, 2017; Zwier, 2018). Aristotle thought that some organisms, in particular some small fish, eels and barnacles, do not arise from living parent organisms, but are instead spontaneously generated from inanimate materials. A series of experiments in the 17th through 19th centuries eventually convinced scientists that spontaneous generation does not occur. Although living creatures like maggots or worms could be observed to appear, apparently spontaneously, on meat, or in a broth, when an effort was made to isolate the medium from all possible sources of contamination by living organisms, the production of living creatures was no longer observed. Thus, the general Aristotelian hypothesis that living organisms could be produced *either* from other living organisms, *or* by spontaneous generation, was replaced by a single-cause explanation: living organisms could only be produced from other living organisms.

These kinds of examples have been used to argue that scientific inference can involve competition between logically compatible hypotheses. In our first example, it is possible for both bolide impact and volcanic activity to have had a causal effect on the extinction, so these hypotheses appear not to be mutually exclusive. In fact, the conjunctive hypothesis might be taken to entail the single-cause hypothesis. In the second example, the disjunctive hypothesis may be taken to be entailed by the single-cause hypothesis. If such entailments hold, then the competing hypotheses are logically consistent with one another. This has been raised as a puzzle for IBE, since IBE, like other theories of scientific inference, is often cast as involving a competition between mutually exclusive alternatives (Schupbach, 2019). If scientific inference really does involve competition between logically compatible hypotheses, the question is how this can be handled by IBE.

The possibility of apparent logical relations between competing hypotheses also raises a related problem for Bayesianism. In this context, the problem arises from the observation that logical relations should constrain probabilistic relations. Let h_1 and h_2 be two specific hypotheses where h_1 entails h_2. According to the probability calculus $p(h_1) \leq p(h_2)$. The same inequality holds also for conditional probabilities, thus we have the following constraint also on the posterior probabilities

$$p(h_1|D) \leq p(h_2|D)$$

Such a constraint means that one should never give a higher probability to h_1 than to h_2. That is, a logically stronger hypothesis should not get a higher probability. Yet in scientific practice, it seems to be quite common for such a preference to be manifested. Furthermore, in the practice of Bayesian model comparison, such preferences are apparently permitted. For example, the problem has been raised for the case of curve-fitting. Suppose we measure the relationship between two variables X and Y. Here X might be the period, and Y might be the length of a pendulum of fixed mass. Suppose our data consists of pairs of observations, where the first is an observation of period and the second a measurement of length. We would like to discover which kind of 'model' best accounts for the data. Should it be a linear model comprising all curves of the form $y = \alpha_0 + \alpha_1 x$ (we denote this as LIN), or a parabolic model comprising all curves of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (we denote this as PAR)? LIN can be regarded as a conjunction of PAR and $\beta_2 = 0$. For each specific curve in PAR, if we set the adjustable parameter for the quadratic term β_2 to zero, we get a linear curve. LIN is clearly a subset of PAR – or, in other words, LIN entails PAR. Since LIN entails PAR, it follows that $p(\text{LIN}) \leq p(\text{PAR})$. Thus it appears that probabilistic comparisons cannot ever favour LIN. Yet LIN is a simpler and more falsifiable hypothesis than PAR, and in practice a scientist would often choose the linear curve if it fits the data adequately.[1]

[1] It is also worth noting that people have a tendency to violate logical constraints in the probabilities that they assign to particular hypotheses. This is the well-known 'conjunction fallacy' (Kahneman et al., 1982). In a famous experiment, people were found to have a tendency to attribute a higher probability, given certain information about Linda, to the proposition that h_1 that 'Linda is a feminist bank-teller' than to the proposition h_2 that 'Linda is a bank-teller', even though h_1 entails h_2. However, people also recognise that they have made an error once it is pointed out, and when the problem is formulated differently – in terms of frequencies, rather than judgments of likelihood – the tendency to commit the fallacy disappears (Gigerenzer, 1991).

3 Approach 1: accept non-mutually exclusive competitors

Some of the solutions to the problem of logical constraints both for IBE and for Bayesianism are based on the acceptance that examples like those described do genuinely show that competing theories can be logically compatible. Various proposals have then been made regarding the implications of this for theories of scientific inference like IBE and Bayesianism. As we shall see, these solutions include abandoning the theories of scientific inference altogether, drastically reducing their scope, or finding ways to show that they can indeed be applied to non-mutually exclusive competing hypotheses. I will now give a brief outline of some of these suggestions.

3.1 Proposed solutions for IBE

3.1.1 Restrict application of IBE to mutually exclusive hypotheses

In the case of IBE, one proposal has been to restrict application of the inference only to mutually exclusive hypotheses. This is suggested by Lipton who says

> [IBE] is meant to tell us something about how we choose between *competing* explanations: we are to choose the best of these. But among compatible explanations we need not choose. (Lipton (2001), p. 104)

However, Lipton's proposal has been criticised by Schupbach on the grounds that it seems to rule out too much. Schupbach argues that 'Many (indeed plausibly *most*) canonical instances of IBE compare potential explanations that are compatible with one another' (Schupbach (2019), p. 147). If this is so, then Lipton's manoeuvre would amount to a very significant restriction on the scope of IBE.

3.1.2 Dissolving the difficulty

Instead of restricting IBE, Schupbach argues for a different solution. He accepts that logically compatible hypotheses are compared in scientific inference[2] but argues that IBE can be framed in a way that allows it to deal with compatible hypotheses. Schupbach allows that the set of potential hypotheses may contain various logically compatible hypotheses, including conjunctions and disjunctions of individual hypotheses also under consideration. However, he suggests that IBE be regarded as choosing which hypothesis is 'explanatorily best', and this can be done among a set or 'lot' of hypotheses where some are compatible with each other:

> the present proposal amounts to thinking of the lot of potential explanations as the set containing the[se] considered hypotheses along with their Boolean combinations. What matters is which combination of considered hypotheses best explains the explanandum, not what logical form the various options take (Schupbach (2019), p. 160).

Thus, according to this proposal, we can apply IBE to logically compatible hypotheses, because the logical relations between hypotheses can be ignored when determining the best explanation.

[2] In another paper, he devises a probabilistic account of how non-mutually exclusive hypothese may compete (Schupbach and Glass, 2017).

3.2 Proposed solutions for Bayesianism

3.2.1 Restrict application of Bayesian comparison to disjoint sets

Just as for IBE, in the case of Bayesianism, one proposal has been to restrict application of the inference. It has been suggested to apply Bayesian model selection only to mutually exclusive hypotheses. In a case like curve-fitting then, this would mean adjusting the hypotheses. Rather than LIN and PAR, the two hypotheses to compare would be LIN and PAR*, where PAR* is the set of all specific curves with a genuinely non-trivial quadratic term: $y = \beta_0 + \beta_1 x + \beta_2 x^2$, where $\beta_2 \neq 0$ (Howson, 1988).

