Jan Sprenger Stephan Hartmann

Bayesian Philosophy of Science

Variations on a Theme by the Reverend Thomas Bayes

$$p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H})\,\frac{p(\mathrm{E}|\mathrm{H})}{p(\mathrm{E})}$$

Ich bin ein Esel, und will getreu,
wie meine Väter, die Alten,
an der alten, lieben Eselei,
am Eseltume halten.

Und weil ich ein Esel, so rat ich euch,
den Esel zum König zu wählen;
wir stiften das große Eselreich,
wo nur die Esel befehlen.

Wir sind alle Esel! I-A! I-A!
Wir sind keine Pferdeknechte.
Fort mit den Rossen! Es lebe—Hurra!—
Der König vom Eselsgeschlechte!

Heinrich Heine

Contents

Theme:
Bayesian Philosophy of Science

Among the greatest achievements of science are the laws, models and theories it has came up with. Newton's laws of mechanics, Bohr's model of the atom and Einstein's Theory of Relativity gave us unprecedented insights into the nature of physical reality. Similarly, Mendel's laws of inheritance, Darwin's theory of natural selection, and Crick and Watson's innovations in molecular biology elucidated how species develop and how traits and properties are passed on from one generation to the next. Pioneers of rational choice theory like Von Neumann, Savage and Arrow explained behavior in terms of beliefs and preferences and came up with powerful axioms for rational decision-making under uncertainty.

This enumeration does not intend to depreciate the value of experimental work in science. Rather, we would like to motivate why theories and models are central for understanding phenomena, predicting future events and transferring scientific knowledge to other domains. Therefore the **assessment of scientific theories and models** is a central part of scientific reasoning.

This book investigates how Bayesian inference can contribute to this goal. While we do not want to claim that scientific reasoning is essentially Bayesian, we claim that Bayesian models can elucidate diverse aspects of scientific reasoning, increasing our understanding of how science works and why it is so successful. The book is written a cycle of variations on this theme; it applies Bayesian inference to eleven different aspects of scientific reasoning.

In this introduction, we explain the constitutive principles and philosophical foundations of Bayesian inference, as well as some particular rea-

soning techniques (e.g., Bayesian networks) that we need in the remainder of the book. Then we describe our methodological approach—Bayesian philosophy of science—in slightly more detail. The level is introductory; no knowledge of calculus or higher mathematics is required.

The Static Dimension of Bayesian Inference: Probability and Degrees of Belief

In science as well as in ordinary life, we routinely make a distinction between more and less credible hypotheses. Consider, for example, the question which nation will win the 2016 European Football Cup. We may consider Albania a very unlikely candidate, England not completely implausible, and we may find it likely that the winner will be either France or Germany. This example illustrates that the epistemic standing of a hypothesis is no all-or-nothing affair, but a matter of gradation. Here the Bayesians step in: they use the concept of **degree of belief** to describe epistemic attitudes about uncertain propositions, and they represent these degrees of belief by a particular mathematical structure: probability functions. These two modeling assumptions are the central elements of Bayesian inference. In other words, Bayesians regard probabilities as expressions of subjective uncertainty—an interpretation that goes back to the English mathematician, reverend and philosopher Thomas Bayes (1701–1761) (Bayes, 1763).

The probability calculus has a long history as a tool for handling subjective uncertainty. In particular, it is one of the dominant paradigms in the psychology of human reasoning (e.g., Oaksford and Chater, 2000). Other scientific applications abound, such as Bayesian inference in phylogenetics, Bayesian interpretations of quantum mechanics, statistical inference and causal induction (e.g., Bernardo and Smith, 1994; Spirtes et al., 2000). Bayesian reasoning is also widely applied in philosophy: it is a standard tool in various branches of epistemology (e.g., Bovens and Hartmann, 2003; Pettigrew, 2015) and in the foundations of decision theory and rational choice (e.g., Jeffrey, 1971; Savage, 1972). The prominence of Bayesian inference in established scientific and philosophical theories recommends it as the default model for modeling uncertain reasoning. We do not claim that it is on foundational grounds superior to alternatives such as ranking functions (Spohn, 2012) or Dempster-Shafer theory (Shafer, 1976): we only claim that the past successes and the wide scope of Bayesian mod-

els recommend them as an excellent tool for studying scientific reasoning. Practical considerations, such as the simplicity of the probability axioms and the existence of a well-developed mathematical theory underneath them, support the case for Bayesian philosophy of science, too.

The distinctive feature of Bayesian inference is the central role of degree of belief. More traditional descriptions of epistemic attitudes, e.g., theories that just distinguish between belief, disbelief and suspension of judgment, struggle to adequately describe graded epistemic attitudes. This fails to account for many cases of scientific reasoning where we do hold graded beliefs. But psychological realism is not the only force that pulls into the direction of a graded theory of epistemic attitudes. An all-or-nothing account of epistemic attitudes also leads into *philosophical* trouble, such as in the famous lottery paradox (Kyburg, 1961). For instance, if just one out of 1.000.000 tickets in a lottery is winning, then for each single ticket #i ($i \in \{1, 2, \ldots, 1.000.000\}$), we seem to be entitled to believe that it is not the winning one. However, we also believe that there is a winning ticket in the lottery. Hence, the set of the propositions that we believe ("There is a winning ticket", "ticket #1 does not win", "ticket #2 does not win", etc.) is inconsistent, which is at least prima facie an undesirable consequence. For theories of graded belief, no such inconsistency arises (for discussion of the full vs. graded belief relationship, see Leitgeb, 2014; Fitelson et al., 2016).

If we want to use the concept of degree of belief to describe which scientific theories are more credible than others, we have to say something about the rules that rational degrees of belief have to satisfy. After all, the objects of degrees of belief—propositions—are related to each other in manifold ways, and these relations constrain the set of degrees of belief that we can rationally entertain. For example, it seems that we cannot believe proposition A to a higher degree than the proposition A $\wedge$ B since A $\wedge$ B cannot be true without A being true. For Bayesians, the probability calculus captures all relevant constraints that rational degrees of belief over a set of logically interconnected propositions have to satisfy.

Let L be a propositional language and let $\mathcal{L}$ be a σ-algebra over well-formed formulae of L: that is, a set of propositions that contains the tautology and contradiction, is closed under logical negation and under infinite disjunction of propositions. This method of construction ensures that the algebra contains all truth-functional compounds of propositions which are

already in the algebra.

A probability function $p : \mathcal{L} \to [0,1]$ operates on such an algebra of propositions and takes values in the unit interval $[0,1]$. That is, if we assign degrees of belief to two propositions, we also have degrees of belief in their truth-functional compounds. More precisely, truth-functional operators affect the assignment of degrees of belief according to the following axioms of probability (Kolmogorov, 1933):

Probability Function *For a propositional language $\mathcal{L}$ with a σ-algebra $\mathcal{A}$, $p : \mathcal{A} \to [0,1]$ is called a* probability function *if and only if it satisfies the following three properties:*

1. $p(\top) = 1$.
2. $p(\neg \mathrm{A}) = 1 - p(\mathrm{A})$.
3. *For mutually exclusive propositions* $\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3, \ldots$:

$$p\left(\bigvee_{n \in \mathbb{N}} \mathrm{A}_n\right) = \sum_{n=1}^{\infty} p(\mathrm{A}_n) \tag{1}$$

In this model, degrees of belief correspond to numbers in the interval $[0,1]$, where zero denotes minimal and one denotes maximal degree of belief. It is not hard to motivate the three above constraints for rational agents: Each tautology is assigned maximal degree of belief. If A is strongly believed, its negation $\neg$A is weakly believed, and vice versa. In particular, the degree of belief in A and $\neg$A add up to unity. Finally, the degree of belief in the disjunction of mutually exclusive propositions corresponds to the sum of the individual propositions. This can be understood as summing up the weight of the possible worlds where $\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3$ etc. obtain. Prima facie, these constraints capture our everyday use of the word "probable".

It is notable that the third condition uses an infinite instead of a finite sum. Indeed, this choice is controversial in the literature. There is a substantial debate about whether probabilistic degrees of belief should satisfy this condition of **countable additivity** instead of the weaker requirement of **finite additivity**: $p(\mathrm{A}) + p(\mathrm{B}) = p(\mathrm{A} \vee \mathrm{B})$ for two mutually exclusive propositions A and B. Several authors argue that accepting countable additivity amounts to making substantial and unwarranted epistemological assumptions (de Finetti, 1972, 1974; Kelly, 1996; Howson, 2008). Jaynes

(2003) responds that countable additivity naturally flows from a proper mathematical modeling of uncertainty. Kadane et al. (1999) discuss further consequences of choosing finitely instead of countably additive probability functions. Fortunately, this choice does not make a difference for most applications in this book. Since countable additivity is standardly assumed in statistical inference, which is one of the focal points of this book (Variations 9–11), we take all probability functions to be countably additive.

Following Hailperin (1984, 1996) and Popper (2002), we conceptualize probability functions as operating on a σ-algebra of propositions. The alternative to this sentential approach consists in Kolmogorov's measure-theoretic approach: probabilities operate on σ-algebras of sets, and the objects of degrees of belief correspond to the epistemic possibilities that an agent considers (e.g., Easwaran, 2011a). In this context, one usually speaks of probability measures—our use of the term "probability function" indicates that we are not considering probability in the general set-theroetic sense. The sentential interpretation strikes us as more natural, simpler and closer to the purpose of this book, namely to model the assessment of scientific theories. As we will see, it is also well suited for justifying why degrees of belief should be probabilistic.

Let us get back to the three probability axioms. While their qualitative motivation is highly plausible, their quantitative form is harder to justify. Why should rational degrees of belief satisfy these peculiar axioms rather than another set of axioms with identical qualitative properties? Three types of arguments have been proposed as an answer:

1. Dutch Book arguments, associated with the names of Ramsey, De Finetti, and Jeffrey;

2. Decision-theoretic arguments à la Savage and von Neumann/Morgenstern;

3. Epistemic arguments due to Cox, Joyce, Pettigrew, and others.

It is important to mention upfront that all these arguments contain some form of idealization: the rational agents who are supposed to conform to the axioms of probability are ideally rational agents, that is, agents who are immune to trivial reasoning fallacies that real agents commit from time to time. Instead of claiming that the degrees of belief of *real agents*

conform to the probability axioms, the arguments below aim to show that *ideally rational agents* should have probabilistically coherent degrees of belief. After all, philosophical theories of degrees of belief are normative in the first place. They provide a logic of uncertain reasoning in the same sense that propositional logic does for classical deductive reasoning.

We begin with the famous **Dutch Book Arguments**. Frank Ramsey (1926) observed that many of our actions are based on our degrees of belief. He regarded human action as a kind of betting, similar to accepting a bet on the next European football champion:

> [...] all our lives we are in sense betting. Whenever we go to the station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home. (Ramsey, 1926, 85)

The most pervasive example of the link between belief and betting are perhaps transactions on financial markets—traders buy and sell stocks, certificates and options, according to their degrees of belief that these will rise and fall. Someone with a high degree of belief that an option will become worthless will sell it more eagerly, and for a lower price, than someone who is convinced that it will increase in value.

To distinguish rational from irrational degrees of belief, Ramsey uses the instrumental, economic conception of rationality and focuses on a particular type of belief-guided action: bets. If we accept and decline bets according to our degrees of belief, only probabilistic degrees of belief will avoid a sure loss of money, or so Ramsey argues. To this end, Ramsey defines degree of belief in the following way: having a degree of belief $p(\mathrm{A}) = x$ means that we consider a bet with stake $€x$ fair if it pays $€1$ if A is true, and nothing if A is false. This definition can be operationalized further: having a degree of belief $p(\mathrm{A}) = x$ implies that the agent is indifferent between taking the role of the bettor and the bookie in a bet on proposition A with betting odds $1/p(\mathrm{A})$. This technique resembles the famous veil of ignorance for disclosing judgments about the fair distribution of goods in a society (Rawls, 1971): the agent does not know whether she will end up as the bettor or as the bookie. As an alternative proposal, consider a fully dispositional, behaviorist definition of degrees of belief (e.g., de Finetti, 1937): agent S's degree of belief in proposition A is equal to p if and only if p is the price at which S would sell or buy a bet on A that pays $€1$ if A occurred.

Ramsey then continues that no system of degrees of belief can be fair if it gives rise to a system of bets (combining the roles of bettor and bookie) that implies a sure loss for the agent. Such a system of bets is called a Dutch book. By establishing an isomorphism between bets and degrees of belief, Ramsey grounds the famous **Dutch Book Argument**: he demonstrates that degrees of belief that violate the axioms of probability will give rise to Dutch Books. Conversely, all probabilistic systems of degrees of belief are immune to Dutch books. Hence, Ramsey infers that only probabilistic degrees of belief are rational (see also Kemeny, 1955).

The cogency of such arguments has been debated in various places (e.g., de Finetti, 1972; Howson, 2008; Hájek and Hartmann, 2010; Hartmann and Sprenger, 2010; Easwaran, 2011a,b). For the sake of simplicity, suppose a fully behaviorist interpretation of degrees of belief. The Dutch Book Theorem assumes that the agent accepts all bets where the proposed odds are higher than her personal odds (viz., degrees of belief), and is ready to act as bookie on all bets where the proposed odds are lower than her personal odds. Real agents, however, are often risk-averse and the stake may influence their willingness to take a side in the bet. In other words, an agent's degree of belief in a proposition may depend on the amount of money that she has to bet. Moreover, the agent may be unwilling to engage in any bet if the stakes are high enough, and be willing to suffer a Dutch book if the stakes are low enough. None of these behaviors strikes us as blatantly irrational unless we presuppose what is to be shown: that rationality equals immunity to Dutch Books.

These objections suggest that a straightforward operationalization of degrees of belief in terms of betting behavior or fairness judgments, about bets, is problematic. With a crumbling link between degrees of belief and fair betting odds, the Dutch Book Argument loses its a part of normative force. Sure, in many situations we can still argue for a strong dependency between degrees of belief and fair betting odds, but it might go too far to identify both concepts with each other. For this reason, we now move to the second, **decision-theoretic argument** in favor of the thesis that degrees of belief should be probabilistic.

The idea of this argument is that probabilistic degrees of belief can represent the epistemic state of an agent who bases her choices on rational preferences. First, a number of axioms are imposed on rational preferences, represented by the binary relation $\preceq$. For example, it is usually

assumed that such preferences are *transitive*: if an agent prefers apples to bananas and bananas to cherries, then she will also prefer apples to cherries. Similarly, it is often assumed that such preferences are *complete*; that the agent either strictly prefers one of two options or she is indifferent between them.

In his 1954 hallmark book "The Foundations of Statistics", Leonard J. Savage sets up an entire system of such axioms, called P1-P7 (for an accessible introduction, see Karni, 2005). They contain transitivity and completeness as well as more demanding axioms, such as the *Sure Thing Principle*: preference between two acts merely depends on their consequences in those states of the world where they have different payoffs.

Savage then proceeds to proving his famous representation theorem. If the preferences of an agent X satisfy the axioms P1-P7, there is a probability function p and a real-valued utility function u (unique up to affine transformation) such that for any two acts f and g, with respect to a state space S:

$$f \preceq g \;\Leftrightarrow\; \int_S u(f(S))dp(s) \;\leq\; \int_S u(g(S))\,dp(s)$$

In other words, act g is preferred to act f if and only if the expected utility of g, relative to the agent's subjective degrees of belief, exceeds the expected utility of f. In other words, we can represent an rational agent as maximizing the subjective expected utility of her actions.

Savage's approach has been very influential in economics, and it bridges the gap between epistemology and decision theory in an attractive and elegant way. However, it is not without drawbacks. First, the probability function $p(\cdot)$ describing the agent's degrees of belief is not unique by itself: it is only jointly unique together with the utility function. This weakens the appeal of Savage's results for models of scientific reasoning, where pragmatic utility considerations are often thought to be secondary to pursuit of truth. Second, Savage's axioms on rational preferences are not all equally compelling. For instance, Maurice Allais (1953) and Daniel Ellsberg (1961, 2001) conducted influential experiments that challenged one of Savage's axioms, the Sure-Thing Principle (see also Allais and Hagen, 1979). Both results have been replicated over and over again. Hence, the decision-theoretic justification of probabilistic degrees of belief also fails to be conclusive.

All this prompts the question of whether there can be a purely **epistemic argument** for the probabilistic nature of degrees of belief, free of op-

erationalist and decision-theoretic considerations. The first attempt along these lines was made by the physicist Richard Cox in 1946, using the word plausibility instead of degree of belief. He demonstrated that any real-valued function $p(\cdot)$ representing the plausibility of a proposition is isomorphic to a probability function if the following two assumptions (plus minor technical requirements) are satisfied:

Complementarity There is a decreasing function $f : \mathbb{R} \to \mathbb{R}$ such that $p(\neg \mathrm{A}) = f(p(\mathrm{A}))$.