Just as in the case of IBE, many have seen this as an unwelcome restriction of Bayesianism's application. Although it solves the problem in some sense, it appears to be a rather artificial solution, since it does away by fiat with problems that scientists or statisticians are actually interested in. Several authors argue that this solution effectively amounts to changing the subject. For instance, Forster and Sober say 'this ad hoc maneuver does not address the problem of comparing (LIN) versus (PAR), but merely changes the subject.' (Forster and Sober (1994), p. 23). Bengt Autzen says that since people making use of the Bayesian model selection methodology 'are genuinely interested in comparing models with non-trivially overlapping parameter ranges, restricting the Bayesian analysis to models with non-overlapping parameter ranges amounts to substantively changing the inference problem.' (Autzen (2019), p. 326).

Another approach is suggested in (Romeijn and Schoot, 2008). This is based on relabelling the nested sets of hypotheses.

> nothing prevents us from using two distinct sets of hypotheses ... which are different from a set-theoretical point of view by virtue of being labeled differently, even while they have the same likelihood functions over the data. (Romeijn and Schoot (2008), p. 353)

The relabelling strategy maintains the set-based view of statistical models, but the claim is that we could relabel LIN and PAR, for example, such that they become disjoint sets.

Autzen has criticised this relabelling view on the grounds that the relabelled sets still have the same likelihood functions as the original sets, thus

> simple relabelling seems to amount to a case of mislabelling. Assigning different labels to sets of pairwise identical probability distributions, gives a misleading picture regarding the possible hypotheses about the data generating mechanism. (Autzen (2019), p. 329)

Autzen is suggesting that multiplying labels creates the impression that there are different data generating mechanisms, whereas in fact there are not. The basic problem here is that although the problem of logical constraints can be technically solved in this way, it remains unclear what the independent reasons would be for doing the relabelling.

3.2.2 Abandon Bayesianism

Some have turned to even more drastic solutions and have seen the problem of logical constraints as a reason to abandon a Bayesian approach to theory comparison. Karl Popper, for example, emphasised the importance of falsifiability in theory choice, where falsifiability often tracks simplicity or informative content. He saw the problem of logical constraints as a reason to think that ‘the scientist does not and cannot aim at a high degree of probability’ (Popper (1959), p. 400). Popper then resisted attempts to characterise scientific theory preferences in terms of probabilities. Others have seen the problem as a reason to resist the Bayesian approach to model selection. For example, Forster and Sober argue that Bayesians are, in cases like curve-fitting, unable to explain why scientists sometimes prefer LIN over PAR. They favour instead a non-Bayesian approach to model selection – particularly recommending the methods based on the Akaike Information Criterion (Forster and Sober, 1994). This approach has tended to be attractive to those who already have other reasons to be uncomfortable with Bayesian methodology – such as the general problem of assigning priors.

There are some significant problems with this approach. Giving up on Bayesian methods means giving up on a methodology which has in practice been very successful in a number of domains. Moreover, there are close connections between the Bayesian approach to model selection and non-Bayesian methods (Claeskens and Hjort, 2008; Grünwald and Roos, 2019), which gives support to the idea that Bayesian methodology is not fundamentally flawed. Furthermore, as we will see in section 6, non-Bayesian methods do not evade the problem entirely. Non-Bayesians have their own problems with justifying the scientific preferences we see in cases like curve-fitting.

4 Approach 2: maintain mutual exclusivity

I will not attempt to respond in detail to the above suggestions, but will rather present an alternative approach. My response to the problem of logical constraints will be based on the idea that it is possible to maintain, despite the appearances of the above examples, that competing theories in scientific inference are mutually exclusive. If true, this would solve the problem of logical constraints both for IBE and for Bayesianism. If the competing theories are mutually exclusive, then they can be compared according to a standard understanding of IBE. Furthermore, there are no logical relations which constrain probability assignments to the hypotheses, so the problem for Bayesianism also disappears. My approach here elaborates and generalises a solution already sketched for a particular example in Henderson et al. (2010).

My argument will be developed in light of a hierarchical view of how scientific theories are compared. The general idea is that scientific theories can be regarded as hierarchically structured with more general or abstract ‘framework’ theories at higher level, and more specific or concrete hypotheses at lower levels. Theory comparison then takes place at multiple levels, and at each level the competing hypotheses are mutually exclusive. I will suggest that the kinds of examples that gave rise to the problem of logical constraints are ones in which higher-level theories are competing. If one identifies these higher-level theories with sets of lower level theories, they can appear to be non-mutually exclusive. However, I will argue that this identification should not be made, and the higher-level theories can be seen as genuinely mutually exclusive alternatives.

I will first provide a brief outline of the hierarchical view of scientific theory comparison. The recognition of the point that scientific theories are hierarchically structured has been a common theme in historically inspired accounts of theory change ((Kuhn, 1962; Laudan, 1977; Lakatos, 1978)). According to the hierarchical view, we may distinguish between general theories, which we will denote using upper case T, and more specific hypotheses, which we will denote as lower case h. The general theories amount to something like a schema or framework. For example, in the comparison between geocentric and heliocentric models of the planetary system in the time of Copernicus, we might consider a general heliocentric model which places the sun at the centre of the planetary system as constituting a general schema T_{Hel}. Another schema would be a geocentric model which places the Earth at the centre T_{Geo} (Henderson, 2014). Each of these schemas contains a number of details and parameters which are not yet filled in: for example, the number of planets, the radii and periods of the orbits, etc. By filling in these details, we obtain a specific hypothesis h which instantiates the general schema. Given certain assumptions, the theory schemas can be said to 'generate' sets of specific hypotheses. For example, the Copernican schema generates a set of possible specific Copernican models.

When we ask for the best explanation of some phenomena, we are often effectively asking which of two general schemas provides the best explanation, rather than which of two specific hypotheses does so. We can ask, for instance, whether phenomena like retrograde motion of the planets is better explained by a heliocentric model or by a geocentric model, without yet getting into the details of delineating specific periods of orbits, etc. Of course, it is also possible to deploy IBE at the level of specific hypotheses also, but this is often done within the general framework provided by an accepted schema.