Compositionality There is a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $p(\mathrm{A} \wedge \mathrm{B}) = g(p(\mathrm{A}), p(\mathrm{B}|\mathrm{A}))$ where $p(\mathrm{B}|\mathrm{A})$ denotes the plausibility of B if we already know A.

In other words, if (i) the plausibility of the negation of a proposition is a decreasing function of the plausibility of the proposition, and (ii) the plausibility of $\mathrm{A} \wedge \mathrm{B}$ is determined by the plausibility of A and B given A, then plausibility measures obey the mathematical probability structures—at least up to mathematical isomorphy (Cox, 1946). $p(\mathrm{B}|\mathrm{A})$ denotes the degree of belief in A if we suppose that A is true, if we take A as given. While Complementarity is uncontroversial, the real philosophical issue is whether the plausibility of $\mathrm{A} \wedge \mathrm{B}$ is indeed a mere function of the plausibility of A and the plausibility of B if we suppose A. We will say more on this in the next section.

In more recent times, the epistemic approach to justifying probabilistic degrees of belief has been resuscitated from a different perspective. James Joyce (1998, 2009) has made major contributions to this research project, based on earlier work by Brier (1950) and Rosenkrantz (1981). As a criterion for the rationality of a system of degrees of belief, Joyce evaluates their *inaccuracy*: that is, he compares our degrees of belief $p(\mathrm{A})$ to the actual truth values of the believed propositions $r(\mathrm{A}) \in \{0, 1\}$. The conventional measure of inaccuracy is the *Brier score* taken over all propositions in a σ-algebra $\mathcal{A}$:

$$B = \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} (p(\mathrm{A}) - r(\mathrm{A}))^2 \tag{2}$$

Generally, degrees of belief match the truth or falsity of the believed propositions quite well in some possible worlds (namely, in those where probable propositions are true) and less well in others (namely, in those where probable propositions are false). Joyce shows that if a system of degrees

of belief does not satisfy the probability axioms, then there is a probability function that is no less accurate in all possible worlds, and more accurate in others, as measured by the Brier score. That is, belief functions that violate the axioms of probability are *dominated* by probabilistic belief functions (Joyce, 1998).

Like the other two types of arguments and Cox's variant of the epistemic argument, Joyce's epistemic inaccuracy argument is not without controversial assumptions: Hájek (2008) points out that one of Joyce's assumption is to assume the converse side of the Dutch Book Theorem, namely that probabilistic belief functions cannot be dominated accuracy-wise (see Maher, 2002, for further criticism). Moreover, the results are sensitive to the choice of the scoring rule (here: the Brier score) for calculating the inaccuracy of a belief function. Leitgeb and Pettigrew (2010a,b) and Pettigrew (2016) develop this research program further.

We have seen that none of the above three arguments succeeds at providing a watertight justification for why degrees of belief should obey the axioms of probability. However, they go a long way along this road. It is also notable that the same result is reached from completely different perspectives and methodological approaches. This suggests that the different attempts provide substantial cumulative justification for modeling degrees of belief as probabilities. Moreover, given the fact that probability is a very specific mathematical structure, it is perhaps not surprising that we have to make substantial and sometimes contentious assumptions in order to obtain a unique representation of degree of belief. We will now proceed to the next concept which plays a central role in Bayesian inference: conditional probability.

Conditional Degree of Belief and Bayes' Theorem

The previous section has motivated why the degrees of belief of a rational agent at a particular time should satisfy the axioms of probability. What about the dynamics of these degrees of belief? How should they change in the light of incoming evidence? To answer this question, Bayesians make use of the concept of **conditional degree of belief**, which we have also seen in Cox's representation theorem above: the rational degree of belief in a scientific hypothesis H after learning evidence E is the conditional degree of belief in H given E, mathematically represented by the **conditional**

probability $p(\mathrm{H}|\mathrm{E})$.

Before we can say more about the dynamics of Bayesian inference, we have to clarify what the concept of conditional degree of belief is about, and why conditional probability is the right explication of this concept. Conditional degree of belief captures the idea that we sometimes judge the plausibility of a proposition B in the light of another proposition A. For example, what is our degree of belief that Real Madrid will win the next Champions League if we suppose that Cristiano Ronaldo is injured for a period for six months? What is our degree of belief that crop yields will decrease by more than 50% if we suppose that there is a draught this summer? What is our degree of belief that there is an intelligent form of life outside the solar system if we make an assumption on the number of terrestrial planets in the galaxy? And so on.

In other words, we adopt the **counterfactual or suppositional interpretation of conditional degree of belief**. This interpretation actually goes back to Frank P. Ramsey (1926) and Bruno de Finetti (1937). Here is Ramsey's famous analysis of conditional degrees of belief:

> If two people are arguing 'if H will E?' and both are in doubt as to H, they are adding H hypothetically to their stock of knowledge and arguing on that basis about E. (Ramsey, 1926)

The above quote is ambiguous: is it about conditional (degree of) belief or about the truth or acceptability conditions for indicative conditionals? The next sentence clarifies Ramsey's project:

> We can say that they are *fixing their degrees of belief in* E *given* H. (ibid., our emphasis)

This makes clear that regardless of the possible link to the epistemology of conditionals, Ramsey intended that hypothetically assuming H would determine one's conditional degrees of belief in E, given H—see also de Finetti (1972, 2008).

Ramsey's analysis creates a direct link between conditional degree of belief and statistical reasoning. For instance, statisticians describe the probability of observing k heads in N independently and identically distributed (i.i.d.) tosses of a fair coin by the Binomial probability density function $B_{N,1/2}(k) = \binom{N}{k}(1/2)^N$. By Ramsey's definition, our conditional degrees of belief in observing a certain number of heads or tails follow

these probability densities—this is just what it *means* to suppose that the coin is fair and that the tosses have been i.i.d. By supposing the fairness of the coin and the i.i.d.-ness of the tosses, we fix our degrees of belief in observing two heads in three tosses at $\binom{3}{2}(1/2)^3 = 3/8$. That is, when interpreted counterfactually, conditional degrees of belief follow objective probabilities in statistical reasoning.

Due to the conditional, and sometimes outright counterfactual nature of these statements, the concept of conditional degree of belief is qualitatively different from the ordinary concept of belief. So why should conditional degrees of belief satisfy the axioms of probability, and how do they relate to unconditional degrees of belief?

To answer the first question, consider degrees of belief in the propositions B_1, B_2, etc., all of them conditional on another proposition A. Then the same arguments as in the previous section apply (or, from the point of view of a sceptic, don't apply). After all, the Dutch Book argument in favor of probabilistic representation of degree of belief does not make a difference between whether or not a proposition A is presupposed. Similarly for the decision-theoretic and the epistemic arguments: all of them retain their normative force if applied to degrees of belief conditional on proposition A and the probabilistic representation $p(\cdot|\mathrm{A})$. Supposing a proposition A creates a new family of degrees of belief, together with a probability function $p(\cdot|\mathrm{A})$ that describes their coherence.

But how do these probability functions square with the unconditional probabilities $p(\cdot)$? Standardly, the conditional probability of an event E given H is defined as the ratio of the probability of the conjunction of both events, divided by the probability of H (assuming $p(\mathrm{H}) > 0$):

$$p(\mathrm{E}|\mathrm{H}) = \frac{p(\mathrm{E} \wedge \mathrm{H})}{p(\mathrm{H})} \qquad \text{(Ratio Analysis)}$$

This move is primarily motivated from the mathematical development of the theory of probability by Kolmogorov (1933), but it can also be justified by a Dutch Book argument. Imagine that all the probabilities in Ratio Analysis represent the degrees of belief of a rational agent and correspond to a system of associated bets. The bet on the conditional event E given H, abbreviated $\mathrm{E}\,|\,\mathrm{H}$, is a so-called *conditional bet*: it is a regular bet on E if H turns out to be true, and the bet is called off—that is, the stake is returned to the bettor with no further consequences—if H is false (de Finetti, 1937). It can then be shown that failure to comply with Ratio Analysis leads to

a Dutch Book. Consider the system of bets where an agent wages $\in x$ on H and $\in y$ on E | H with betting odds $1/p(\mathrm{H})$ and $1/p(\mathrm{E}|\mathrm{H})$, while acting as a bookie for a bet on $\mathrm{E} \wedge \mathrm{H}$ with stake $\in z$ and odds $1/p(\mathrm{E} \wedge \mathrm{H})$. The odds correspond, of course, to the degrees of belief specified by $p(\cdot)$ and $p(\cdot|\mathrm{H})$. Assume further that the stakes satisfy the equations $x = z$ and $y = x/p(\mathrm{H})$. It can then be shown (proof omitted) that this system of bets is fair if, and only if, Ratio Analysis is satisfied. Otherwise, the agent will either be left with a sure loss or with a sure gain: the paradigm case of a Dutch book. For these reasons, Ratio Analysis is unanimously accepted as a constraint on conditional degrees of belief.

A consequence of Ratio Analysis is Bayes' famous theorem. If we combine

$$p(\mathrm{E}|\mathrm{H}) = \frac{p(\mathrm{E} \wedge \mathrm{H})}{p(\mathrm{H})}$$

with the cognate equation

$$p(\mathrm{H}|\mathrm{E}) = \frac{p(\mathrm{E} \wedge \mathrm{H})}{p(\mathrm{E})}$$

then we obtain, by a matter of simple substitution, **Bayes' Theorem:**

$$p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H})\frac{p(\mathrm{E}|\mathrm{H})}{p(\mathrm{E})} \tag{3}$$

This equation will accompany us throughout the book—it describes how the degree of belief in H given E relates to the unconditional degrees of belief in H and E, and to the conditional degree of belief in E given H. Note that it does *not* describe how agents should change their degrees of belief in H when learning E: it relates the probability function $p(\cdot)$, that represents the agents' unconditional degrees of belief, to the probability functions $p(\cdot|\mathrm{H})$ and $p(\cdot|\mathrm{E})$, which represent the agent's conditional degrees of belief. Acting as such an epistemic coordination principle is the philosophical significance of Bayes' theorem.

It is also possible to write the right hand side of Bayes' Theorem slightly differently:

$$p(\mathrm{H}|\mathrm{E}) = \left(1 + \frac{p(\neg\mathrm{H})}{p(\mathrm{H})} \cdot \frac{p(\mathrm{E}|\neg\mathrm{H})}{p(\mathrm{E}|\mathrm{H})}\right)^{-1} \tag{4}$$

In this formulation, the dependency of $p(\mathrm{H}|\mathrm{E})$ on $p(\mathrm{E})$ is replaced by a dependency on $p(\mathrm{E}|\neg\mathrm{H})$. In many cases of scientific and in particular statistical inference, the latter quantity can be accessed and calculated more

easily than $p(\mathrm{E})$. We will make use of both (3) and (4) frequently throughout the book.

We would also like to discuss a tempting proposal, namely to read Ratio Analysis as a *definition* of conditional probability, rather than a mathematical constraint. In fact, this is a road taken my many textbooks (e.g., Earman, 1992; Skyrms, 2000; Howson and Urbach, 2006). Transferred to conditional degree of belief, this would mean that the conditional degree of belief in E given H is just the ratio of the unconditional degrees of belief in $\mathrm{E} \wedge \mathrm{H}$ and H. In this case, Bayes' Theorem is indeed a theorem of mathematics without philosophical import.

In our view, a conditional probability that is reduced to unconditional probability would have trouble to describe conditional degrees of belief. First of all, the conditional probability of E given H cannot be calculated when the unconditional probability of H, $p(\mathrm{H})$, is zero. But intuitively, such conditional probabilities make sense when H is an idealized, but almost certainly false hypothesis (e.g., "this particular coin is fair", "these two random variables have the same variance", etc.). Similarly, intuitively meaningful questions such as "What is the probability that a point on Earth is in the Western hemisphere (H), given that it lies on the equator (E)?" cannot be answered if Ratio Analysis is an exhaustive analysis of conditional probability. At least, more technical detail has to be provided. Hájek (2003) gathers a lot of such criticisms in order to make a case against Ratio Analysis as a definition of conditional probability while Easwaran (2011c) and Myrvold (2015) explore avenues for parrying Hájek's criticism.

Second, Ratio Analysis fails to grasp the normative role of conditional degree of belief in Bayesian inference. Often, it is part of the *meaning* of H to constrain $p(\mathrm{E}|\mathrm{H})$ in a unique way. For determining our rational degree of belief that a fair coin yields a particular sequence of heads and tails, it does not matter whether the coin in question is actually fair. Regardless of our degree of belief in that proposition, we all agree that the probability of two heads in three tosses is $3/8$ *if we suppose that the coin is fair*. This sentence has a distinctly analytical flavor whereas the degrees of belief in $\mathrm{E} \wedge \mathrm{H}$ and H are prima facie unrestricted. If Ratio Analysis is all that we can say about conditional probability and conditional degree of belief, then this feature of conditional degree of belief drops out of the picture (see also Edgington, 1995).

Third, as a matter of psychological fact, we do not form conditional

degrees of belief via the conjunction of both propositions. It is cognitively very demanding to elicit our degrees of belief in $\mathrm{E} \wedge \mathrm{H}$ and H, and to calculate their ratio. Indeed, recent experimental evidence suggests that Ratio Analysis is a poor description of how people reason with conditional probabilities, pointing out the necessity of finding an alternative account (Zhao et al., 2009).

For all these reasons, Ratio Analysis is not suitable as a definition of conditional probability and conditional degree of belief. Arguably, the suppositional understanding of conditional probability is also better suited for scientific reasoning, e.g., degrees of belief in the outcomes of an experiment that is described by a statistical hypothesis. Indeed, several mathematicians, epistemologists and philosophers of science have proposed to understand conditional probability as a primitive concept (Renyi, 1970; Popper, 2002; Hájek, 2003; Maher, 2010). The unconditional probability of A can then be defined as the probability of A conditional on a tautological proposition. This move does justice to the intuition that if conditional degree of belief cannot be reduced to unconditional degree of belief, as we have argued above, then there is actually a *set* of probability functions which describe the agent's epistemic attitudes: $p(\cdot|\mathrm{A})$, $p(\cdot|\mathrm{B})$, and so on. Ratio Analysis and Bayes' Theorem are then more than a mathematical fact about a probability function: they can be interpreted as requirements on how conditional and actual degrees of belief, and the various probability functions that represent them, should cohere. By taking conditional probability as a primitive concept, the variety of probability functions is unified in a single two-ary function $p(\cdot|\cdot)$. Another option consists in unifying conditional and unconditional probabilities under the umbrella of the concept of a conditional expectation, and a σ-algebra that is conditional on a random variable. For this rather technical route, see Gyenis et al. (2016); Rédei and Gyenis (2016).

The suppositional interpretation of conditional probability has far-reaching philosophical implications which deserve a detailed treatment, but cannot be covered in this book (for discussion, see Sprenger, 2016a). As later chapters will show, the suppositional approach allows for fruitful applications of Bayesian reasoning to the topics of confirmation, old evidence, causality, and scientific objectivity. We will now return to the question of how the concept of conditional degree of belief provides Bayesian inference with a rule for updating degrees of belief.

The Dynamic Dimension of Bayesian Inference: Bayesian Conditionalization

The previous sections have explained the static dimensions of Bayesian inference: representing degrees of belief in terms of probabilities and coordinating unconditional with conditional degrees of belief. To obtain a full-fledged theory of reasoning with degrees of belief, we also need a principle that states how degrees of belief are changed in the light of incoming information. This is actually very simple: The rational degree of belief in hypothesis H after learning evidence E is expressed by the conditional probability of H given E.

Bayesian Conditionalization The rational degree of belief in a proposition H after learning evidence E, represented by the probability function $p'_E(\cdot)$, is the conditional probability of H given E according to the agent's original degrees of belief represented by the probability function $p(\cdot)$: $p'_E(\mathrm{H}) = p(\mathrm{H}|\mathrm{E})$.

The principle of Bayesian Conditionalization often figures as a cornerstone of Bayesian reasoning (e.g., Earman, 1992, 34). It is inspired by the same idea that motivated conditional degrees of belief: when we learn a piece of evidence E, we add it to our background knowledge and see which consequences this addition has for the rest of our epistemic attitudes. This is why the new degree of belief in H is set equal to the conditional probability of H given E.

By means of Bayes' Theorem, presented in the previous section, we can express Bayesian Conditionalization as follows (see Equations (3) and (4)):

$$\begin{aligned} p'_{\mathrm{E}}(\mathrm{H}) &:= p(\mathrm{H})\frac{p(\mathrm{E}|\mathrm{H})}{p(\mathrm{E})} \\ &= \left(1 + \frac{p(\neg\mathrm{H})}{p(\mathrm{H})} \cdot \frac{p(\mathrm{E}|\neg\mathrm{H})}{p(\mathrm{E}|\mathrm{H})}\right)^{-1} \end{aligned}$$

In this equation, $p(\mathrm{H})$ and $p(\mathrm{H}|\mathrm{E})$ are called the **prior probability** and **posterior probability** of H, while $p(\mathrm{E}|\mathrm{H})$ and $p(\mathrm{E}|\neg\mathrm{H})$ are called the **likelihoods** of H and $\neg$H on E, that is, the probability of the observed evidence E under a specific hypothesis, in this case H or $\neg$H. This version of Bayesian Conditionalization is especially useful for applications in statistical inference, where the statistical model often provides us with the

likelihood function $p(\mathrm{E}|\mathrm{H}_\theta)$, for hypotheses indexed by some parameter θ. Bayesian Conditionalization provides a way of learning novel evidence by means of trading off the prior probability of hypothesis H with the likelihoods of H and $\neg$H on the evidence.

Agents who initially have different degrees of belief, represented by probability functions $p_1(\cdot)$ and $p_2(\cdot)$, are brought closer to each other if they both follow Bayesian Conditionalization as an updating rule. As long as they agree on the propositions which obtain measure zero (in other words, $p_1(\mathrm{X}) = 0 \Rightarrow p_2(\mathrm{X}) = 0$, and vice versa), the distance between p_1 and p_2 will approach zero. In other words, individual differences will eventually cancel out if both agents are Bayesian conditionalizers. For more discussion of this convergence of priors literature, see Blackwell and Dubins (1962), Gaifman and Snir (1982) and Earman (1992).

Essentially, Bayesian Conditionalization forges together learning E, as described by $p'_{\mathrm{E}}(\cdot)$, and supposing E, as described by $p(\cdot|\mathrm{E})$. However, in spite of the intuitive similarity between learning E and supposing E, it is not easy to justify this equality. After all, there are nontrivial psychological differences: Zhao et al. (2012) found, in a recent experiment, that participants who *learned* evidence E (e.g., by observing relative frequencies) submitted different probability estimates of H than participants who had to *suppose* that E occurred. Given this discrepancy on the descriptive level, we have to come up offer a convincing normative argument in favor of Bayesian Conditionalization.

A standard proposal, much similar to what we have seen before, consists in **dynamic Dutch Book Arguments** (Teller, 1973). Consider an agent who knows in advance that an observable X will take one of the values $x_1, x_2, x_3, \ldots$ For instance, E could be the outcome of the toss of a die with possible results being in the set $\{1, \ldots, 6\}$. Assume that $p(X = x_1) > 0$ and that $p'(\mathrm{A}) < p(\mathrm{A}|X = x_1)$ for some proposition A, where $p'(\cdot)$ describes the agent's degrees of belief in case $X = x_1$ is learned. Assume further that the agent engages on the following system of bets: she buys a conditional bet on A given $X = x_1$, described by the odds $1/p(\mathrm{A}|X = x_1)$, she buys a bet on $X = x_1$ with a very small stake, and she will sell a bet on A if X should happen to take the value x_1. This last bet is described by $p'(\mathrm{A})$—the agent's degrees of belief after learning $X = x_1$—, and its stake is slightly higher than the conditional bet on A. Teller (1973) showed that such a system of bets leads to a sure loss for the agent: either she will lose

her bet on $X = x_1$ and the other two bets will be called off, or she wins this bet, but the gain will be compensated by the safe loss that the two other bets yield. See also Easwaran (2011a, 316).

One problem with the dynamic Dutch Book Argument consists in the fact that it requires the agent to fix in advance which bets she is going to accept in the future if she happens to learn a certain fact about the world. In other words, the dynamic Dutch Book Argument is a sanity check for the preferences and commitments of an agent, instead of a proof of the irrationality of following another updating rule. Moreover, the scope of Teller's argument and its successors (e.g., van Fraassen, 1989; Lewis, 1999) is often restricted.

As a tool for learning from experience, Bayesian Conditionalization is also somewhat restricted. For example, it does not describe how we update our degrees of belief in the light of information whose propositional status is unclear, e.g., indicative conditionals. We will address this particular challenge in the first variation. Moreover, sometimes we do not learn that evidence E has occurred with certainty, just that it is highly likely. For instance, a look at the weather forecast may shift our probability distribution over E = "it will rain tonight" and its negation from $p(\mathrm{E}) = 1/2$ to $p'(\mathrm{E}) = 9/10$. How should our belief in other propositions, such as H = "the sun will shine tomorrow" change in the face of such evidence? Of course, we could update on the second-order proposition that the probability of E has changed, but such a move would involve great complications and leave the object language L. And even then, the implications for the posterior probability of H would not be clear.

To solve this challenge, Jeffrey (1971) has argued that the posterior probability of hypothesis H after learning E, $p'(\mathrm{H})$, should obey the equation

$$p'(\mathrm{H}) = p'(\mathrm{E})\, p(\mathrm{H}|\mathrm{E}) + p'(\neg\mathrm{E})\, p(\mathrm{H}|\neg\mathrm{E}) \qquad \text{(JC)}$$

whenever the following two equations are satisfied:

$$p(\mathrm{H}|\mathrm{E}) = p'(\mathrm{H}|\mathrm{E}) \qquad p(\mathrm{H}|\neg\mathrm{E}) = p'(\mathrm{H}|\neg\mathrm{E}) \qquad \text{(Rigidity)}$$

The first equation computes the new degree of belief in H as the weighted average of the conditional degrees of belief of H given E and given ¬E, weighted with the degree of belief that E occurred. (JC) or **Jeffrey Conditionalization** follows from the Law of Total Probability together with (Rigidity). In a recent paper, Schwan and Stern (2016) argue that (Rigidity)

holds whenever E screens off H from the the propositional content D of the learning experience, that is, $p(\mathrm{H}|\mathrm{D}, \pm\mathrm{E}) = p(\mathrm{H}|\pm\mathrm{E})$. Obviously, Jeffrey Conditionalization reduces to Bayesian Conditionalization when E is known for certain, that is, when $p'(\mathrm{E}) = 1$.

Diaconis and Zabell (1982) have also demonstrated that Bayesian Conditionalization is just a special case of a more general updating principle, namely minimizing the Kullback-Leibler divergence between prior and posterior probability distribution under the constraint $p'(\mathrm{E}) = 1$. Variation 1 reproduces Diaconis and Zabell's proof, discusses their approach in more detail and applies it to learning conditional information. A similar result can be shown for Jeffrey Conditionalization when $p'(\mathrm{E}) < 1$. That is, Bayesian learning can be represented as conservative belief revision: Bayesians change their degrees only in so far as newly learned constraints on their degrees of belief (e.g., $p'(\mathrm{E}) = 1$) force them to do so. Or in other words, they stay as close as possible to their prior degrees of belief as these constraints allow them to do. This principle is also entrenched in non-quantitative theories of dynamic reasoning, such as the AGM-model of belief revision (Alchourrón et al., 1985), which operates on the binary level of belief and disbelief. Coherence with established reasoning models in other domains of science and philosophy may be seen as a distinct plus for Bayesian inference.

Hence, while it is difficult to give a fully compelling and conclusive justification of changing degrees of belief in a particular way, the above arguments provide a strong cumulative case for Bayesian Conditionalization. This is especially interesting since the motivations come from different directions: one comes from the operationalist, decision-theoretic corner, and another one from a principle of epistemic conservativity which is also used in qualitative models of belief revision.

Let us wrap up by adding a third dimension to Bayesian inference. We have talked a lot about the mathematical axioms that govern the statics and dynamics of rational degree of belief. But so far, we were silent on how degrees of belief should inform rational decisions. In many applications of Bayesian reasoning, it is emphasized that the goal of a Bayesian inference is the calculation of **posterior probabilities**. This is also the main result of many approaches to rational choice in economics: posterior probabilities are combined with subjective utilities in order to make an optimal choice (Savage, 1972). Under a certain set of assumptions on rational preferences,

it can be shown that all relevant information for making rational decisions (in the economic, instrumental sense of rationality) is contained in an agent's posterior probability distribution. This is the action-related dimension of Bayesian inference. We therefore identify the core of Bayesian inference as the conjunction of the following principles:

Static Dimension Rational degrees of belief of an agent are represented by probability functions.

Dynamic Dimension Bayesian Conditionalization (or some generalization thereof) prescribes how a rational agent should revise her degrees of belief.

Action Dimension The posterior probability distribution is the rational basis for assessing evidence, accepting hypotheses and making decisions.

At least one of these principles is endorsed by anyone who calls herself a Bayesian. However not all Bayesians, or scientists who apply Bayesian methods, agree with all three principles (for a survey, see Weisberg, 2009). Bayesian statisticians sometimes refuse to interpret probabilities in a subjective sense: **objective prior probabilities**, which are based on symmetry, invariance or information minimization principles (Jeffreys, 1961; Bernardo, 1979a; Vassend, 2016), do not represent any agent's degrees of belief and are supposed to screen off Bayesian inference from the charge of arbitrariness. What is more, Bayesian statisticians frequently use improper prior distributions which fail to sum up to one and violate the axioms of probability (Bernardo and Smith, 1994). However, objective Bayesians typically accept the other two principles—Bayesian Conditionalization and decision-making based on posterior probabilities.

Ironically, there are also varieties of objective Bayesian inference who accept the first, static principle and who reject the second, dynamic principle, that is, Bayesian Conditionalization (Jaynes, 1968, 2003; Williamson, 2007, 2010). Jon Williamson suggests the following approach: rational degrees of belief should satisfy the axioms of probability as a matter of coherence. Moreover, they should be in sync with empirical constraints, that is, knowledge about the external world. Such knowledge can consist in propositions that the agent has come to know. But it also includes the expectation and variance of random variables, or other constraints that

are difficult to express in a simple propositional language. For this reason, Williamson's approach is particularly apt for statistical reasoning. Of course, there will typically be more than one probability function which satisfies these constraints. Williamson recommends to choose the most equivocal, that is, the most middling distribution. This choice can be motivated in different ways, one of them including risk aversion. See Williamson (2007) and Williamson (2010, Chapter 2 and 3) for foundational motivation and Seidenfeld (1979, 1986) for two classic criticisms.

Finally, also the link between (posterior) degree of belief and decision-making need not be strict. It is possible to reason with subjective degrees of belief and to change them by Bayesian Conditionalization, but to make decisions in a different way, e.g., based on frequentist statistics. Indeed, scientists often admit that Bayesian inference is a foundationally sound framework for modeling rational degree of belief, but they prefer to work with frequentist or descriptive statistics. A reason for that preference is the difficulty of coming up with meaningful prior distributions, and because of concerns relating to scientific objectivity (e.g., US Food and Drug Administration, 2010; Gelman and Shalizi, 2012; Cumming, 2014; Trafimow and Marks, 2015). Similarly, one may decide to assess theories on behalf of their confirmational track record, which may be derived from a Bayesian framework without being equal to a theory's posterior probability. A recent representative of such an approach is Brössel (2016) who has introduced the concept of confirmation commitments (see also Hawthorne, 2005). More on the notion of confirmation will be said in Variation 2.

These remarks conclude our discussion of the foundations of Bayesian inference. We now introduce a powerful practical tool for making Bayesian inference: Bayesian networks. Philosophically, this does not add any substantial assumptions, but since Bayesian Networks will be one of our main tools in the remainder of the book, it is useful to explain the principles behind them.

Bayesian Networks

A Bayesian network is a directed acyclical graph (DAG) which represents conditional and unconditional probabilistic independencies between various propositions. It is a useful graphical tool for describing inferential relations between propositions—or in a causal reading, how events affect

each other. Indeed, they greatly facilitate causal inference in science (Pearl, 2000; Spirtes et al., 2000).

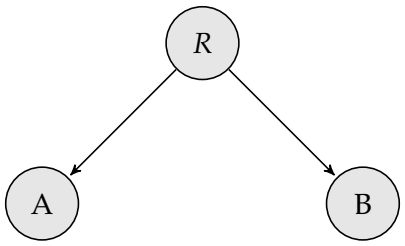

Figure 0.1: The Bayesian Network for the Risotto Example.

A canonical example of reasoning with a Bayesian Network is given in Figure 0.1. R represents the proposition that Alice and Bob have prepared a risotto from poisonous mushrooms. A represents the proposition that Alice has stomach pain after eating the risotto, and B represents the proposition that Bob has stomach pain after eating the risotto. Assume that there is a probability distribution $p(\cdot)$ over the propositional variables $A \in \{\mathrm{A}, \neg\mathrm{A}\}$, $B \in \{\mathrm{B}, \neg\mathrm{B}\}$ and $R \in \{\mathrm{R}, \neg\mathrm{R}\}$. Throughout the book, we follow the convention that propositional variables are printed in italic script, while their instantiations are printed in roman script (Bovens and Hartmann, 2003). The arrows in the graph then correspond to probabilistic dependencies between the variables.

In this graph, A would be called a *descendant* of R, and R would be a *parent* of A. Likewise for the relation between B and R. Also descendants of A and B would be among the descendants of R; however, R would not be their parent. Lack of an arrow between two nodes of the network indicates that the two variables do not depend on each other directly, but only via one or several intermediate variables. Conditional on these intermediate variables, they are probabilistically independent.

In general, the relationship between DAGs and probability distributions can be formalized as follows:

Parental Markov Condition The probability distribution $p(\cdot)$ is Markov relative to a directed acyclical graph G if and only if every variable is probabilistically independent of all its non-descendants in G, conditional on its parents.

That is, the probability distribution over A, B and R is Markov relative

to the graph in Figure 0.1 if and only if

$$p(\mathrm{A},\mathrm{B}|\mathrm{R}) = p(\mathrm{A}|\mathrm{R}) \cdot p(\mathrm{B}|\mathrm{R}) \quad (5)$$

$$p(\mathrm{A},\mathrm{B}|\neg\mathrm{R}) = p(\mathrm{A}|\neg\mathrm{R}) \cdot p(\mathrm{B}|\neg\mathrm{R}) \quad (6)$$

where R acts as the parent node of A, and B is a non-descendant of A. Our shorthand notation for the conditional independence of A and B, given R, is $(A \perp\!\!\!\perp B)|R$.

This property is plausible for the causal interpretation that we have given to the network. If we already know that Alice and Bob ate a poisonous mushroom risotto, learning about Alice's stomach pain does not raise or lower our probability that Bob has stomach pain. In other words, eating the poisonous mushroom risotto is a *common cause* of the stomach pain of both Alice and Bob. When the Bayesian network correctly represents the causal relations between different variables, with arrows denoting paths of causal influence, the Parental Markov Condition is transformed into the philosophically more substantive **Causal Markov Condition**: a phenomenon is independent of its noneffects, given its direct causes.

The Parental Markov Condition specifies the constraints that the relations between the nodes in the Bayesian Network G place on the probability distribution $p(\cdot)$. This allows for easily reading off the conditional independencies. Moreover, with the help of the Parental Markov Condition, we can calculate joint and marginal probabilities in a straightforward way. For instance, in the above example, we can use Equations (5) and (6) to write $p(\mathrm{A},\mathrm{B},\mathrm{R})$ as

$$p(\mathrm{A},\mathrm{B},\mathrm{R}) = p(\mathrm{A},\mathrm{B}|\mathrm{R})p(\mathrm{R}) = p(\mathrm{A}|\mathrm{R})p(\mathrm{B}|\mathrm{R})p(\mathrm{R}) \quad (7)$$

and analogously for all other conjunctions of $\pm A$, $\pm B$, and $\pm R$. Similarly, the marginal probability of A (and likewise for B) can be written as

$$\begin{aligned} p(\mathrm{A}) &= \sum_{\pm B,\pm R} p(\mathrm{A},\mathrm{B},\mathrm{R}) \\ &= \sum_{\pm B,\pm R} p(\mathrm{A}|\mathrm{R})p(\mathrm{B}|\mathrm{R})p(\mathrm{R}) \end{aligned}$$

where the sum is taken over the the different possible values of B and R (here: true or false). We have used the law of total probability in the first line and Equation (7) in the second line.