Bayesian comparison can also be applied not only to specific hypotheses but also to competing general theories or schemas (Henderson et al., 2010). When Bayesian comparison is applied to competing schemas $\{T_i\}$, these are assigned prior probabilities $p(T_i)$, and then updated by Bayesian conditionalisation, given the evidence D. This results in posterior probabilities given by Bayes' rule as

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

A general theory T_i may have adjustable parameters which we denote by a vector $\tilde{\theta}$. Then, in the equation 1, $p(D|T_{\mathrm{i}})$ is a 'marginal likelihood', obtained by integrating over the likelihoods for all the specific hypotheses allowed by the general theory

$$p(D|T_i) = \int p(D|\widetilde{\theta})p(\widetilde{\theta}|T_i)d\widetilde{\theta} \tag{2}$$

Here $p(\widetilde{\theta}|T_i)$ is the prior over the adjustable parameters, given a particular theory T_i. In this methodology, the competing general theories are usually treated as mutually exclusive alternatives.

A Bayesian may also compare the specific hypotheses given by particular choices of parameter values, given a particular theory schema. This is done by a Bayesian update on the prior for the parameters $p(\widetilde{\theta}|T_i)$ to the posterior given by

$$p(\widetilde{\theta}|T_i, D) = \frac{p(D|\widetilde{\theta}, T_i)p(\widetilde{\theta}|T_i)}{p(D|T_i)}$$

Thus, we can have Bayesian evaluation at two levels – that of the general theory schema, and that of the specific hypotheses within a certain schema (Henderson et al., 2010).

If the general theory specifies not a deterministic, but a probabilistic relationship between the variables, it may constitute a 'statistical model'. A statistical model is a mathematical model which tells us about the process by which the data is generated. A simple example of a statistical model is the binomial model. Suppose we have a simple system, like a coin, which may give one of two possible outcomes in an experiment. A coin may land heads or it may land tails when it is tossed, for instance. Then if we assume that there is a fixed chance q that the coin lands heads, the probability of throwing n heads in a series of N tosses is given by the binomial distribution

$$p(n) = \frac{N!}{n!(N-n)!}q^n(1-q)^{N-n} \tag{3}$$

We say that the data is generated by a binomial model $B(N, q)$, which has parameters N and q. The data regarding the number of heads thrown is then distributed according to an equation of the form 3.

We have distinguished above between the comparison of higher-level theory schemas, and the comparison of particular hypotheses within a schema. Statisticians also distinguish between the task of finding the right model, and finding the best specific hypotheses within a model. The task of finding the right model is called 'model selection', and it contrasts with 'parameter-learning', which involves fitting parameters to a particular model. There are a number of approaches to statistical model selection – employing both non-Bayesian and Bayesian methodologies (J. Friedman, Hastie, Tibshirani, et al., 2001; Grünwald, 2007; Claeskens and Hjort, 2008). The Bayesian approach involves following the same procedure as we saw above for general theory schemas. We take a hypothesis space of different candidate models $\{\mathcal{M}_i\}$ and compute the posterior probabilities

$$p(\mathcal{M}_i|D) = \frac{p(D|\mathcal{M}_i)p(\mathcal{M}_i)}{p(D)} \tag{4}$$

Marginal likelihoods for the models are again computed as in equation 2:

$$p(D|\mathcal{M}_i) = \int p(D|\widetilde{\theta})p(\widetilde{\theta}|\mathcal{M}_i)d\widetilde{\theta} \tag{5}$$

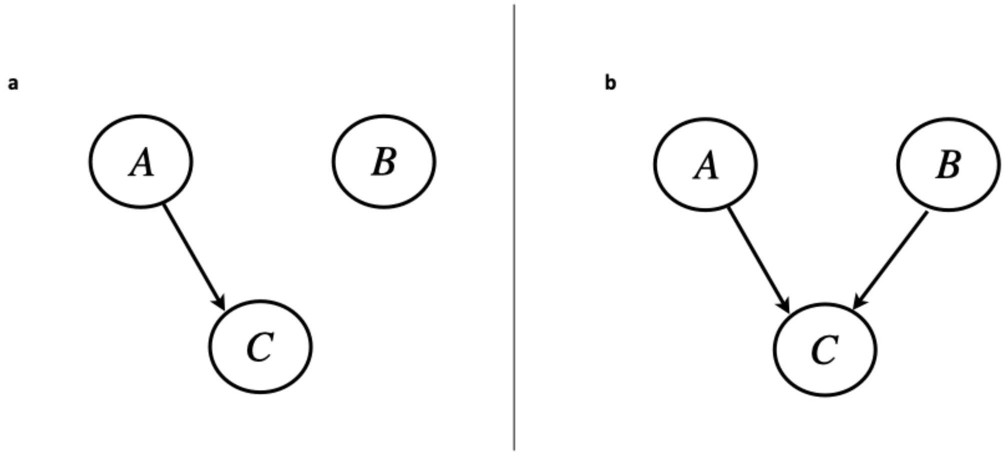
Bayesian model selection is a method extensively used in the applied sciences (for examples, see Jefferys and Berger (1992), Griffiths, Kemp, and Tenenbaum (2008), and Gelman et al. (2013)).

In previous work, I have argued that the hierarchical view of theory comparison allows us to connect IBE and Bayesianism (Henderson, 2014; Henderson, 2017). When IBE involves comparison of general theory schemas or models, rather than specific hypotheses, what is generally valued as explanatory is the ability of a general schema to account for the data on the basis of its core principles without relying too heavily on special choices of auxiliary hypotheses or parameters. This is often expressed by saying that the explanation provided is 'simpler' or 'more unified'. Bayesian methods applied to theory schemas also effectively penalise schemas which are non-explanatory in this sense, since such fine-tuning tends to reduce the marginal likelihood of a model (other things being equal). This occurs via the marginal likelihood in equation 2. Given natural choices of priors over the adjustable parameters, the marginal likelihood effectively penalises theory schemas or models which only fit the data in a small range of parameter values (Henderson, 2014). This is the well-known 'Bayesian Occam's razor' effect (Jefferys and Berger, 1992; MacKay, 2003). Thus, the key considerations which go into IBE are reflected in Bayesian calculations, and the two approaches to theory comparison should be regarded as compatible with one another. In fact, according to the view which I have called 'emergent compatibilism', IBE can be explicated in Bayesian terms (Henderson, 2014; Henderson, 2017). This close relation between IBE and Bayesianism makes it not unexpected that the solution to the problem of logical constraints is essentially the same in both cases. The hierarchical picture of theory comparison sketched here is key to the solution of the problem in both guises.

5 Solution for IBE

As we have seen, the problem of logical constraints has been raised for IBE as the concern that the competing hypotheses may not be mutually exclusive. In this section, I will use the framework of

Figure 1: (a) Graph 1: A has a causal influence on C, but B does not. (b) Graph 2: A and B both have a causal influence on C.