The above equations suggest that a joint or marginal probability can always be reduced to a combination of probabilities conditional on parent variables and probabilities of root variables, that is, variables that do not have parents. Indeed, in general, it will always be the case that for a graph G with variables $\{A_1, \ldots, A_n\}$:

$$p(\mathrm{A_1}, \ldots, \mathrm{A_n}) = \prod_{i=1}^{n} p(\mathrm{A_i}|\mathrm{Par}(\mathrm{A_i}))$$

That is, if we reason about probabilities in a Bayesian network, it suffices to know the base rates of the root variables and the conditional probability of any variable given its parents. Often, these values are much easier to elicit than the joint or marginal probabilities.

The scope of applications of Bayesian Networks in science is huge, and it goes beyond the scope of this extremely brief introduction to list even the most important ones. In fact, our use of Bayesian Networks in this book will remain on an elementary level: to represent causal relations and conditional independencies between propositional variables, and to calculate joint and marginal probabilities in an efficient way. We now articulate our view of Bayesian philosophy of science.

Bayesian Philosophy of Science

In a classical introduction to Bayesian inference, it is claimed that "scientific reasoning is essentially reasoning in accordance with the formal principles of probability" (Howson and Urbach, 2006, xvii)—see also Earman (1992, 142). Personally, we find this claim too strong: as an impressive number of works in philosophy of science have shown, successful scientific reasoning patterns are extremely diverse and vary with the disciplinary context where they are applied, as well as with the type of problems they address (e.g., Hempel, 1965; Cartwright, 1979; van Fraassen, 1980; Hacking, 1983). There are also principled reasons why a purely probabilistic logic of scientific inference would have to be incomplete (Norton, 2016). What we hope to show in this book is that there are *some* general aspects of scientific reasoning that are well captured by Bayesian inference.

In other words: describing scientists as having degrees of belief which are updated by Bayesian Conditionalization helps us to better understand how they reason about their theories, why they accept some and reject others. In particular, we are interested in Bayesian models of cognitive

values—that is, values that are characteristic of a good scientific theory (Kuhn, 1977a; McMullin, 1982; Douglas, 2009a, 2013)—such as predictive accuracy, explanatory power and simplicity.

We use a spectrum of Bayesian models across the book, and we do so in a straightforwardly eclectic fashion. Some of our models are based on Bayesian Conditionalization and others on more general forms of belief change. Sometimes we import models from Bayesian statistics, sometimes from the philosophical literature on Bayesian inductive logic. To our mind, this is not a problem. After all, we are not interested in defending Bayesian inference as the uniquely *correct* theory of epistemic attitudes, but in showing that it is a *fruitful* one. Hence, does Bayesian inference provide unexpected insights into scientific reasoning? Does it solve problems that other models struggle with? Does it suggest interesting experiments or questions for future research? We believe that Bayesian inference fares well with respect to those criteria which are also standardly used for the evaluation of scientific models (Weisberg, 2007, 2012; Frigg and Hartmann, 2012). However, no single Bayesian model will be able to succeed at modeling phenomena as diverse as scientific confirmation, explanation, intertheoretic reduction and causal effect. Diverse phenomena ask for a diversity of (Bayesian) models, and of course, Bayesian models may often have to be complemented by non-Bayesian approaches.

We understand Bayesian philosophy of science as **the use of Bayesian principles and methods for modeling scientific reasoning**. It involves two different goals and methods: on the one hand, we explicate central concepts of scientific reasoning in a Bayesian language; on the other hand, we apply Bayesian inference to scientific reasoning, e.g., by building Bayesian models of the No Alternative Argument and the No Miracles Argument.

In the **explicative project**, we follow Carnap's methodology of replacing a vague concept, the *explicandum*, by an exact one, the *explicatum*:

> If a concept is given as explicandum, the task consists in finding another concept as its explicatum which fulfils the following requirements to a sufficient degree.
>
> (1) The explicatum is to be *similar* to the explicandum in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close

similarity is not required, and considerable differences are permitted.

(2) The characterization of the explicatum, that is, the rules of its use (for instance, in the form of a definition), is to be given in an *exact* form, so as to introduce the explicatum into a well-connected system of scientific concepts.

(3) The explicatum is to be a *fruitful* concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a non-logical concept, logical theorems in the case of a logical concept).

(4) The explicatum should be as *simple* as possible; this means as simple as the more important requirements (1), (2), and (3) permit. (Carnap, 1950, 7)

In the context of this book, this means that we provide a quantitative dimension for central concepts in scientific reasoning, such as confirmation, explanatory power and causal effect. Explication involves a tight interconnection of conceptual analysis and formal methods: conceptual analysis leads us to adequacy conditions which the explicatum has to satisfy, while formal reasoning helps us to characterize the set of explicata that satisfy these conditions, and to prove existence and uniqueness theorems.

Bayesian model-building, on the other hand, takes a more applied focus. Rather than explicating a particular concept by the axiomatic method of adequacy conditions and representation theorems, we identify a set of variables that matter for a peculiar case of scientific reasoning (e.g., the No Miracles Argument). Then we postulate relations between them and we investigate what kind of interesting findings we can make on the basis of these assumptions. The fact that our models are framed in terms of rational degrees of belief makes them Bayesian. This means that our work in this book has a more foundational side, connected to the explicative project, and a more applied side, connected to Bayesian model-building. Notably, both the explicative and the model-building project make use of empirical and computational methods where appropriate: experimental findings are evaluated in order to judge the adequacy of an explicatum with respect to the concept that it targets, and computational methods are used for exploring the consequences of our models in cases where we fail to achieve strictly analytical solutions.

An Outline of the Book

We conclude this introduction with an overview of the chapters of the book, which may be thought of as variations on the themes presented in this exposition.

The first five variations center around a common theme: the confirmation of scientific theories. Scientific theories are valued to the extent that they make accurate predictions, and degree of confirmation quantifies the extent to which theories have been predictively successful. Measuring degree of confirmation is a classical task for Bayesian philosophy of science, since confirmation can be straightforwardly explicated in terms of increase in probability. But our approach is broader: we also address challenges to Bayesian Confirmation Theory, and we demonstrate how certain argument patterns in science (e.g., the No Alternatives Argument and the No Miracles Argument) can be recast as confirmatory arguments for the theory in question.

Variation 1 describes how learning conditional information (e.g., if intervention X is made, result Y will occur) may confirm or disconfirm a scientific theory. For instance, how should the belief in a theory T change if we learn that it makes a particular prediction (e.g., $p(\mathrm{E}|\mathrm{T}) = 1$)? To solve this challenge, we use a generalization of Bayesian Conditionalization and conceptualize rational degree of belief change as minimizing the divergence between prior and posterior distribution, conditional on preserving the causal and inferential structure of the involved propositions. This allows us to capture several (counter)examples that haunt other accounts of learning conditional information.

Variation 2 is devoted to a quantitative analysis of confirmation in Bayesian terms, in particular confirmation as increase in firmness: evidence E confirms theory T if and only if it raises the (subjective) probability of T. We motivate and describe the transition from qualitative theories of confirmation in first-order logic to quantitative, Bayesian models of confirmation. We characterize these models and explain their advantages vis-à-vis qualitative models, especially with respect to classical challenges such as the paradox of the ravens, the tacking paradox and the grue paradox. This brings us to an analysis of the problems of Bayesian confirmation theory, such as the measure sensitivity of confirmation-theoretic analysis, and more generally, the plurality of Bayesian confirmation measures.

We finally discuss how conceptual analysis and empirical evidence can be combined to narrow down the class of adequate confirmation measures.

Variation 3 discusses one of the major challenges to Bayesian confirmation theory: the Problem of Old Evidence. How do Bayesians describe the confirmatory power of the discovery that a theory T implies evidence E when E has been known for a long time? According to the standard Bayesian model of confirmation, evidence E confirms theory T if and only if learning E raises the probability of T. But this is impossible if the evidence is already known ($p(\mathrm{E}) = 1$). We resolve this problem by means of two different Bayesian models that demonstrate how explaining old evidence raises the rational degree of belief in theory T.

Variation 4 deals with the No Alternatives Argument. Does the failure to find alternatives to a scientific theory confirm it? Arguments of this kind are often employed in support of string theory or other theories that lack strong empirical support (e.g., in paleontology). After all, there have been enormous efforts to find a viable alternative, and the failure to find one may license an explanatory inference for the truth (or empirical adequacy) of that theory. By framing the argument within a probabilistic model, we can show that longstanding failure to find alternatives supports a theory—even if the strength of the argument (i.e., the degree of confirmation it provides) is context-sensitive and depends on the exact circumstances.

Variation 5 develops a probabilistic assessment of the famous No Miracles Argument in favor of scientific realism. That argument contends, in a nutshell, that the truth of scientific theories is the only viable explanation of their success. We frame the No Miracles Argument as a confirmatory argument: does the success of scientific theories make them more probable? We set up various Bayesian models to answer this question, which also correspond to different ways of interpreting the No Miracles Argument. These models take into account factors that have been neglected in reconstructions of the No Miracles Argument: the stability of theories in a specific discipline, and their success rate. Thus we get a better grip on the circumstances when the success of science supports realist inclinations, and when it doesn't.

The second set of variations abandons the topic of confirmation in favor of central concepts in scientific reasoning. Some of them (e.g., explanatory power, simplicity, corroboration, intertheoretic reduction) are also often cited as virtues of a good scientific theory.

Variation 6 develops a Bayesian analysis of causal effect, building on the scientific literature on causal Bayes nets and combining it with methods from Bayesian confirmation theory. First, we defend the choice of a framework where different measures can be embedded and compared: causal Bayes nets. Second, we derive representation theorems for various measures of causal effect, that is, theorems that characterize a measure of causal effect as the only measure (up to ordinal equivalence) that satisfies a certain set of adequacy conditions. Third, we make an argument for preferring a particular measure. Finally, we apply that measure to a case from epidemiology, demonstrating how closely scientific and philosophical reasoning about causal effect are intertwined.

Variation 7 is devoted to the topic of explanatory power, a classical cognitive value in scientific reasoning. Hempel (1965) famously postulated a structural identity between prediction and explanation: explanations show why a particular phenomenon occurred by deriving it from the theory, and explanatory power is proportional to the ability of the explanans to account for the explanandum (see also Hempel and Oppenheim, 1948). We explore to what extent this classical view, fallen out of fashion in modern philosophy of science, can be rescued in a Bayesian framework, where explications of explanatory power are based on considerations of statistical relevance. Then we compare several of these explanatory power measures and their respective strengths and weaknesses.

Variation 8 provides a Bayesian account of intertheoretic reduction. Ceteris paribus, theories which have broad scope and cohere with theories at other levels of description have more value than isolated theories. For example, models of statistical mechanics reduce to thermodynamics equations in the mathematical limit. Such reductive relationships between theories at the phenomenal and fundamental level are described by the models by Nagel (1961) and Schaffner (1967). We defend these models against the standard criticism and show how the establishment of reductive relationships can raise the probability of the involved theories. That is, we do not only show how reduction unifies different theories, but we also demonstrate that it has a positive effect on the assessment of the involved theories.

Variation 9 brings together Bayesian models of theory assessment with (frequentist) hypothesis testing in science and Popper's critical rationalism. In hypothesis tests, we often observe a failure to reject the null (=de-

fault) hypothesis at a statistically significant level. Does this mean that the null hypothesis is confirmed, or as Popper said, corroborated by the results? Can we characterize the conditions when corroboration takes place, and give a quantitative dimension to corroboration judgments? We first show why a confirmation-theoretic framework cannot provide such an explication. Then we derive an axiomatic measure of corroboration from a more general set of probabilistic constraints and relate it back to principles of Bayesian inference.

Variation 10 analyzes the value of simplicity in statistical inference. We analyze the general question of whether simplicity is a good reason to prefer a theory to an alternative in the context of statistical model selection. In particular, we are concerned with Forster and Sober's (1994) thesis that simpler models are more likely to be true, or at least predictively accurate. We analyze model selection criteria that could support their claim, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) and compare them to genuinely Bayesian methods, such as model selection on the basis of Bayes factors. We demonstrate that a link between simplicity and predictive success cannot be established on the basis of these criteria. Furthermore, we show that Bayesian methods are often used in an instrumental way that is detached from the philosophical foundations of Bayesian inference.

Variation 11, finally, deals with the question of whether subjective Bayesian inference can ever achieve a sufficient degree of objectivity to counter the charge of arbitrariness, and to maintain the epistemic authority of science. In particular, the irreducibly subjective nature of prior probabilities, and their inevitable impact on (supposedly objective) measures of evidence is often cited as a reason to mistrust Bayesian inference. However, such arguments often presuppose an unrealistically strong and outdated conception of scientific objectivity. Therefore, our strategy for countering this criticism is twofold: we combine an up-to-date conceptual analysis of scientific objectivity with formal arguments that Bayesian inference is, on these accounts, no less objective than its competitors.

The book concludes with a short recapitulation of the original theme: we count the successes and failures of Bayesian philosophy of science, make up the balance, and sketch future research projects.

Variation 1: Learning Conditional Information

Indicative conditionals of the form "if A, then B" constitute a substantial part of scientific evidence. Many experiments report that a certain intervention leads to a certain effect. If sugar is thrown into a glass of water, it dissolves. If a mouse is infected with a certain virus, it develops a certain type of symptoms. If a consumer is frequently exposed to a commercial, he or she may be more likely to buy the advertized product.

Scientific evidence thus often comes in the form of conditional statements. But how should we change our degrees of belief in scientific theories when we learn such conditionals? This question has prompted a large amount of literature, but without conclusive results. Douven (2012) concludes that a general and feasible account of learning conditionals is still to be formulated. Indeed, all accounts that have been proposed so far face problems. Here are three popular attempts.

First and most straightforwardly, one might identify the natural language indicative conditional $\mathrm{H} \rightarrow \mathrm{E}$ with the material conditional $\mathrm{H} \supset \mathrm{E}$, which is equivalent to $\neg\mathrm{H} \vee \mathrm{E}$. Then, one may conditionalize on that proposition. Popper and Miller (1983) challenged this proposal with an argument based on the probability calculus. It goes as follows. Consider two propositions H and E and a probability distribution p with $0 < p(\mathrm{H}) < 1$ and $p(\mathrm{E}|\mathrm{H}) < 1$. We now learn the indicative conditional $\mathrm{H} \rightarrow \mathrm{E}$, which we express as the material conditional $\neg\mathrm{H} \vee \mathrm{E}$. To update our beliefs, we use Bayesian Conditionalization, i.e. we calculate the posterior probability $p'(\mathrm{H}) := p(\mathrm{H}|\neg\mathrm{H} \vee \mathrm{E})$. Interestingly, it turns out that $p'(\mathrm{H}) < p(\mathrm{H})$. The proof is elementary; so we reproduce it here. Bayes' Theorem implies that

$$p(\mathrm{H}|\neg\mathrm{H} \vee \mathrm{E}) = p(\neg\mathrm{H} \vee \mathrm{E}|\mathrm{H}) \cdot \frac{p(\mathrm{H})}{p(\neg\mathrm{H} \vee \mathrm{E})}$$

and hence, it is sufficient to show that $p(\neg \mathrm{H} \vee \mathrm{E}|\mathrm{H}) < p(\neg \mathrm{H} \vee \mathrm{E})$:

$$\begin{aligned} p(\neg \mathrm{H} \vee \mathrm{E}|\mathrm{H}) - p(\neg \mathrm{H} \vee \mathrm{E}) &= p(\neg \mathrm{H}|\mathrm{H}) + p(\mathrm{E}|\mathrm{H}) - p(\neg \mathrm{H} \wedge \mathrm{E}|\mathrm{H}) \\ &\quad -(1 - p(\mathrm{H}) + p(\mathrm{H} \wedge \mathrm{E})) \\ &= p(\mathrm{E}|\mathrm{H}) - (1 - p(\mathrm{H})) - p(\mathrm{E}|\mathrm{H})p(\mathrm{H}) \\ &= p(\mathrm{E}|\mathrm{H})\,(1 - p(\mathrm{H})) - (1 - p(\mathrm{H})) \\ &= (1 - p(\mathrm{H}))\,(p(\mathrm{E}|\mathrm{H}) - 1) \\ &< 0 \end{aligned}$$

In other words, learning "if E, then H" always decreases the probability of H if one interprets the above sentence as the material conditional. However, there are also cases where the posterior probability of a hypothesis is intuitively judged to be greater than or equal to the prior probability upon learning that it makes a particular prediction. We give some examples in Section 1.2. The naïve Bayesian proposal to identify learning an indicative conditionals with learning the associated material conditional has trouble to account for such evidence: it does not do justice to the variety of conditional information that we encounter in nature.

Second, David Lewis (1976) proposed an account called *imaging*, which requires a possible worlds semantics with similarity relations between different possible worlds. On Lewis' account, an indicative conditional is true if the consequens holds true in the closest possible world where the antecedens is true. It turns out, however, that this proposal also fails to do justice to some of our intuitive judgments (cf. Douven and Dietz, 2011). Moreover, Lewis' imaging is an account of the *semantics* of conditionals; it is unclear how it should guide our *reasoning* with conditionals and the revision of our beliefs.

Third, one may conclude that Bayesian Conditionalization is, as an updating rule, too restricted to account for learning indicative conditionals, or more generally, conditional information. Does this mean farewell to Bayesian philosophy of science? Not necessarily so. We may try to find a conservative extension of Bayesian Conditionalization: an updating rule which preserves the probabilistic nature of degrees of belief, which agrees with Bayesian Conditionalization for learning first-order propositions, and which *also* covers the learning of conditionals. This is our approach in this variation. We propose **minimizing Kullback-Leibler divergence between posterior and the prior probability distribution** as an extension

of Bayesian Conditionalization that is able to account for the persistent problem of learning conditional information (e.g., indicative conditionals). More precisely, we argue that minimizing divergence between prior and posterior probability distribution delivers intuitively correct results for learning conditional information if the posterior distribution is also required to **respect causal and inferential constraints** provided by the context of the examples in question.

The remainder of this variation is organized as follows. Section 1.1 introduces KL divergence minimization and shows its equivalence to Bayesian Conditionalization for updating on standard evidence (i.e., learning a first-order proposition E). Section 1.2 challenges this account by developing three examples which the divergence minimization method struggles to model adequately. Section 1.3 shows how these challenges can be met if the divergence minimization method is extended and properly applied. The entire variation, and this section in particular, builds on the theoretical innovations presented in Hartmann and Rafiee Rad (2016) and applies them to scientific reasoning. Finally, Section 1.4 takes stock and comments on the scope of our proposal while Section 1.5 contains the proofs of our results. From now on, we drop the adjective "indicative" and the noun "conditional" is always meant to refer to an indicative conditional.

The Kullback-Leibler Divergence and Probabilistic Updating

Bayesian Conditionalization is a powerful, but somewhat limited tool for changing one's belief in the light of new evidence. As we have motivated at the beginning of this variation, not all scientific evidence comes in the form of a first-order proposition that we can easily learn and represent in our object language. Indicative conditionals, whose propositional status is very much disputed in the literature (e.g., Edgington, 1995, 2014), are a case in point. Other examples are learning the mean value of a random variable, learning measurement variance, etc. It is far from clear how conditionalization on sentences such as "the mean of variable X is between 13.2 and 15.4" should work. This calls for a more general updating rule where these sentences, instead of being the objects of conditionalization, constrain probability functions that represent our rational degrees of be-

lief.

The Kullback-Leibler (KL) divergence $D_{KL}(p'||p)$ plays such a constraining role. It has been introduced in the context of transmitting electric signals using a binary code (Shannon, 1949). In the next two paragraphs, we motivate the particular mathematical form of KL divergence. Notably, the results in this variation also hold for divergence functions that differ from KL divergence, e.g., Hellinger distance. Hence, nothing substantial depends on the choice of this particular measure.

Kullback-Leibler divergence is relevant in the context of finding an cost-efficient code for transmitting a string S whose tokens (e.g., letters of the alphabet) occur with varying frequency, expressed by probability distribution p'. The less tokens we need to transmit the string, the better. Assume that we use a binary code C where frequently occurring letters such as "e" and "a" are coded by short sequences of bits such as "0" and "10", and infrequently occurring letter such as "x" are coded by long sequences such as "11111110". The length of the bit sequences implicitly defines a probability distribution p over the elements of our code: namely that distribution of tokens which is optimally encoded by C. In the above example, where a '0' denotes transition to the next token, that would be $p(\text{“}e\text{''}) = 1/2$, $p(\text{“}a\text{''}) = 1/4$, $p(\text{“}x\text{''}) = 1/64$. You get the idea. KL divergence measures the loss in efficiency by using a code with probability distribution p, when the real frequency of the tokens in S follows distribution p'. Efficiency is expressed by the expected excess length of transmitting S by our code C, instead of using an optimal code.

Assume that $s_1, \ldots, s_n$ denote the tokens of the string S which we try to transmit. How should we measure the loss in efficiency when transmitting S with code p although the true frequency distribution of the s_i is described by p'? The Kullback-Leibler divergence calculates this difference as follows:

$$D_{KL}(p'||p) := \sum_{i=1}^{n} p'(s_i) \log \frac{p'(s_i)}{p(s_i)} \,. \tag{1.1}$$

In this formular, we quantify the loss in efficiency for each token s_i by means of the logarithmic difference $\log p'(s_i) - \log p(s_i)$, which is equivalent to $\log(p'(s_i)/p(s_i))$, and we use the actual frequency of each token for weighting these losses. If the logarithm is taken at base two, this means that $D_{KL}(p'||p)$ expresses the expected number of bytes that we will have to invest for transmitting S with p instead of the optimal code p'.

Kullback-Leibler divergence can be generalized beyond the information-theoretic context and be regarded as a general expression of divergence between two probability distributions. Note that the base of the logarithm will not matter for the results that we show in this variation. Moreover, $D_{KL}(p'||p)$ will be zero if and only if p' and p agree on all elements of the state space.

The Kullback-Leibler divergence implicitly defines a method for updating degrees of belief: when learning a set of constraints $\mathcal{E}$, one should adopt the probability distribution $p'(\cdot)$ which has, among all distributions that satisfy $\mathcal{E}$, the smallest KL-divergence to the old distribution $p(\cdot)$. In other words, we change beliefs as much as required by the learned set of empirical constraints, but no more: like theories of belief revision such as AGM (Alchourrón et al., 1985; Makinson, 1985), minimizing Kullback-Leibler divergence is inherently conservative. Moreover, the divergence minimization method can also handle much more general constraints than Bayesian Conditionalization. Yet, the two forms of updating are closely related and often equivalent, as we shall show in this section. This result was first proved by Diaconis and Zabell (1982).

Consider two binary propositional variables, H and E. We represent the probabilistic dependence between H and E in the Bayesian Network depicted in Figure 1. To complete it, we fix the prior probability of the root node H, i.e.

$$h := p(\mathrm{H}) \tag{1.2}$$

and the conditional probabilities of E, given the values of its parent H:

$$p := p(\mathrm{E}|\mathrm{H}) \quad , \quad q := p(\mathrm{E}|\neg\mathrm{H}) \tag{1.3}$$

Next, we learn that the evidence E obtains. This is a constraint on the

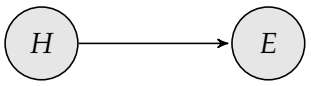

Figure 1.1: The Bayesian Network representation of the relation between H and E.

posterior probability distribution p' which amounts to

$$p'(\mathrm{E}) = 1\,. \tag{1.4}$$

Now we make analogous definitions for the variables h', p' and q':

$$h' := p(\mathrm{H})$$

$$p' := p'(\mathrm{E}|\mathrm{H}) \qquad\qquad q' := p'(\mathrm{E}|\neg\mathrm{H})$$

Calculating the Kullback-Leibler divergence between p' and p yields

$$\begin{aligned} D_{KL}(p'||p) &:= \sum_{H,E} p'(H,E) \log \frac{p'(H,E)}{p(H,E)} \\ &= h' \log \left(\frac{h'}{h\,p} \right) + \overline{h'} \log \left(\frac{\overline{h'}}{\overline{h}\,q} \right) \\ &= h' \log \frac{h'}{h} + \overline{h'} \log \frac{\overline{h'}}{\overline{h}} + h' \log \frac{q}{p} + \log \frac{1}{q}. \end{aligned} \tag{1.5}$$

where we have used the convenient shorthand $\overline{h} := 1 - h$, which we will use throughout the book. We have also used the shorthand notation $p(\mathrm{H}, \mathrm{E})$ for $p(\mathrm{H} \wedge \mathrm{E})$ which we will also use below when appropriate.

With the help of Equation (1.5), we can show the following theorem:

Theorem 1.1 (Diaconis and Zabell, 1982) *Let $p(\cdot)$ be a probability distribution over propositions of a propositional language L. Suppose we learn sentence* E*. Then the following two updating rules for the posterior distribution p' are equivalent:*

1. *Bayesian Conditionalization on* E*: $p'(\cdot) = p(\cdot|\mathrm{E})$.*
2. *Minimizing Kullback-Leibler divergence $D_{KL}(p'||p)$ subject to the constraint that $p'(\mathrm{E}) = 1$.*

This result, although far from being novel, is deep and interesting. It shows that Bayesian Conditionalization on proposition E is the updating method which minimizes the change from the current probability distribution if (i) the posterior probability distribution p' is constrained by $p'(\mathrm{E}) = 1$; (ii) the amount of change is measured by Kullback-Leibler divergence. This result can be interpreted as saying that Bayesian Conditionalization does not require us to change our beliefs more than necessary.

Let us now explore whether this method can also be applied to finding a suitable probability distribution after having learned a conditional. To apply the proposed method, one has to derive a probabilistic statement from the learned conditional. Here we follow Douven (2012) and others and represent learning the conditional $\mathrm{H} \rightarrow \mathrm{E}$ as a constraint on the conditional posterior probability $p'(\mathrm{E}|\mathrm{H})$. In particular, we assume that learning $\mathrm{H} \rightarrow \mathrm{E}$ implies $p'(\mathrm{E}|\mathrm{H}) = 1$. That is, we are not interested in

the probability of the conditional itself. Instead, we are interested in *the effects of learning a conditional on the relevant conditional probabilities*. This assumption is compatible with agnosticism about the propositional status of conditionals. It is also weaker than Stalnaker's Thesis which identifies the probability of a conditional with its conditional probability (Stalnaker, 1968, 1970, 1975), and which is vulnerable to triviality arguments in the style of Lewis (1976). Our approach can also be motivated from the Ramsey test for conditional probabilities, which evaluates them by hypothetically adding the antecedens to the background knowledge—see page 11 in the introduction. In other words, if we already know $\mathrm{H} \to \mathrm{E}$ and add H to our stock of background knowledge, E will be a certainty by Modus Ponens. Hence, $p(\mathrm{E}|\mathrm{H}) = 1$.

To test this method, consider the Bayesian Network depicted in Figure 1.1. In this scenario, we learn the conditional $\mathrm{H} \to \mathrm{E}$, which implies that

$$p'(\mathrm{E}|\mathrm{H}) := p' = 1. \tag{1.6}$$

The Kullback-Leibler divergence between p' and p is then given by

$$\begin{aligned} D_{KL}(p'||p) &:= \sum_{H,E} p'(H,E) \log \frac{p'(H,E)}{p(H,E)} \\ &= h' \log\left(\frac{h'}{h\,p}\right) + \overline{h'}\left(q' \log\left(\frac{\overline{h'}\,q'}{\overline{h}\,q}\right) + \overline{q'} \log\left(\frac{\overline{h'}\,\overline{q'}}{\overline{h}\,\overline{q}}\right)\right) \qquad (1.7) \\ &= h' \log\frac{h'}{h} + \overline{h'} \log\frac{\overline{h'}}{\overline{h}} + h' \log\frac{1}{p} + \overline{h'}\left(q' \log\frac{q'}{q} + \overline{q'} \log\frac{\overline{q'}}{\overline{q}}\right). \end{aligned}$$

A simple algebraic proof then suffices for the following theorem:

Theorem 1.2 *Let* E, H *be two sentences of a propositional language L with probability distribution* $p(\cdot)$. *Suppose we learn* $\mathrm{H} \to \mathrm{E}$ *and we construct the posterior probability distribution by minimizing Kullback-Leibler divergence* $D_{KL}(p'||p)$ *in* p', *subject to the constraint that* $p'(\mathrm{E}|\mathrm{H}) = 1$. *Then, if* $p(\mathrm{E}|\mathrm{H}) < 1$, *it will be the case that* $p'(\mathrm{H}) < p(\mathrm{H})$.

This result may sound wrong at first sight. After all, we only learn that H has E as a consequence and nothing else. So why should this prompt us to change our belief in H? And why should the probability of H decrease? Note, however, that H becomes more informative after having learned the conditional. It makes a prediction on E. It is therefore natural to set the new probability of H to a lower value as more informative hypotheses take

more risk to be mistaken. This point was already made by Popper (2002) and Hempel and Oppenheim (1948). They stressed that being informative is a scientific virtue which may contribute to the acceptance of H, but it is not correlated with posterior probability, quite to the contrary. The result of Theorem 1.2 agrees, by the way, with Popper and Miller's diagnosis for learning the material conditional $H \supset E$ by means of Bayesian Conditionalization (Popper and Miller, 1983).

We have seen that minimizing Kullback-Leibler divergence leads to reasonable results for situations involving two propositional variables. But does it also work for more complicated scenarios?

Three Challenges for Minimizing Divergence

In a variety of papers, Richard Dietz, Igor Douven and Jan-Willem Romeijn developed examples which challenge the divergence minimization method for learning conditionals. Below, we adapt those examples to a context of scientific reasoning. Each example starts with a story that sets up the scene. Then a conditional is learned which may prompt some previously held beliefs to change.

1. **The Medicine Example.** A general practitioner has to choose whether or not to give drug D to a patient. She will administer D if and only if (i) D is effective against the strains of bacteria that the patient is infected with; (ii) the patient has no medical condition that makes him susceptible to serious side effects of D. The GPs's assistant checks the patient's medical history and tells her boss: "If D is effective against that strain of bacteria, then we should administer D." Upon learning this conditional, the GP does not change her belief in the efficacy of D—rather, she reasons that her assistant has checked whether the patient is sensitive to side effects of D. Thus learning the conditional leaves the probability of the antecedens unchanged. This counterexample to Theorem 1.2 is adapted from Douven and Romeijn (2011).

2. **The Astronomy Example.** The astronomic community in the 16th century considered two general theories, the Copernican model and the Ptolemaic model. An astronomer observes the movements of the planets Mars, Jupiter and Saturn over an extended period and notes down his observations. He finds an agreement between periods of

retrograde motion and relative brightness. He now works himself through the implications of the Copernican model and he realizes: "If the Copernican model is true, then the outer planets (=Mars, Jupiter and Saturn) will display retrograde motion when they are close to Earth." Already knowing that the (apparent) retrograde motion of these planets would agree with his actual observations because brightness is an indicator of spatial proximity, he now finds it more likely that the Copernican model is true. In this example, learning the conditional information should intuitively increase the probability of the antecedens of the conditional. This example is adapted from Douven and Dietz (2011).

3. **The Economics Example.** An economist is interested in whether a country is recovering economically. During the Christmas period, she surveys the sales volume of several warehouses. It turns out to be low. She asks a colleague about the consequences of economic recovery on consumer income. Her colleague answers: "If there is an economic recovery going on, consumer's income has increased", e.g., because of generous end-of-the-year bonuses. Upon learning this conditional, the economist thinks it is doubtful (even if not wholly excluded) whether an economic recovery is currently going on. As a result, she lowers her degree of belief for economic recovery and thus decreases the probability of the antecedens of the conditional. This example is adapted from Douven (2012).

These three cases describe different ways how learning a conditional may affect the probability of the antecedens: it may be lowered, increased, or remain unchanged. This is bad news for the divergence minimization method which claims that the probability of the antecedens always decreases—see Theorem 1.1. Does this mean that the project of finding a general theory of learning conditionals is futile or doomed? We disagree and show how the divergence minimization method can successfully deal with the above examples when plausible causal and inferential constraints are imposed on the posterior probability distribution.

Meeting the Challenges

To meet the three challenges presented in the previous section, we adopt the following methodology. First, we identify all relevant variables of the problem at hand and the causal and inferential relations that hold between them. Second, we represent causal and inferential relations between the variables by (conditional) independencies in a Bayesian Network and fix the prior probability distribution p that is associated with that network. Third, we express the learned conditional as a constraint on the posterior probability distribution p' and *assume that the relevant independencies are not changed by learning the conditional*. That is, they are constraints on the prior and the posterior probability distribution, e.g., because they express a certain causal structure. From the story it is clear that the incoming information does not overturn the structure; hence, we should preserve it as a constraint on the posterior probability distribution. Fourth, we obtain the posterior probability distribution p' by minimizing the Kullback-Leibler divergence $D_{KL}(p'||p)$ to the prior distribution p. Fifth, we check whether the results comply with our intuitions. To repeat, in comparison to standard updating by minimizing Kullback-Leibler divergence, we have now imposed the additional constraint on the posterior distribution that causal structure (e.g., which interventions affect which variables) and inferential relations (e.g., which variables are probabilistically independent of others) remain intact.