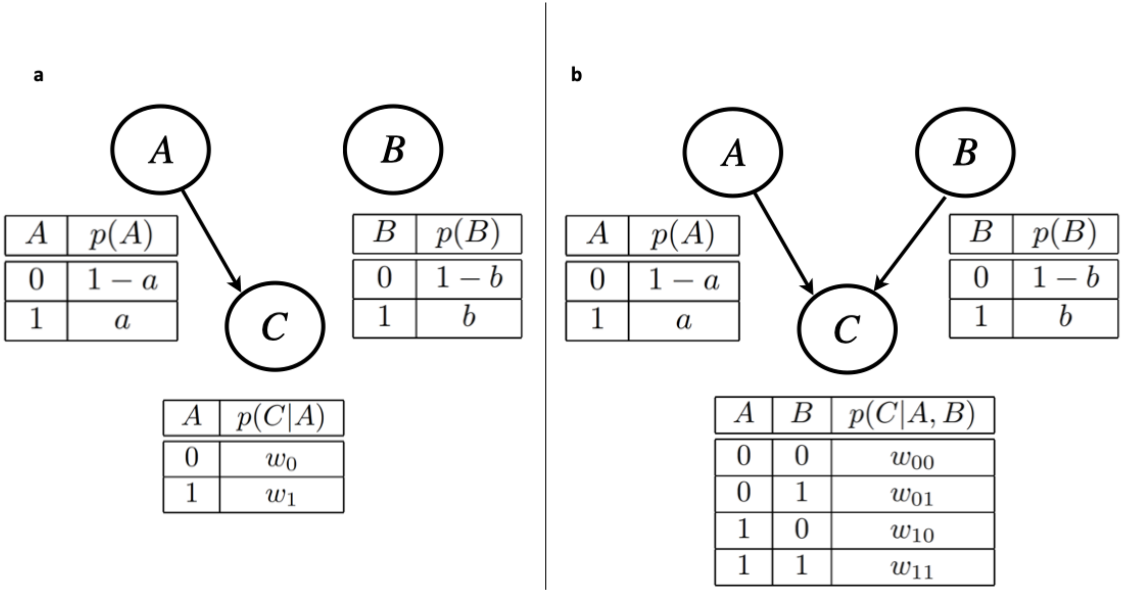


causal graph theory to formalise the hypotheses which are under comparison in causal examples like those described in section 2. When this is done, it becomes clear that the competing hypotheses are mutually exclusive after all.

Causal graph theory is a well-established formalism for representing hypotheses about causal relationships between variables (Spirtes et al., 2000; Pearl, 2009). In this formalism, a causal graph is used to represent the causal structure relating a set of variables. In such a 'Directed Acyclic Graph' or 'DAG', the nodes represent variables, and arrows between the nodes represent causal relations between the variables. The graph must be 'acyclic', meaning that it is not possible to go in a cycle by following arrows. As an example of how DAGs can represent causal structure, consider the two graphs shown in Figure 1. In both cases, there are three variables A, B and C. In Graph 1, variable A has a causal influence on variable C, but there are no causal relations between B and the other variables. In Graph 2, both A and B have a causal influence on C. When there is a causal arrow from a variable A to a variable C, we say that A is a 'parent' of C. Any variables which can be reached from A by a directed path of arrows are called 'descendants' of A.

Different causal structures can be expected to produce different data, where the data may consist either of observations of correlations between variables, or results of interventions where one or more variables is set to a particular value and the values of the other variables observed. The connection between a particular causal graph and the expected probability distribution over the variables $\{X_i\}$ is made using the Causal Markov Condition. The Causal Markov Condition is the assumption that each variable X_i is probabilistically independent of all its non-descendants, given its parents. Thus the causal graph tells us about the causal relations and which variables are probabilistically independent of which. Without further information, however, it does not tell us about the exact relation between the causal relations. For example, graph B could be used to represent either a conjunctive causal structure, where both causes A and B are needed to produce the effect C, or a disjunctive causal structure where either A or B is needed to produce C. Each of those possibilities would be associated with a different specific probability distribution over the variables. A given causal graph is compatible with a number of different specific probability distributions which satisfy

Figure 2: Parametrised graphs. (a) Graph 1 with parameters $\{a, b, w_0, w_1\}$. (b) Graph 2 with parameters $\{a, b, w_{00}, w_{01}, w_{10}, w_{11}\}$.



A	$p(A)$
0	$1-a$
1	a

B	$p(B)$
0	$1-b$
1	b

A	$p(C\|A)$
0	w_0
1	w_1

A	$p(A)$
0	$1-a$
1	a

B	$p(B)$
0	$1-b$
1	b

A	B	$p(C\|A, B)$
0	0	w_{00}
0	1	w_{01}
1	0	w_{10}
1	1	w_{11}

the independence relations given by the Causal Markov Condition.

However, a particular probability distribution is specified, once we are given what is known as the 'parameters of the graph'. Given the Causal Markov Condition, the probability distribution over $\{X_i\}$ factorises as

$$p(X_1, X_2, ..., X_n) = \prod_i p(X_i|Parent(X_i))$$

where $Parent(X_i)$ is the set of parents of X_i. In order to have the full probability distribution $p(X_1, X_2, ..., X_n)$ over all the variables then we need to know the conditional probability for each variable conditional on its parents. Note that when a node has no parents, the conditional probability just becomes a prior probability on the node (such as in the case of variables A and B above). These conditional probabilities are called the 'parameters of the graph'.

For example, the probability distribution associated with Graph 1 is

$$p(A, B, C) = p(C|A)p(A)p(B)$$

and the probability distribution associated with Graph 2 is

$$p(A, B, C) = p(C|A, B)p(A)p(B)$$

Figure 2 shows Graphs 1 and 2, together with the parameters which need to be defined in each case.

It is possible to restrict the parametrisation to a particular type of relationship. For example, in Graph 2, if there is a conjunctive relationship between the two causes A and B, in the sense that both contribute causally to C, then the table of conditional probabilities is shown in Table 1.

Table 1: Parameters specifying a conjunctive relationship between the two causes in Graph 2.

A	B	$p(C\|A,B)$
0	0	0
0	1	0
1	0	0
1	1	1

Table 2: Parameters specifying a disjunctive relationship between the two causes in Graph 2.

A	B	$p(C\|A,B)$
0	0	0
0	1	1
1	0	1
1	1	1

On the other hand, a disjunctive relationship (where either cause A or cause B produces the effect C) is shown in Table 2. Another simple functional form is a Noisy-OR, shown in Table 3. This applies in the situation where both A and B increase the probability that C occurs, but each cause acting alone does not give probability one that C occurs.

We can deploy this framework to formally represent the different hypotheses which are competing in examples such as the explanation of the Cretaceous-Paleogene mass extinction. The hypothesis that the extinction was caused simply by a bolide impact can be represented by a causal structure such as in Figure 1(a). Here A would represent bolide impact and C would represent the extinction. The hypothesis that multiple causes were involved can be represented instead by a causal structure such as in Figure 1(b). In this case, the variable B represents an additional cause such as volcanic activity. Of course, for both options, there are many details to be filled in to give a plausible specific hypothesis. But when we ask whether the mass extinction is best explained by the impact hypothesis as opposed to the multiple-cause hypothesis, this can be seen as the question of which theory schema out of Figure 1(a) and Figure 1(b) best accounts for the evidence. The important point is that the competing hypotheses are distinct causal structures, and it is legitimate to treat these as mutually exclusive alternatives. Formally, the alternatives are different directed acyclic graphs or (DAG) representing different hypotheses about the causal structure. These DAGs are not identical to any particular set of probabilitiy densities over the variables in the graph. As we have seen, the DAG can indeed generate the joint probability density over all the variables, but only on the basis of certain assumptions such as the Causal Markov Condition.

Table 3: Parameters specifying a noisy-OR relationship between the two causes in Graph 2.

A	B	$p(C\|A,B)$
0	0	0
0	1	w_B
1	0	w_A
1	1	$w_A + (1 - w_B)w_A$

It is of course possible to adjust the causal parameters of the multi-cause schema to accommodate data which fit well to a single-cause schema – for example, by setting the strength of the causal arrow between B and C to zero. This means that the set of probability densities generated from the single-cause schema is indeed a subset of the set that can be generated from the multiple-cause schema. But this does not mean that the single-cause schema itself need be regarded as a special case of the multiple-cause schema. If we do not identify the schemas with the set of probability densities they generate, then the schemas themselves can still be regarded as mutually exclusive alternatives.

Other cases of IBE applied to apparently non-mutually exclusive hypotheses can be treated similarly. For example, the disjunctive schema representing the Aristotelian point of view allows for the possibility of two different kinds of generating process for organisms: spontaneous generation and generation from parent organisms. The Aristotelian theory can be represented by a schema of the form in Figure 1(b), whereas the modern theory would be represented by a schema of the form in Figure 1(a). In this example, A would represent parent organisms which have a causal effect on B, the production of offspring. The variable C would represent an alternative cause for B consisting of a certain configuration of non-organic conditions. It is reasonable to treat these DAGs as mutually exclusive alternatives, because they represent different causal structures and thus distinct possible ways that the world might be. Graph 2 represents a world where spontaneous generation is an actual possibility, whereas graph A represents the world we think we live in, where living organisms can only be produced by reproduction from other living creatures.

Overall, then, using causal graph theory to formalise instances of IBE which appear to involve competition with conjunctive or disjunctive explanations shows that the higher-level schemas which are competing to provide the best explanation can be seen as causal structures which are represented by DAGs and which may plausibly be regarded as mutually exclusive alternatives.

6 Solution for Bayesianism

The problem of logical constraints for Bayesian model selection also arises only if we identify the models to be compared with sets of specific probability densities generated by those models. This 'set-based' way of understanding what a statistical model is is fairly common in statistics. For example, in his textbook *All of Statistics,* Larry Wasserman defines statistical models as follows:

> A statistical model $\mathcal{F}$ is a set of distributions (or densities or regression functions). A parametric model $\mathcal{F}$ is a set that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is
>
> $$\mathcal{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \mu \in \mathbb{R}, \sigma > 0 \right\}$$
>
> This is a two-parameter model. We have written the density as $f(x;\mu,\sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters. In general, a parametric model takes the form
>
> $$\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$$
>
> where θ is an unknown parameter (or vector of parameters) that can take values in the parameter space Θ. (Wasserman (2013), pp. 87-88)

According to this view, in the curve-fitting case, the linear model is regarded as a set of all the possible probability densities of normal form parametrised by α_0, α_1 and σ_1

$$\mathcal{M}_{\mathrm{LIN}} :\{\mathcal{N}(\alpha_0 + \alpha_1 x, \sigma_1), \alpha_0 \in \mathbb{R},\ \alpha_1 \in \mathbb{R}, \sigma_1 > 0\} \tag{6}$$

and the quadratic model as a set of all the possible probability distributions allowed by the parameters:

$$\mathcal{M}_{\mathrm{PAR}} :\{\mathcal{N}(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma_2), \beta_0 \in \mathbb{R},\ \beta_1 \in \mathbb{R},\ \beta_2 \in \mathbb{R},\ \sigma_2 > 0\} \tag{7}$$

Here $\mathcal{N}(\mu, \sigma)$ denotes a normal distribution with mean μ and standard deviation σ.[3] In this case, $\mathcal{M}_{\mathrm{LIN}}$ is again a subset of $\mathcal{M}_{\mathrm{PAR}}$:, and thus $p(\mathcal{M}_{\mathrm{LIN}}) \leq p(\mathcal{M}_{\mathrm{PAR}})$. Thus, again it would not seem possible to assign a higher probability to the simpler, linear model.

An alternative to the set-based view of models was suggested in Henderson et al. (2010). According to this 'generative' view, the general theory schemas or models are mathematical objects which can be used together with further assumptions to generate the set of specific hypotheses, but which should not be identified with such sets (even if that set is supplemented with a different label). Thus, we called the general theory schemas or models (that enter into equation 4) 'generators'. The $\{\mathcal{M}_i\}$ which are fed into equation 4 are general hypotheses independent of particular assumptions about the adjustable parameters.

I have already suggested this way of looking at causal models. There are several levels at which we learn about the causal model, and each can be conducted according to Bayesian principles. At the higher level, we compare different causal structures by assigning prior probabilities to the different graphs $\{\mathcal{G}_i\}$ and then calculating their posterior probabilities according to

$$p(\mathcal{G}_{\mathrm{i}}|D) = \frac{p(D|\mathcal{G}_{\mathrm{i}})p(\mathcal{G}_{\mathrm{i}})}{p(D)}$$

The marginal likelihood here is obtained by integrating over all the values the parameters of the graph could assume, and we thus compare the different graphs without needing to assign any particular choice of parameter values. Learning the causal structure constitutes a form of Bayesian model selection (Heckerman, Geiger, and Chickering, 1995; Koller and N. Friedman, 2009), and this is commonly distinguished in practice from the task of parameter estimation given a particular causal graph structure.