The Medicine Example

We introduce three binary propositional variables to represent the medicine example. The variable E has the values E: "Drug D is effective for the bacteria the patient is infected with", and ¬E: "Drug D is not effective for these bacteria". The variable S has values S: "The patient is susceptible to serious side effects when taking drug D", and ¬S: "The patient is not susceptible to serious side effects when taking drug D". Finally, the variable A has the values A: "Administer drug D", and ¬A: "Do not administer drug D".

Before we proceed, let us show that using the material conditional and Bayesian Conditionalization leads to an intuitively wrong result. To do so, remember that the learned conditional is "if drug D is effective against these bacteria, then administer it", expressed as E $\rightarrow$ A. We interpret this

as $\mathrm{E} \supset \mathrm{A}$ which is equivalent to $\neg\mathrm{E} \vee \mathrm{A}$. Assuming $0 < p(\mathrm{E}), p(\mathrm{E}, \neg\mathrm{A}) < 1$ and using the ratio analysis of conditional probability, we then obtain

$$\begin{aligned} p(\mathrm{E}|\mathrm{E} \supset \mathrm{A}) &= p(\mathrm{E}|\neg\mathrm{E} \vee \mathrm{A}) = \frac{p(\mathrm{E} \wedge (\neg\mathrm{E} \vee \mathrm{A}))}{p(\neg\mathrm{E} \vee \mathrm{A})} = \frac{p(\mathrm{E} \wedge \mathrm{A})}{p(\neg\mathrm{E} \vee \mathrm{A})} \\ &= \frac{p(\mathrm{E}) - p(\mathrm{E}, \neg\mathrm{A})}{1 - p(\mathrm{E}, \neg\mathrm{A})}. \end{aligned} \tag{1.8}$$

It is then easy to verify that Equation (1.8) requires $p'(\mathrm{E}) = p(\mathrm{E}|\mathrm{E} \supset \mathrm{A}) < p(\mathrm{E})$, which conflicts with our intuitive judgment that the probability of the efficacy of the drug should remain unchanged.

Let us now show how our suggested methodology deals with the case. The story suggests a number of dependencies and independencies between the various variables. The Bayesian Network in Figure 3 represents the probabilistic dependencies and independencies between these variables. The arrow represent the causal relations and the effect of interventions on the variables.

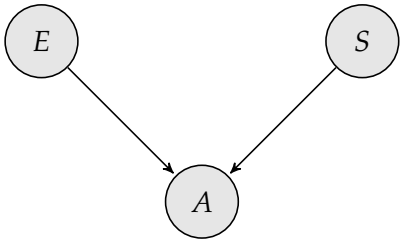

Figure 1.2: The Bayesian Network for the Neuroscience Example.

To complete the Bayesian Network, we have to fix the prior probability of the root nodes and the conditional probabilities of all other nodes, given the values of their parents. We set

$$e := p(\mathrm{E}) \qquad\qquad s := p(\mathrm{S})$$

and

$$\begin{aligned} \alpha &:= p(\mathrm{A}|\mathrm{E}, \mathrm{S}) & \beta &:= p(\mathrm{A}|\mathrm{E}, \neg\mathrm{S}) \\ \gamma &:= p(\mathrm{A}|\neg\mathrm{E}, \mathrm{S}) & \delta &:= p(\mathrm{A}|\neg\mathrm{E}, \neg\mathrm{S}) \end{aligned}$$

Given the punch line of the story, we may assume that

$$\beta = p(\mathrm{A}|\mathrm{E}, \neg\mathrm{S}) = 1\,. \tag{1.9}$$

That is, if the drug is effective and the patient not susceptible to side effects, we will administer the drug. All other conditional probabilities (i.e. β, γ and δ) are in the open interval $(0, 1)$. Let us now consider the posterior probability distribution p', which is defined over the same variables as the prior distribution. The constraint $\beta = 1$, due to Equation (1.9), should be preserved in that distribution, too. Hence we conclude

$$\beta' := p'(\mathrm{A}|\mathrm{E}, \neg\mathrm{S}) = 1 \qquad \overline{\beta'} := p'(\neg\mathrm{A}|\mathrm{E}, \neg\mathrm{S}) = 0 \tag{1.10}$$

Another constraint on the posterior probability distribution is the learned conditional "if the drug is effective against these bacteria, then administer it", which implies that

$$p'(\mathrm{A}|\mathrm{E}) = 1 \tag{1.11}$$

and hence $p'(\neg\mathrm{A}|\mathrm{E}) = 0$. Assuming that all unconditional probabilities are in the open interval (0,1), we can apply the ratio analysis of conditional probability and infer that $p'(\neg\mathrm{A}, \mathrm{E}) = 0$ which implies in turn that

$$\begin{aligned} p'(\neg\mathrm{A}, \mathrm{E}, \neg\mathrm{S}) &= p'(\neg\mathrm{A}|\mathrm{E}, \neg\mathrm{S})\, p'(\mathrm{E})\, p'(\neg\mathrm{S}) = 0 \\ p'(\neg\mathrm{A}, \mathrm{E}, \mathrm{S}) &= p'(\neg\mathrm{A}|\mathrm{E}, \mathrm{S})\, p'(\mathrm{E})\, p'(\mathrm{S}) = 0 \end{aligned}$$

The first equation is satisfied since $\overline{\beta'} = p'(\neg\mathrm{A}|\mathrm{E}, \neg\mathrm{S}) = 0$, as we noted in Equation (1.10). Regarding the second equation, one of the factors on the right hand side must be zero. We can safely assume that $p'(\mathrm{E}) > 0$ (why should learning the conditional $\mathrm{E} \rightarrow \mathrm{A}$ rule out that the drug is effective?) and also that $\overline{\alpha'} = p'(\neg\mathrm{A}|\mathrm{E}, \mathrm{S}) > 0$: if the patient is susceptible to side effects, it is not clear whether the GP still administers drug D. Hence $s' := p'(\mathrm{S}) = 0$. This makes sense: the information received from the assistant suggests that the patient is not susceptible to serious side effects. We can now show the following theorem:

Theorem 1.3 *Consider the Bayesian Network in Figure 1.2 with the prior probability distribution p from equations* (1.41)*. We furthermore assume that*

(i) the posterior probability distribution p' is defined over the same Bayesian Network and respects the same independence assumptions;

(ii) the learned conditional is modeled as the constraint (1.9) *on p', that is, $p(\mathrm{A}|\mathrm{E}, \mathrm{S}) = 1$, implying $p'(\mathrm{S}) = 0$;*

(iii) p' *minimizes the Kullback-Leibler divergence to* p.

Then $p'(\mathrm{E}) = p(\mathrm{E})$.

We conclude that the proposed method yields the intuitively correct result in this case: the probability of the drug being effective is invariant under learning the conditional $\mathrm{E} \to \mathrm{A}$.

The Astronomy Example

Again, we introduce three binary propositional variables to represent the astronomy example. The variable C has values C: "The Copernican model is true", and ¬C: "The Copernican model is false". The variable M has the values M: "The outer planets display retrograde motion when close to Earth", and ¬M: "The outer planets do not display retrograde motion when close to Earth". Finally, the variable O has the values O: "Periods of retrograde motion and relative brightness agree", and ¬O: "Periods of retrograde motion and relative brightness do not agree". The Bayesian Network in Figure 4 represents the probabilistic dependencies and independencies between these variables: O depends on C only via M, or in other words: $O \perp\!\!\!\perp C|M$.

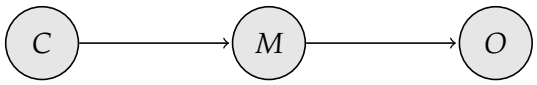

Figure 1.3: The Bayesian Network for the Astronomy Example.

To complete the Bayesian Network, we have to fix the prior probability of C, i.e.

$$\alpha := p(\mathrm{C}), \tag{1.12}$$

and the conditional probabilities

$$\begin{aligned} p_1 &:= p(\mathrm{M}|\mathrm{C}) & \qquad q_1 &:= p(\mathrm{M}|\neg\mathrm{C}) \\ p_2 &:= p(\mathrm{O}|\mathrm{M}) & \qquad q_2 &:= p(\mathrm{O}|\neg\mathrm{M}) \end{aligned}$$

We can now calculate the prior probability distribution over the variables C, M and O:

$$\begin{aligned} p(\mathrm{C},\mathrm{M},\mathrm{O}) &= \alpha\, p_1\, p_2 \quad , \quad p(\mathrm{C},\mathrm{M},\neg\mathrm{O}) = \alpha\, p_1\, \overline{p_2} \\ p(\mathrm{C},\neg\mathrm{M},\mathrm{O}) &= \alpha\, \overline{p}_1 q_2 \quad , \quad p(\mathrm{C},\neg\mathrm{M},\neg\mathrm{O}) = \alpha\, \overline{p}_1\, \overline{q}_2 \end{aligned}$$

$$\begin{aligned} p(\neg C, M, O) &= \overline{\alpha}\, q_1\, p_2 \quad , \quad p(\neg C, M, \neg O) = \overline{\alpha}\, q_1\, \overline{p_2} \\ p(\neg C, \neg M, O) &= \overline{\alpha}\, \overline{q}_1 q_2 \quad , \quad p(\neg C, \neg M, \neg O) = \overline{\alpha}\, \overline{q}_1\, \overline{q}_2 \end{aligned} \tag{1.13}$$

Next we learn two items of information. First, we learn that O obtains. Assuming that the conditional independencies depicted in Figure 4 do not change, this means that we learn that

$$p'(O) = \alpha' \left(p'_1\, p'_2 + \overline{p'}_1 q'_2\right) + \overline{\alpha'} \left(q'_1\, p'_2 + \overline{q'}_1 q'_2\right) = 1\,, \tag{1.14}$$

where we have replaced all variables by the corresponding primed variables. Second, we learn the conditional "if the Copernican model is true, then the outer planets will display retrograde motion when they are close to Earth", which implies that

$$p'(M|C) = p'_1 = 1. \tag{1.15}$$

Inserting Equation (1.15) into Equation (1.14), we obtain

$$\alpha'\, p'_2 + \overline{\alpha'} \left(q'_1\, p'_2 + \overline{q'_1}\, q'_2\right) = 1. \tag{1.16}$$

This equation only holds for $\alpha' \in (0, 1)$, if

$$p'_2 = 1 \tag{1.17}$$

and if

$$q'_1\, p'_2 + \overline{q'_1}\, q'_2 \equiv q'_1 + \overline{q'_1}\, q'_2 = 1. \tag{1.18}$$

It has the solutions (i) $q'_1 = 1$ and (ii) $q'_2 = 1$. As solution (i) does not make sense (if the Copernican model is false, the retrograde motion pattern stays unexplained), we conclude that

$$q'_2 = 1. \tag{1.19}$$

Equations (1.17) and (1.19) make sure that $p'(O) = 1$. Inserting conditions (1.15), (1.17) and (1.19) into the analogues of equations (1.13), we can calculate the posterior probability distribution:

$$\begin{aligned} p'(C, M, O) &= \alpha' \quad , \quad p'(C, M, \neg O) = 0 \\ p'(C, \neg M, O) &= 0 \quad , \quad p'(C, \neg M, \neg O) = 0 \\ p'(\neg C, M, O) &= \overline{\alpha'}\, q'_1 \quad , \quad p'(\neg C, M, \neg O) = 0 \\ p'(\neg C, \neg M, O) &= \overline{\alpha'}\, \overline{q'}_1 \quad , \quad p'(\neg C, \neg M, \neg O) = 0 \end{aligned} \tag{1.20}$$

We can now show the following theorem:

Theorem 1.4 *Consider the Bayesian Network in Figure 1.3 with the prior probability distribution from Equation* (1.13)*. Let*

$$\eta := \frac{p_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2}. \tag{1.21}$$

We furthermore assume that

(i) the posterior probability distribution p' is defined over the same Bayesian Network;

(ii) the learned information constrains p' via Equations (1.14) *and* (1.15)*;*

(iii) p' minimizes the Kullback-Leibler divergence to p.

Then $p'(\mathrm{C}) > p(\mathrm{C})$ if and only if $\eta > 1$.

That is, our rational degree of belief in the Copernican model is increased when the condition $\eta > 1$ holds. When will this be the case? This depends on how we flesh out the details of the historical story and the background assumptions of the astronomer. The prior degree of belief in the Copernican model might have been small back in the days since the Copernican model did not produce more accurate predictions than the Ptolemaic model and did not provide explanations for many physical phenomena, such as the movement of the Earth. Hence α is small. Moreover, the observed correlation between brightness and retrograde motion is njot explained by any alternative model which speaks for a small probability of

$$p(\mathrm{O}) = \alpha\,(p_1\, p_2 + \overline{p_1}\, q_2) + \overline{\alpha}\,(q_1\, p_2 + \overline{q_1}\, q_2) \tag{1.22}$$

As α is small and $\overline{\alpha}$ is large, we conclude that $\epsilon := q_1\, p_2 + \overline{q_1}\, q_2$ must be small, too.

From the story it is also clear that $p_2 = p(\mathrm{O}|\mathrm{M})$ is fairly large: Given the postulated relation between a planet's position and the pattern of retrograde motion, agreement between brightness and retrograde motion is to be expected. At the same time, q_2 will be very small as there is no reason to assume such a striking agreement if planets do not display retrograde motion pattern when close to Earth. Finally, p_1 may not be very large, but the previous considerations suggest that $p_1 \gg \epsilon$. We conclude that

$$\eta = \frac{p_1}{\epsilon} \cdot p_2 \tag{1.23}$$

will typically be greater than 1. If $\eta \leq 1$, then the probability of N will not increase after learning the two pieces of information.

We conclude that the proposed method yields the intuitively correct result in this case. Of course, the exact result will depend of the specific details of the story, but this appears to be a very sensible feature of our approach: we have already seen before that contextual factors may determine whether learning a conditional raises or lowers the probability of the antecedens.

The Economics Example

Finally, we turn to the economics example. To represent the scenario, we introduce the following propositional variables. The variable R has the values R: "An economic recovery is going on", and ¬R: "No economic recovery is going on". The variable I has the values I: "Consumer income is increased", and ¬I: "Consumer income is not increased". The variable S has the values S: "The level of spending in warehouses is low", and ¬S: "The level of spending in warehouses is high". The Bayesian Network in Figure 5 represents the probabilistic dependencies and independencies between these variables, as well as their causal relations. Note that the Bayesian Network in Figure 5 has the same structure as the Bayesian Network in Figure 4. Our calculation therefore proceeds as in the previous example.

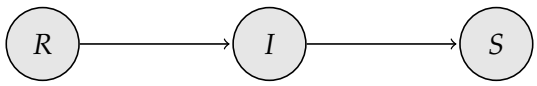

Figure 1.4: The Bayesian Network for the Economics Example.