Models representing single causes, as well as conjunctive and disjunctive causes, can thus be compared in a Bayesian fashion. There are often computational challenges in calculating the relevant marginal likelihoods, but there are algorithmic techniques which have been developed for this purpose (MacKay, 2003). When we are interested in for example comparing whether the single-cause Graph 1 is better supported by data than a two-cause model of noisy-OR form, we would use the parametrisation of Graph 2 given in Table 3, and perform the integration over that family of causal models (Steyvers et al., 2003; Griffiths and Tenenbaum, 2005). Thus, this methodology provides a systematic way to address examples such as the extinction case, or the case of spontaneous generation. Depending on the data, a conjunctive model can be favoured over a single-cause model, or vice versa, and similarly for comparisons of a disjunctive model to a single-cause model, or indeed to a conjunctive model. For example, a single-cause model has fewer adjustable parameters than

[3] This would be described by a density function
$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}, \mu \in \mathbb{R}, \sigma > 0.$

a multiple-cause model (see Figure 2). If it nonetheless can account for the data, then it will be preferred by virtue of its greater simplicity. This occurs naturally in the calculation of the Bayesian posterior, since the marginal likelihood of the multiple-cause model is lower than the single-cause model on account of the need to integrate over a larger area of its parameter space where the fit is not good. However, whether or not there is a preference for the simpler hypothesis depends on the data. There are also situations where the multiple-cause model gets better support from the data than the one-cause model.

The generative view can also be applied to cases like curve-fitting. A curve-fitting problem may involve comparing two theory schemas H_1 and H_2, each specifying different functional forms for the relationship between variables X and Y:

H_1: $y = \alpha_0 + \alpha_1 x$

H_2: $y = \beta_0 + \beta_1 x + \beta_2 x^2$

The schema H_1 specifies a linear relationship between X and Y, whereas the schema H_2 gives a quadratic relationship. Each of these schemas concerns the general form of the relationship, rather than any specific curve holding between X and Y. For each schema, a number of specific curves can be generated. For instance, the curve $y = 3 + 2x$ is one of the possible specific curves generated by H_1, obtained by setting $\alpha_0 = 3$ and $\alpha_1 = 2$. The curve $y = 1 + 4x + 0.3x^2$ is one of the specific curves generated by H_2, obtained by setting the adjustable parameters $\beta_0 = 1, \beta_1 = 4$ and $\beta_2 = 0.3$.

According to the 'set-based' view of models, the model consists of the set of all specific hypotheses which it describes. In this case, for example, the two relevant sets would be all the specific curves that take the form $y = \alpha_0 + \alpha_1 x$ (that we previously denoted as LIN), and all specific curves that take the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (denoted as PAR). As we saw earlier, since LIN entails PAR, the probability of LIN cannot be greater than the probability for PAR.

However, if we regard H_1 and H_2 as generative schemas, and do not identify them with the sets LIN and PAR, we do not have to see H_1 as entailing H_2. Rather, H_1 and H_2 represent different theories about what the basic generating mechanism is that produces the data. They might represent different physical processes. Suppose, for example, that the linear and quadratic models are used to describe a situation where X represents the concentration of a particular reactant and Y represents the rate at which a chemical reaction proceeds. Depending on how exactly the molecules combine with one another, the rate of a chemical reaction may be linearly dependent on the concentration of a particular reactant. However, if the rate is quadratically dependent on the concentration of a reactant, that may signal the presence of a different kind of reaction – namely a 'second-order' reaction (Atkins, De Paula, and Keeler, 2006). The two schemas thus correspond to quite different physical situations.

Bengt Autzen raises an objection to the generative view as follows (Autzen, 2019). If H_1 is taken to say that 'the curve specifying the relation between X and Y has a linear form', and H_2 is taken to say that 'the curve specifying the relation between X and Y has a quadratic form', then indeed H_1 would still entail H_2, and the problem of logical constraints would not be evaded. However, this is not how we should understand what these schemas amount to. Schemas like H_1 and H_2 provide a specification of the physical possibilities for a situation. H_2 represents a physical situation where a process described by a quadratic equation actually is possible, whereas H_1 rules that kind of process out. In the case of the chemical reactions, H_2 allows that there can be a second-order reaction going on in the system – even if it makes a negligible contribution to the rate, and even in the case where its presence might be hard to detect. H_1 on the other hand, claims that no such reaction is possible. So understood, H_1 and H_2 do describe mutually exclusive ways that the world might be.

The generative view preserves what is correct about the problem of logical constraints. For genuinely nested sets of functions or distributions such as LIN and PAR, probabilities are indeed constrained to obey the inequality $p(\text{LIN}){\leq}\, p(\text{PAR})$. But such an inequality at the level of the specific hypotheses is compatible with the generators being mutually exclusive. A common assumption in Bayesian model selection is to set the priors for the different models equal. So here for example, we might set the prior probability of the generator H_1 equal to that of the generator H_2 (supposing that these are the only alternatives, they both have prior probability 0.5). Now H_1 only assigns probability to specific hypotheses of the form $y = \alpha_0 + \alpha_1 x$, whereas H_2 assigns probability to specific hypotheses of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$, of which some are linear (if $\beta_2 = 0$) and some are non-trivially quadratic ($\beta_2 \neq 0$). Now consider the probability of the set of specific hypotheses LIN

$$p(\text{LIN}) = \sum_i p(\text{LIN}|H_i)p(H_i)$$

If H_1 is the generator, the specific curve produced will definitely be in LIN: $p(\text{LIN}|H_1) = 1$, whereas if H_2 is the generator there is some (probably small) probability p to produce a curve in LIN: $p(LIN|H_2) = p$. Thus $p(LIN) = 0.5.1 + p.0.5$. On the other hand

$$p(\text{PAR}) = \sum_i p(\text{PAR}|H_i)p(H_i)$$

and no matter which of H_1 and H_2 is the generator, the specific curve produced will definitely be in PAR. Thus p(PAR)=1. Thus, for any $p < 1$, the inequality $p(\text{LIN}) \leq p(\text{PAR})$ will be satisfied, even though the generator H_1 could in principle be assigned a higher prior probability than H_2, or as in this case, equal prior probability to H_2.

I now compare the generative view with an alternative suggested in Autzen (2019). Autzen agrees that the set-based view of models is responsible for the problem of logical constraints in Bayesian model selection. However, he proposes a different view of models than the one presented here. Autzen argues that besides the set-based view, there is another usage of the term 'model' to be found in Bayesian statistics. This is what he calls a 'Bayesian model'. A Bayesian model is not simply $\{p(y|\theta) : \theta \in \Theta\}$ (or $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ for probability densities) but also includes the prior over the adjustable parameters. Thus a Bayesian model is $(\{p(y|\theta) : \theta \in \Theta\}, p(\theta))$ with $p(\theta)$ denoting the prior probability density of θ. Autzen says

> By including the prior of the adjustable parameter into the model, it becomes clear how models that contain pairwise identical probabilistic hypotheses about the data-generating mechanism can have different empirical content. (Autzen (2019), p. 330)

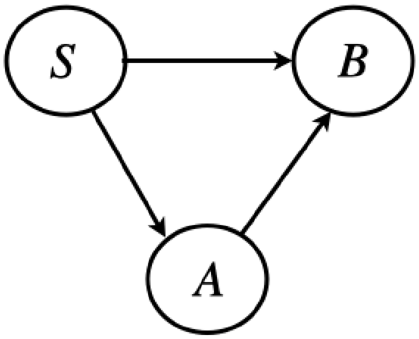
Applying Autzen's idea to the curve-fitting example, the models that are compared in equation 4 are not $\mathcal{M}_{\text{LIN}}$ and $\mathcal{M}_{\text{PAR}}$, but $\mathcal{M}^*_{\text{LIN}}$ and $\mathcal{M}^*_{\text{PAR}}$ defined as

$$\mathcal{M}^*_{\text{LIN}} : (\{N(\alpha_0 + \alpha_1 x, \sigma_1), \alpha_0 \in \mathbb{R},\, \alpha_1 \in \mathbb{R}, \sigma_1 > 0\}, \nu(\alpha_0, \alpha_1, \sigma_1))$$

$$\mathcal{M}^*_{\text{PAR}} : \left(\{N(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma_2), \beta_0 \in \mathbb{R},\, \beta_1 \in \mathbb{R},\, \beta_2 \in \mathbb{R},\, \sigma_2 > 0\}, \nu(\beta_0, \beta_1, \beta_2, \sigma_2)\right)$$

where $\nu(\alpha_0, \alpha_1, \sigma_1)$ and $\nu(\beta_0, \beta_1, \beta_2, \sigma_2)$ are priors over the adjustable parameters of $\mathcal{M}_{\text{LIN}}$ and $\mathcal{M}_{\text{PAR}}$ respectively. The problem of logical constraints is avoided, because $\mathcal{M}^*_{\text{LIN}}$ is no longer a subset of $\mathcal{M}^*_{\text{PAR}}$, thanks to the inclusion of these priors in the definition of the model.

Figure 3: Smoking (S) has a negative causal effect on bellysize (B) and a positive causal effect on activity (A),which in turn has a positive effect on B. In general for this causal structure, B is probabilistically dependent on S, but with a violation of Faithfulness, the causal parameters can be chosen such that B is independent of S. Then the path $S \to B$ exactly cancels the path $S \to A \to B$.



Autzen's Bayesian model approach also solves the problem of logical constraints in a technical sense. However I think the generative view is preferable because it corresponds better to the way the models $\{\mathcal{M}_i\}$ are actually regarded in practice. Models, I maintain, are treated as separate mathematical entities – such as causal DAGs, which provide schemas for the construction of theories. Furthermore the generative approach does not require that the prior over parameters be incorporated into the definition of the models themselves. Doing so brings problems of its own, as Autzen acknowledges, since there may well be cases where it is unclear how exactly to specify the prior. Thus, on the Bayesian model approach, all the problems associated with the assignment of Bayesian priors enter into the definition of the competing higher-level models.

Finally, we are now in a position to see why abandoning Bayesianism cannot be the right solution to the problem of logical constraints. There are of course a number of non-Bayesian approaches to model selection which do not assign probabilities to the competing models. Thus this might appear to be a reason to opt for non-Bayesian approaches, rather than Bayesian ones. However, the problem of logical constraints does not disappear in non-Bayesian methodology. Rather it appears in a different guise. The non-Bayesian must also address the problem of why you would ever prefer the simpler hypothesis, given that if the models are nested, you can always adjust the parameters of the more complex model so that it coincides with the simpler model as a special case. You can, for example, always adjust the quadratic model to put $\beta_2 = 0$, and then the question is why would you prefer the linear model to the adjusted quadratic model? In non-Bayesian approaches to causal structure learning, this preference has been enforced by adopting a special principle known as Faithfulness. Let $\mathcal{G}$ be a causal graph and P a probability distribution generated by $\mathcal{G}$. In general, $\mathcal{G}$ may contain other probabilistic independences than those that the Causal Markov Condition implies. $\mathcal{G}$ and P satisfy the Faithfulness Condition if and only if every conditional independence relation true in P is entailed by the Causal Markov Condition. For example, we could have a case where smoking S has a negative effect on bellysize B, but it also happens that smoking makes a person more active A, and this has a positive effect on bellysize (see Figure 3). According to the

CMC, there are no conditional independencies between any of the variables in this graph. Thus in general we expect to see dependence between smoking and bellysize. However, it is possible for S and B, for example, to be independent of one another for particular choices of the causal parameters. This could occur, for instance, if the parameters are such that the correlation induced by the common cause S exactly cancels the direct causal path from A to B. In this case the causal parameters would be 'fine-tuned' to produce the independence, rather than the structure of the causal graph itself being responsible. The Faithfulness condition essentially rules out such fine-tuning.

To justify invoking Faithfulness, Spirtes *et al.* argue that it is an instance of a more general principle of scientific inference which they call 'Spearman's principle' (Spirtes et al., 2000). If we are comparing two models which both account for the data, on the basis of which we judge certain 'constraints' to hold in the system in question (such as probabilistic independencies in the population of interest), then Spearman's principle says that we should prefer (other things being equal) the model which generates these constraints no matter what values are assigned to that model's 'free parameters' over the model which yields the constraints only for particular values of its free parameters. There has been discussion of the justification for special principles such as this (Woodward, 1998; Weinberger, 2018). Marc Lange, for example, has argued that there appear to be cases where Spearman's Principle should not hold (Lange, 1995). I will not pursue this issue further here. My main observation is that giving up on Bayesian methodology (as in the solution suggested in section 3.2.2) does not entirely solve the problem generated by the comparison of nested theories, since non-Bayesian methodology also has to deal with the problem of justifying the special principles it invokes to explain and justify a preference for simpler theories.

7 Conclusion

In a number of scientific inferences the competing hypotheses appear to be logically consistent. For example, there are cases in which single-cause hypotheses compete with hypotheses involving either conjunctions or disjunctions of causes. Since theories of scientific inference such as IBE and Bayesianism usually assume that competing hypotheses are mutually exclusive, this presents a challenge which I have called the 'problem of logical constraints'. For Bayesianism, the problem manifests itself in the apparent need to constrain probability assignments by the logical relations between the competing hypotheses. My solution to the problem of logical constraints is to see that, despite appearances, the competing hypotheses actually are mutually exclusive alternatives. This is motivated by a hierarchical view of theory comparison, which is also key to my 'emergent compatibilist' view that IBE can be explicated in Bayesian terms (Henderson, 2014). From this point of view, it is to be expected that there should be a common solution to the problem of logical constraints for both IBE and Bayesianism.