To complete the Bayesian Network, we have to fix the prior probability of R, i.e.

$$p(\mathrm{R}) = r, \tag{1.24}$$

and the conditional probabilities

$$p_1 := p(\mathrm{I}|\mathrm{R}) \qquad q_1 := p(\mathrm{I}|\neg\mathrm{R})$$
$$p_2 := p(\mathrm{S}|\mathrm{I}) \qquad q_2 := p(\mathrm{S}|\neg\mathrm{I})$$

We can now calculate the prior probability distribution over the vari-

ables R, I and S:

$$\begin{aligned} p(\mathrm{R},\mathrm{I},\mathrm{S}) &= r\, p_1\, p_2 \quad , \quad p(\mathrm{R},\mathrm{I},\neg\mathrm{S}) = r\, p_1\, \overline{p_2} \\ p(\mathrm{R},\neg\mathrm{I},\mathrm{S}) &= r\, \overline{p_1} q_2 \quad , \quad p(\mathrm{R},\neg\mathrm{I},\neg\mathrm{S}) = r\, \overline{p_1}\, \overline{q_2} \\ p(\neg\mathrm{R},\mathrm{I},\mathrm{S}) &= \overline{r}\, q_1\, p_2 \quad , \quad p(\neg\mathrm{R},\mathrm{I},\neg\mathrm{S}) = \overline{r}\, q_1\, \overline{p_2} \\ p(\neg\mathrm{R},\neg\mathrm{I},\mathrm{S}) &= \overline{r}\, \overline{q_1}\, q_2 \quad , \quad p(\neg\mathrm{R},\neg\mathrm{I},\neg\mathrm{S}) = \overline{r}\, \overline{q_1}\, \overline{q_2} \end{aligned} \tag{1.25}$$

Next we learn two items of information and our probability distribution changes from p to p'. First, we learn that S obtains. Assuming that the causal structure depicted in Figure 5 does not change, this means that we learn that

$$p'(\mathrm{S}) = r'\,(p'_1\, p'_2 + \overline{p'}_1 q'_2) + \overline{r'}\,(q'_1\, p'_2 + \overline{q'}_1 q'_2) = 1\,, \tag{1.26}$$

where we have replaced all variables by the corresponding primed variables. Second, we learn the conditional "if there is an economic recovery going on, consumers income is increased", which implies that

$$p'(\mathrm{I}|\mathrm{R}) = p'_1 = 1. \tag{1.27}$$

Inserting Equation (1.27) into Equation (1.26), we obtain:

$$r'\, p'_2 + \overline{r'}\, \left(q'_1\, p'_2 + \overline{q'_1}\, q'_2\right) = 1 \tag{1.28}$$

This equation only holds for $r' \in (0,1)$, if

$$p'_2 = 1 \tag{1.29}$$

and if

$$q'_1\, p'_2 + \overline{q'_1}\, q'_2 \equiv q'_1 + \overline{q'_1}\, q'_2 = 1. \tag{1.30}$$

It has the solutions (i) $q'_1 = 1$ and (ii) $q'_2 = 1$. As solution (i) does not make sense—why should we be certain that consumer income is increased?—, we conclude that

$$q'_2 = 1. \tag{1.31}$$

Inserting conditions (1.27), (1.29) and (1.31) into the analogues of Equation (1.25), we can calculate the posterior probability distribution:

$$\begin{aligned} p'(\mathrm{R},\mathrm{I},\mathrm{S}) &= r' \quad , \quad p'(\mathrm{R},\mathrm{I},\neg\mathrm{S}) = 0 \\ p'(\mathrm{R},\neg\mathrm{I},\mathrm{S}) &= 0 \quad , \quad p'(\mathrm{R},\neg\mathrm{I},\neg\mathrm{S}) = 0 \\ p'(\neg\mathrm{R},\mathrm{I},\mathrm{S}) &= \overline{r'}\, q'_1 \quad , \quad p'(\neg\mathrm{R},\mathrm{I},\neg\mathrm{S}) = 0 \end{aligned} \tag{1.32}$$

$$p'(\neg \mathrm{R}, \neg \mathrm{I}, \mathrm{S}) = \overline{r'}\,\overline{q_1'} \qquad , \qquad p(\neg \mathrm{R}, \neg \mathrm{I}, \neg \mathrm{S}) = 0$$

The structure of the relevant Bayesian Network in the economics example is the same as in the previous astronomy example—see Figure 1.3 and 1.4. Hence, we can apply Theorem 1.4. Whether or not the probability of the antecedens R is raised by learning the conditional depends on whether or not

$$\eta := \frac{p_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2} > 1.$$

In this case, we have evidence to the contrary. It is clear from the story that $q_2 \gg p_2$: the probability of low spending is higher for increased than for unchanged consumer income. Hence,

$$\eta < \frac{p_1 p_2}{q_1 p_2 + \overline{q_1} p_2} = \frac{p_1 p_2}{p_2} = p_1 < 1. \tag{1.33}$$

We conclude that the posterior probability of economic recovery is smaller than the prior probability. Hence, the proposed method again yields the intuitively correct result.

Discussion

Minimizing the Kullback-Leibler divergence between prior and posterior probability distribution, subject to a set of empirical constraints, is an interesting extension of Bayesian Conditionalization. First, it has a wider scope and allows for normatively attractive and computationally feasible learning of a wide range of constraints on the posterior probability distribution, such as the mean or variance of a random variable. The ability to process such evidence is an important feature of any theory of scientific inference. Second, whenever we perform Bayesian Conditionalization on a first-order proposition, minimizing Kullback-Leibler divergence will deliver the same result. Learning by minimizing Kullback-Leibler divergence is thus a conservative extension of Bayesian Conditionalization which does not threaten core Bayesian principles. Rather, it enlarges the scope of Bayesian reasoning in science. Using the divergence minimization method, we can address and resolve challenges that have been put forward against Bayesian Conditionalization and Bayesian reasoning in general.

This variation has applied divergence minimization to learning conditional information. We have focused on evidence in the form of indicative conditionals, e.g., that a certain manipulation reliably yields a certain

result, or that a scientific theory has certain observational consequences. Learning these conditionals is hard to represent in the ordinary Bayesian mechanism, as shown by Popper and Miller's paradoxical results for learning material conditionals. Without delving further into the epistemology of conditionals, we assume that learning a conditional $\mathrm{H} \rightarrow \mathrm{E}$ imposes a constraint on our posterior distribution p': namely that the conditional probability of E, given H, is equal to one. If one would like to attack our work in this variation, one could either doubt that learning a conditional imposes the constraint $p'(\mathrm{E}|\mathrm{H}) = 1$ on the posterior distribution, or require that learning the conditional implies *more* constraints. To us, none of these options look particularly appealing.

The divergence minimization method can now be applied to minimizing the divergence between p and p', subject to the constraint $p'(\mathrm{E}|\mathrm{H}) = 1$. However, direct application of that method does not take account of the variety of the inferences that we make when learning a conditional: sometimes the probability of the antecedens is raised, sometimes it is lowered, sometimes is stays equal.

To deal with all three cases, we have suggested a refinement of the divergence minimization method that adequately deals with these cases: represent the causal and inferential relations among the involved propositions by a Bayesian network with a set of conditional and unconditional independencies. These independencies act as constraints on both the prior and posterior distributions. After all, in the discussed examples, they concern elements of the background story and are not changed by learning the conditional. When Kullback-Leibler divergence is minimized subject to these constraints, the intuitively correct results follow.

Does the proposed method also give the adequate results if more complicated scenarios are considered? We do not see a way how to answer this question in full generality. The set of possible scenarios where conditionals are learned is unrestricted, and one cannot do much apart from studying them case by case. We are, however, optimistic that the proposed method will work for more complicated scenarios involving more than three variables, as our examples represent diverse cases of probabilistic dependencies. The logic behind our approach is simple and intuitive: in moving from a prior to a posterior distribution, one should not only minimize the distance subject to novel evidence, but also subject to those constraints which do not change under the learning conditional informa-

tion. This concerns in particular the causal and inferential structure of the model, e.g., the variables which are affected by an intervention and the set of probabilistic independencies. Whenever the learned evidence does not change these relations, our model provides a general and adequate method of Bayesian updating. We conclude that the scope of evidence that Bayesian reasoners can model is wider than those captured by Bayesian Conditionalization (i.e., first-order propositions). This observation rebuts a large number of criticisms raised against the Bayesian research program in philosophy of science.

Proofs of the Theorems

Auxiliary Lemmata

The following three lemmata will be useful for the proofs presented in the remainder of this Appendix.

Lemma 1: *Let* $f(x) := \log(a\,x), g(x) := x\log(a\,x)$ *and* $h(x) := \overline{x}\log(a\,\overline{x})$. *Then the first derivatives are:* $f'(x) = 1/x, g'(x) = 1 + \log(a\,x)$ *and* $h'(x) = -1 - \log(a\,\overline{x})$.

Proof: Trivial.

Lemma 2: *The function* $f(x) := x\log\frac{x}{x'} + \overline{x}\log\frac{\overline{x}}{\overline{x'}}$ *has a minimum at* $x = x'$.

Proof: Using Lemma 1, we obtain

$$f'(x) = \log\left(\frac{x}{\overline{x}}\cdot\frac{\overline{x'}}{x'}\right). \tag{1.34}$$

Setting this expression equal to zero (i.e. the argument of the logarithm equal to 1), one obtains $x = x'$. As $f''(x) = 1/(x\,\overline{x}) > 0$ for all $x \in (0,1)$, we have indeed found the minimum.

Lemma 3: *Consider the equation* $x'/\overline{x'} = k \cdot x/\overline{x}$ *with* $k > 0$. *Then (i)* $x' > x$ *iff* $k > 1$, *(ii)* $x' = x$ *iff* $k = 1$ *and (iii)* $x' < x$ *iff* $k < 1$.

Proof: This follows from the observation that the function $\varphi(x) := x/\overline{x}$ is strictly monotonically increasing for $x \in (0,1)$.

We now proceed to proving our main results.

Proof of the Theorems

Proof of Theorem 1.1: Let H be any closed sentence of L. The joint posterior probability distribution over H and E will have the following form:

$$\begin{aligned} p'(\mathrm{H},\mathrm{E}) &= h'\,p' \quad , \quad p'(\mathrm{H},\neg\mathrm{E}) = h'\,\overline{p'} \\ p'(\neg\mathrm{H},\mathrm{E}) &= \overline{h'}\,q' \quad , \quad p'(\neg\mathrm{H},\neg\mathrm{E}) = \overline{h'}\,\overline{q'}\,, \end{aligned} \tag{1.35}$$

where we have replaced all variables by the corresponding primed variables. The constraint $p'(E) = 1$ and Equation (1.35) then entail that

$$h'\,p' + \overline{h'}\,q' = 1 \tag{1.36}$$

and, taking into account that all four atoms in Equation (1.35) sum up to 1, that

$$h' \, \overline{p'} = \overline{h'} \, \overline{q'} = 0. \tag{1.37}$$

It is easy to see that Equation (1.37) only holds for all $h' \in (0,1)$ if $p' = q' = 1$. In this case, Equation (1.36) is automatically fulfilled for all h'. The posterior probability distribution then simplifies as follows:

$$\begin{aligned} p'(\mathrm{H},\mathrm{E}) &= h' \quad , \quad p'(\mathrm{H},\neg\mathrm{E}) = 0 \\ p'(\neg\mathrm{H},\mathrm{E}) &= \overline{h'} \quad , \quad p'(\neg\mathrm{H},\neg\mathrm{E}) = 0 \end{aligned} \tag{1.38}$$

To determine the value of h', we differentiate the Kullback-Leibler divergence (see Equation (1.5)) between p' and p with respect to h' and obtain after some algebra:

$$\frac{\partial D_{KL}}{\partial h'} = \log\left(\frac{h'}{\overline{h'}} \cdot \frac{\overline{h}}{h} \cdot \frac{q}{p}\right) \tag{1.39}$$

To find the minimum, we set the latter expression equal to zero (i.e., we set the argument of the logarithm equal to 1) and obtain:

$$h' = \frac{h\,p}{h\,p + \overline{h}\,q}$$

In more familiar form, this can be written as

$$p'(\mathrm{H}) = p(\mathrm{H}|\mathrm{E}),$$

where the right hand side is the posterior probability distribution that follows from Bayesian Conditionalization. The posterior distribution obtained from minimizing Kullback-Leibler divergence subject to the constraint $p'(E) = 1$ is equal to the distribution obtained by conditionalizing on E. To complete the proof, we convince ourselves that

$$\frac{\partial^2 D_{KL}}{\partial h'^2} = \frac{1}{h' \, \overline{h'}} > 0 \tag{1.40}$$

for all $h' \in (0,1)$, which shows that we have indeed found the unique minimum of $D_{KL}(p'||p)$. Hence, Bayesian Conditionalization follows from minimizing the Kullback-Leibler divergence between posterior and prior probability distribution, if one considers the learned information as a constraint on the posterior. □

Proof of Theorem 1.2: The proof runs analogous to the above theorem. To find the minimum of $D_{KL}(p'||p)$, we first differentiate this expression with respect to q' and obtain

$$\frac{\partial D_{KL}}{\partial q'} = \overline{h'} \log\left(\frac{q'}{\overline{q'}} \cdot \frac{\overline{q}}{q}\right).$$

Next, we set this expression equal to zero and obtain $q' = q$. With this, we simplify D_{KL} and obtain

$$D_{KL}(p'||p) = \left(h' \log\frac{h'}{h} + \overline{h'} \log\frac{\overline{h'}}{\overline{h}}\right) + h' \log\frac{1}{p}.$$

Next, we differentiate $D_{KL}(p'||p)$ with respect to h' and obtain

$$\frac{\partial D_{KL}}{\partial h'} = \log\left(\frac{h'}{\overline{h'}} \cdot \frac{\overline{h}}{h} \cdot \frac{1}{p}\right).$$

Setting this expression equal to zero yields

$$\frac{h'}{\overline{h'}} = p \cdot \frac{h}{\overline{h}},$$

and hence

$$h' = \frac{hp}{hp + \overline{h}}.$$

Using Lemma 3, we conclude from Equation (1.5) that $h' < h$, if $0 < p < 1$. □

Proof of Theorem 1.3: With the constraint that $\beta = p(\mathrm{A}|\mathrm{E}, \neg\mathrm{S}) = 1$, the prior probability distribution over A, E and S takes the following form:

$$\begin{aligned}
p(\mathrm{A},\mathrm{E},\mathrm{S}) &= \alpha es & p(\neg\mathrm{A},\mathrm{E},\mathrm{S}) &= \overline{\alpha} es \\
p(\mathrm{A},\mathrm{E},\neg\mathrm{S}) &= e\overline{s} & p(\neg\mathrm{A},\mathrm{E},\neg\mathrm{S}) &= 0 \\
p(\mathrm{A},\neg\mathrm{E},\mathrm{S}) &= \gamma\overline{e}s & p(\neg\mathrm{A},\neg\mathrm{E},\mathrm{S}) &= \overline{\gamma e}s \\
p(\mathrm{A},\neg\mathrm{E},\neg\mathrm{S}) &= \delta\overline{es} & p(\neg\mathrm{A},\neg\mathrm{E},\neg\mathrm{S}) &= \overline{\delta es}
\end{aligned} \tag{1.41}$$

Taking into account the learned conditional $p'(\mathrm{A}|\mathrm{E}) = 1$ and the condition $p'(\mathrm{S}) = 0$ that we derived from that equation, the posterior distribution looks as follows:

$$p'(\mathrm{A},\mathrm{E},\mathrm{S}) = 0 \qquad p'(\neg\mathrm{A},\mathrm{E},\mathrm{S}) = 0$$

$$\begin{aligned} p'(\mathrm{A},\mathrm{E},\neg\mathrm{S}) &= e' & p'(\neg\mathrm{A},\mathrm{E},\neg\mathrm{S}) &= 0 \\ p'(\mathrm{A},\neg\mathrm{E},\mathrm{S}) &= 0 & p'(\neg\mathrm{A},\neg\mathrm{E},\mathrm{S}) &= 0 \\ p'(\mathrm{A},\neg\mathrm{E},\neg\mathrm{S}) &= \delta'\overline{e'} & p'(\neg\mathrm{A},\neg\mathrm{E},\neg\mathrm{S}) &= \overline{\delta e'} \end{aligned} \tag{1.42}$$

Calculating the Kullback-Leibler divergence between p' and p, we obtain

$$\begin{aligned} D_{KL}(p'||p) &:= \sum_{\pm A,E,S} p'(A,E,S)\cdot\log\left(\frac{p'(A,E,S)}{p(A,E,S)}\right) \\ &= e'\log\left(\frac{e'}{e\,\overline{s}}\right)+\overline{e'}\,\delta'\log\left(\frac{\overline{e'}\,\delta'}{\overline{e}\,\delta\,\overline{s}}\right)+\overline{e'}\,\overline{\delta'}\log\left(\frac{\overline{e'}\,\overline{\delta'}}{\overline{e}\,\overline{\delta}\,\overline{s}}\right) \\ &= e'\log\frac{e'}{e}+\overline{e'}\log\frac{\overline{e'}}{\overline{e}}+\overline{e'}\left(\delta'\log\frac{\delta'}{\delta}+\overline{\delta'}\log\frac{\overline{\delta'}}{\overline{\delta}}\right)+\log\frac{1}{\overline{s}} \end{aligned}$$

Next, we differentiate this expression with respect to e' and δ' and obtain

$$\begin{aligned} \frac{\partial D_{KL}}{\partial e'} &= \log\left(\frac{e'}{\overline{e'}}\cdot\frac{\overline{e}}{e}\right)-\left(\delta'\log\frac{\delta'}{\delta}+\overline{\delta'}\log\frac{\overline{\delta'}}{\overline{\delta}}\right) \\ \frac{\partial D_{KL}}{\partial \delta'} &= \overline{e'}\log\left(\frac{\delta'}{\overline{\delta'}}\cdot\frac{\overline{\delta}}{\delta}\right). \end{aligned}$$

Setting the expression in Equation (1.43) equal to zero, we obtain

$$\delta' = \delta. \tag{1.43}$$

Substituting this result into Equation (1.43), we obtain

$$\frac{\partial D_{KL}}{\partial e'} = \log\left(\frac{e'}{\overline{e'}}\cdot\frac{\overline{e}}{e}\right). \tag{1.44}$$