For causal examples, I have argued that the causal models that are competing are actually different causal structures, and should not be identified with sets of specific hypotheses generated by those structures. Even though the causal structures may generate nested sets of specific hypotheses, this does not mean that there are entailment relations between them. Rather, it is legitimate to regard the competing causal structures as mutually exclusive alternatives. I suggest that this solution can be generalised beyond causal examples, if we recognise the hierarchical way in which scientific theory comparison generally takes place. Scientific theory comparison involves comparison between models or theories at higher levels, and more fully specified hypotheses at lower levels. Well-recognised statistical techniques like model selection also proceed in a similar way. In standard examples like curve-fitting, the models compared at the higher level can also be regarded as schemas which represent distinct physical situations and which may thus be regarded as mutually exclusive alternatives. This account makes sense of usual Bayesian model selection practices, in which priors are assigned to the competing models without any concern for logical constraints. Nonetheless, we have also shown that logical constraints at the level of the specific hypotheses are still respected by the probabilities.

References

Alvarez, Luis W (1983). “Experimental evidence that an asteroid impact led to the extinction of many species 65 million years ago”. In: *Proceedings of the National Academy of Sciences of the United States of America* 80.2, pp. 627–642.

Atkins, Peter, Julio De Paula, and James Keeler (2006). *Atkins' Physical chemistry*. Oxford University Press.

Autzen, Bengt (2019). “Bayesian Ockham's razor and nested models”. In: *Economics and Philosophy* 35.2, pp. 321–338.

Claeskens, Gerda and Nils Lid Hjort (2008). *Model selection and model averaging*. Cambridge University Press.

Forster, M and E Sober (1994). “How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions”. In: *The British Journal for the Philosophy of Science* 45, pp. 1–35.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, et al. (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Gelman, Andrew et al. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

Gigerenzer, Gerd (1991). “How to make cognitive illusions disappear: Beyond “heuristics and biases””. In: *European review of social psychology* 2.1, pp. 83–115.

Griffiths, Thomas L., Charles Kemp, and Joshua B. Tenenbaum (2008). “Bayesian models of cognition”. In: *The Cambridge Handbook of Cognitive Psychology*. Ed. by R. Sun. Cambridge University Press.

Griffiths, Thomas L. and Joshua B. Tenenbaum (2005). “Structure and strength in causal induction”. In: *Cognitive psychology* 51.4, pp. 334–384.

Grünwald, Peter (2007). *The minimum description length principle*. MIT Press.

Grünwald, Peter and Teemu Roos (2019). “Minimum description length revisited”. In: *International journal of mathematics for industry* 11.01, p. 1930001.

Heckerman, David, Dan Geiger, and David M Chickering (1995). “Learning Bayesian networks: The combination of knowledge and statistical data”. In: *Machine learning* 20.3, pp. 197–243.

Henderson, Leah (2014). “Bayesianism and inference to the best explanation”. In: *The British Journal for the Philosophy of Science* 65.4, pp. 687–715.

— (2017). “Bayesianism and Inference to the Best Explanation: the case of individual vs. group selection in biology”. In: *Best explanations: new essays on Inference to the Best Explanation*. Ed. by Ted Poston and Kevin McCain. Oxford University Press.

Henderson, Leah et al. (2010). “The structure and dynamics of scientific theories: A hierarchical Bayesian perspective”. In: *Philosophy of Science* 77.2, pp. 172–200.

Howson, Colin (1988). “On the Consistency of Jeffreys's Simplicity Postulate, and its Role in Bayesian Inference”. In: *The Philosophical Quarterly* 38.150, pp. 68–83.

Jefferys, William H and James O Berger (1992). “Ockham’s razor and Bayesian analysis”. In: *American Scientist* 80.1, pp. 64–72.

Kahneman, Daniel et al. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge university press.

Keller, Gerta (2014). “Deccan volcanism, the Chicxulub impact, and the end-Cretaceous mass extinction: Coincidence? Cause and effect”. In: *Geological Society of America Special Papers* 505, pp. 57–89.

Koller, Daphne and Nir Friedman (2009). *Probabilistic graphical models: principles and techniques.* MIT Press.

Kuhn, Thomas S. (1962). *The structure of scientific revolutions.* University of Chicago Press.

Lakatos, Imre (1978). “Falsification and the methodology of scientific research programmes”. In: *The Methodology of Scientific research programmes.* Ed. by John Worrall and G. Currie. Cambridge University Press.

Lange, Marc (1995). “Spearman’s Principle”. In: *The British Journal for the Philosophy of Science* 46.4, pp. 503–521.

Laudan, Larry (1977). *Progress and its problems.* Great Britain: Routledge and Kegan Paul.

Lehoux, Daryn (2017). *Creatures born of mud and slime: the wonder and complexity of spontaneous generation.* JHU Press.

Lipton, Peter (2001). “Is explanation a guide to inference? A reply to Wesley C. Salmon”. In: *Explanation: theoretical approaches and applications.* Ed. by Giora Hon and Sam S. Rakover. Dordrecht: Kluwer Academic, pp. 93–120.

MacKay, David JC (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

Pearl, Judea (2009). *Causality.* Cambridge university press.

Popper, Karl (1959). *The Logic of Scientific Discovery.* London: Hutchinson.

Romeijn, Jan-Willem and Rens van de Schoot (2008). “A philosopher’s view on Bayesian evaluation of informative hypotheses”. In: *Bayesian evaluation of informative hypotheses.* Springer, pp. 329–357.

Schupbach, Jonah N. (2019). “Conjunctive explanations and Inference to the best explanation”. In: *Teorema: International Journal of Philosophy* 38.3, pp. 143–162.

Schupbach, Jonah N. and David Glass (2017). “Hypothesis competition beyond mutual exclusivity”. In: *Philosophy of Science* 84, pp. 810–824.

Sober, Elliott (2015). *Ockham’s razors: a user’s manual.* Cambridge University Press.

Spirtes, Peter et al. (2000). *Causation, prediction, and search.* MIT press.

Steyvers, Mark et al. (2003). “Inferring causal networks from observations and interventions”. In: *Cognitive science* 27.3, pp. 453–489.

Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference.* Springer Science & Business Media.

Weinberger, Naftali (2018). “Faithfulness, coordination and causal coincidences”. In: *Erkenntnis* 83.2, pp. 113–133.

Woodward, James (1998). “Causal independence and faithfulness”. In: *Multivariate Behavioral Research* 33.1, pp. 129–148.

Zwier, Karen R (2018). “Methodology in Aristotle’s theory of spontaneous generation”. In: *Journal of the History of Biology* 51.2, pp. 355–386.