Setting the expression in Equation (1.44) equal to zero, we finally obtain $e' = e$. To show that we have indeed found a minimum, we calculate the Hessian matrix of D_{KL} at $(e',\delta') = (e,\delta)$ and obtain

$$H(D_{KL})|_{e,\delta} = \begin{pmatrix} 1/\overline{e} & 0 \\ 0 & e/(\delta\,\overline{\delta}) \end{pmatrix}. \tag{1.45}$$

This matrix is positive definite, which completes the proof of Theorem 1.3. □

Proof of Theorem 1.4: With the prior probability distribution from Equation (1.13) and the posterior probability distribution from Equation

(1.20), we obtain for the Kullback-Leibler divergence between P' and P:

$$\begin{aligned} D_{KL}(P'||P) &:= \sum_{C,M,O} p'(C,M,O) \cdot \log\left(\frac{p'(C,M,O)}{p(C,M,O)}\right) \\ &= c' \log\left(\frac{c'}{c\, p_1\, p_2}\right) + \overline{c'}\, q_1' \log\left(\frac{\overline{c'}\, q_1'}{\overline{c}\, q_1\, p_2}\right) + \overline{c'}\, \overline{q'}_1 \log\left(\frac{\overline{c'}\, \overline{q'}_1}{\overline{c}\, \overline{q_1}\, q_2}\right) \\ &= c' \log\frac{c'}{c} + \overline{c'} \log\frac{\overline{c'}}{\overline{c}} + \overline{c'}\left(q_1' \log\left(\frac{q_1' p_1}{q_1}\right)\right. \\ &\quad \left. + \overline{q_1'} \log\left(\frac{\overline{q_1'} p_1 p_2}{\overline{q_1} q_2}\right)\right) + \log\frac{1}{p_1 p_2} \end{aligned}$$

Next, we calculate the first derivatives of $D_{KL}(P'||P)$ with respect to c' and q_1' and obtain after some algebra:

$$\begin{aligned} \frac{\partial D_{KL}}{\partial c'} &= \log\left(\frac{c'}{\overline{c'}} \cdot \frac{\overline{c}}{c} \cdot \frac{1}{k_0}\right) - q_1' \log\left(\frac{q_1'}{\overline{q_1'}} \cdot \frac{\overline{q_1}\, q_2}{q_1\, p_2}\right) \\ \frac{\partial D_{KL}}{\partial q_1'} &= \overline{c'} \log\left(\frac{q_1'}{\overline{q_1'}} \cdot \frac{\overline{q_1}\, q_2}{q_1\, p_2}\right) \end{aligned}$$

with

$$k_0 := \frac{p_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2}. \tag{1.46}$$

To minimize $D_{KL}(P'||P)$ we first set (1.46) equal to zero (noting that $c' \in (0,1)$) and obtain

$$q_1' = \frac{q_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2}. \tag{1.47}$$

With this, we simplify the expression in Equation (1.46) and obtain

$$\frac{\partial D_{KL}}{\partial c'} = \log\left(\frac{c'}{\overline{c'}} \cdot \frac{\overline{c}}{c} \cdot \frac{1}{k_0}\right). \tag{1.48}$$

Setting now also the expression in Equation (1.48) to zero, we obtain

$$\frac{c'}{\overline{c'}} = k_0 \cdot \frac{c}{\overline{c}}. \tag{1.49}$$

Using Lemma 3, we conclude that $c' > c$ iff $k_0 > 1$. This completes the proof of Theorem 1.4. (We skip the proof that the corresponding Hessian is positive definite if Equations (1.47) and (1.49) hold.) □

Finally, it is interesting to see that conditionalizing on B and the material conditional $\mathrm{C} \supset \mathrm{M} \equiv \neg\mathrm{C} \vee \mathrm{M}$ yields the same result in this case.

$$\begin{aligned} p(C|C \supset M, O) &= p(\mathrm{C}|\mathrm{O} \wedge (\neg\mathrm{C} \vee \mathrm{M})) = \frac{p(\mathrm{C} \wedge \mathrm{O} \wedge (\neg\mathrm{C} \vee \mathrm{M}))}{p(\mathrm{O} \wedge (\neg\mathrm{C} \vee \mathrm{M}))} \\ &= \frac{p(\mathrm{O} \wedge \mathrm{C} \wedge \mathrm{M})}{p((\mathrm{O} \wedge \neg\mathrm{C}) \vee (\mathrm{O} \wedge \mathrm{M}))} = \frac{p(\mathrm{O}, \mathrm{C}, \mathrm{M})}{p(\mathrm{O}, \neg\mathrm{C}) + p(\mathrm{O}, \mathrm{M}) - p(\mathrm{O}, \neg\mathrm{C}, \mathrm{M})} \\ &= \frac{p(\mathrm{O}, \mathrm{C}, \mathrm{M})}{p(\mathrm{O}, \neg\mathrm{C}) + p(\mathrm{O}, \mathrm{C}, \mathrm{M})}. \end{aligned}$$

With the Bayesian Network depicted in Figure 4 and the prior probability distribution from Equation (1.13), we then obtain

$$p(C|C \supset M, O) = \frac{c\, p_1\, p_2}{c\, p_1\, p_2 + \overline{c}\,(q_1\, p_2 + \overline{q_1}\, q_2)} = \frac{c\, k_0}{c\, k_0 + \overline{c}} \equiv c' = p'(\mathrm{C}). \tag{1.50}$$

From this equation it is easy to see that $P^*(\mathrm{C}) > p(\mathrm{C})$ iff $k_0 > 1$. Hence, both procedures yield exactly the same result in this case.

Variation 2: Confirmation

Confirmation of scientific theories by empirical evidence is a central element of scientific reasoning. Their acceptance and rejection is often based on the track record of experiments that confirmed or undermined them. Eddington's observations of the 1919 solar eclipse confirmed Einstein's General Theory of Relativity (GTR) and strongly contributed to the endorsement of GTR among theoretical physicists. Equally spectacularly, a huge set of observations by CERN researchers confirmed the existence of the Higgs Boson, a fundamental particle hypothesized in the 1960s. In economics, Maurice Allais and Daniel Ellsberg conducted experiments about decision-making under uncertainty that undermined the empirical basis of Rational Choice Theory. But what are the conditions when a piece of evidence confirms or undermines a theory?

Philosophical accounts of confirmation answer this question by characterizing a confirmatory relationship between theory and evidence in logical or probabilistic terms. Such criteria facilitate the analysis and reconstruction of canonical confirmation cases in the history of science, and they also allow for a critical evaluation of experiments and observational studies in modern science. As we will see in Variation 9 and 11, theories of confirmation also connect to hypothesis testing in science.

The concept of confirmation also has numerous relations to other central topics of scientific reasoning. For example, Variations 6 and 7 expose substantial links between degree of confirmation, causal effect and explanatory power, following up on Carl G. Hempel's (1965) postulate of a structural identity between explanation and prediction. In Variation 8, we show how establishing intertheoretic relations between different theories, e.g., Nagelian reduction, may confirm a theory and raise our confidence in it.

Moreover, scientific reasoning can often be cast in confirmatory arguments. In Variation 4, we show how the failure to find satisfactory alterna-

tives may confirm a theory, even if there is no positive empirical evidence in its favor. Furthermore, Bayesian Confirmation Theory allows for a critical analysis of the famous No Miracles Argument (NMA). That argument claims that the astonishing success of science in recent centuries indeed confirms the hypothesis that our best scientific theories genuinely refer and constitute knowledge of the world. More on this is said in Variation 5.

The numerous references to later parts of this book make clear that confirmation is a basic concept in our work. Moreover, Bayesian Confirmation Theory is the oldest and most worked-out branch of Bayesian philosophy of science. It provides an excellent case for motivating why the Bayesian calculus can elucidate scientific reasoning, in particular, how probabilistic accounts of confirmation can address longstanding puzzles about inductive inference (e.g., the tacking paradoxes, the grue paradox, and the paradox of the ravens). Therefore, we deal with this topic in quite some detail. We first explain the benefits of expressing confirmation in Bayesian terms (Section 2.1). Then we introduce different notions of confirmation (firmness vs. increase in firmness—Section 2.2), and we examine the question whether there is a single best confirmation measure (Section 2.3). In the end, we conclude that purely theoretical and conceptual reasons are not sufficient to determine a unique measure: different measures capture different senses of confirmation, and the choice between them may also be influenced by empirical and contextual factors (Section 2.4).

Motivating Bayesian Confirmation Theory

Probability is an extremely natural model for explicating degree of confirmation. This has a number of reasons.

First, probability is, as quipped by Cicero, "the guide to life". Our decisions and actions are often based on which hypotheses are more probable than others: e.g., if there is a high chance of rain, we might cancel a planned beach trip. Confirmation is a guide to probability: better confirmed hypotheses are, ceteris paribus, more probable than others. It is therefore natural to integrate confirmation and probability within a single mathematical formalism.

Second, probability is the preferred tool for expressing uncertainty in science. Probability distributions are used for describing measurement er-

ror and for characterizing the "noise" in a system—the part of the data which cannot be explained by reference to natural laws. By phrasing confirmation in terms of probability, we connect a philosophical analysis of inductive inference to familiar scientific models where probabilistic expression of uncertainty already plays a dominant role (e.g., in linear regression models).

Third, statistics, the science of analyzing and interpreting data and assessing theories on the basis of data, is formulated in terms of probability theory. Statisticians have proved powerful mathematical results on the foundations of inductive inference, such as de Finetti's famous representation theorem for subjective probability (de Finetti, 1974) or the convergence results by Blackwell and Dubins (1962) and Gaifman and Snir (1982). Probabilistic accounts of confirmation can directly make use of these results, leading to a beneficial interaction between philosophical and statistical work (e.g., Howson and Urbach, 2006; Good, 2009). For example, the widespread practice of null hypothesis significance tests (NHST) can be fruitfully reviewed from the standpoint of a probabilistic theory of confirmation (Royall, 1997; Sprenger, 2016b).

These considerations explain why philosophy of science has paid so much attention to probabilistic confirmation theories. Among those theories, Bayesian Confirmation Theory is the most prominent representative. We shall now describe this approach.

Confirmation as (Increase in) Firmness

We remember from the introduction that Bayesians represent subjective degrees of belief by means of a probability function. The basic idea of Bayesian Confirmation Theory is that confirmation judgments are functions of an agent's conditional and unconditional degrees of belief. At first sight, this may appear unpalatably subjective. Two things should be noted, though: First, agents are assumed to be rational: their degrees of belief conform to the axioms of probability, take into account relevant evidence, etc. Second, even when the posterior degree of belief in a hypothesis differs among rational agents, they could still agree on the degree of confirmation that the evidence delivers.

We now engage in a Bayesian explication of degree of confirmation and assume that it only depends on the joint probability distribution of

the hypothesis H and the evidence E. More precisely, we assume that E and H are in the set of propositions $\mathcal{L}$ of a propositional language L that describes our domain of interest. A Bayesian confirmation measure is a function $c : \mathcal{L}^2 \times \mathfrak{P} \to \mathbb{R}$, where $\mathfrak{P}$ is the set of probability measures on the algebra generated by $\mathcal{L}$ that models the agent's degrees of belief. This function assigns a real number $c(\mathrm{H}, \mathrm{E})$ to any pair of propositions (H, E) together with a probability function $p \in \mathfrak{P}$—a number that is interpreted as the degree to which E confirms H. Reference to the probability measure p is omitted as a matter of convenience and in agreement with conventions in the literature. Similarly, we omit reference to specific background assumptions K in the confirmation relation and just assume that they are shared among all rational agents.

The classical method for explicating degree of confirmation is to specify **adequacy conditions** for the concept and to derive **representation theorems** for various confirmation measure. Such theorems characterize the set of measures (or possibly the unique measure) that satisfy these constraints. This approach allows for a sharp demarcation and mathematically rigorous characterization of the explicandum, and at the same time for critical discussion of the explicatum, by means of defending and criticizing the properties which are encapsulated in the adequacy conditions.

For example, a central function of a measure of degree of confirmation is to establish a bridge between qualitative and quantitative theories of confirmation:

Qualitative-Quantitative Bridge Principle for Confirmation For any propositions H, E $\in \mathcal{L}$ and probability measure $p \in \mathfrak{P}$ and a confirmation measure $c : \mathcal{L}^2 \times \mathfrak{P} \to \mathbb{R}$, there is a real number $t \in \mathbb{R}$ such that

- E confirms/supports H if and only if $c(\mathrm{H}, \mathrm{E}) > t$;
- E undermines/disconfirms H if and only if $c(\mathrm{H}, \mathrm{E}) < t$
- E is confirmationally neutral/irrelevant to H if and only if $c(\mathrm{H}, \mathrm{E}) = t$.

In other words, a measure of degree of confirmation should guide our qualitative confirmation in the sense that there is a numerical threshold for telling positive confirmation from disconfirmation (Carnap, 1950, 463). As a matter of convenience, we often drop quantification over the propo-

sitions H, E $\in \mathcal{L}$ and the probability measure $p \in \mathfrak{P}$, following Crupi (2013).

As already explained, Bayesian Confirmation Theory phrases degree of confirmation in terms of probabilistic dependencies between hypothesis H and evidence E. The following adequacy condition makes this approach explicit and contributes to a more precise description of the confirmation measure (e.g., Crupi, 2013):

Formality $c(\mathrm{H}, \mathrm{E})$ is a measurable function from the joint probability distribution over H and E to a real number $c(\mathrm{H}, \mathrm{E}) \in \mathbb{R}$. In particular, there exists a function $f : [0,1]^3 \to \mathbb{R}$ such that $c(\mathrm{H}, \mathrm{E}) = f(p(\mathrm{H} \wedge \mathrm{E}), p(\mathrm{H}), p(\mathrm{E}))$.

Since the three probabilities $p(\mathrm{H} \wedge \mathrm{E})$, $p(\mathrm{H})$, $p(\mathrm{E})$ suffice to determine the joint probability distribution of H and E, we can express $c(\mathrm{H}, \mathrm{E})$ as a function of these three arguments. In other words, Formality creates the common ground on which the various confirmation measures compete.

Another cornerstone for Bayesian explications of confirmation is the following principle:

Final Probability Incrementality For any propositions H, E, and E′ $\in \mathcal{L}$ with probability measure $p \in \mathfrak{P}$,

$$c(\mathrm{H}, \mathrm{E}) > c(\mathrm{H}, \mathrm{E}') \qquad \text{if and only if} \qquad p(\mathrm{H}|\mathrm{E}) > p(\mathrm{H}|\mathrm{E}'). \quad (2.1)$$

According to this principle, E confirms H more than E′ does if and only if it raises the probability of H to a higher level. It is easy to show that Final Probability Incrementality also implies that $c(\mathrm{H}, \mathrm{E}) = c(\mathrm{H}, \mathrm{E}')$ if and only if $p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H}|\mathrm{E}')$. Given the basic intuition that degree of confirmation should co-vary with boost in degree of belief, satisfactory Bayesian explications of degree of confirmation should arguably satisfy this condition.

There are now two main roads for adding more conditions, which will ultimately lead us to two different explications of confirmation: **confirmation as firmness of belief** and as **confirmation as increase in firmness** (Carnap, 1950). The latter is often called the incremental concept of confirmation or confirmation as **evidential support**.

We begin with confirmation as firmness. Consider the football standings from Table 2.1. Three teams in the Italian *Seria A*, AS Roma, FC Internazionale ("Inter"), and Juventus ("Juve") are competing for the *scudetto*,

Rank	Team	Points	Team	Points
	after 36 out of 38 rounds		after 37 out of 38 rounds	
1	Roma	78	Inter	79
2	Inter	76	Roma	78
3	Juve	74	Juve	74

Table 2.1: A motivating example for Conditional Equivalence. Top of the Seria A after 36 and 37 out of 38 rounds, respectively.

the national soccer championship. On the penultimate match day, Inter beats Juve in the *Derby d'Italia* while Roma loses to another team. Call this conjunction of propositions E. Let H = Inter will win the championship and H′ = Roma will be the runner-up. Given E, H and H′ are logically equivalent in the sense that we can derive one from the other given E. It is now very natural to claim that E confirms H and H′ to an equal degree. This intuition is expressed in the following adequacy condition:

Local Equivalence If H and H′ are logically equivalent given E (i.e., $\mathrm{E} \wedge \mathrm{H} \models \mathrm{H}'$, $\mathrm{E} \wedge \mathrm{H}' \models \mathrm{H}$), then $c(\mathrm{H}, \mathrm{E}) = c(\mathrm{H}', \mathrm{E})$.

In other words, E confirms the hypotheses H and H′ to an equal degree if they are indistinguishable conditional on E.

If we buy into this intuition, Local Equivalence allows for a powerful representation theorem by Michael Schippers (2016): all confirmation measures that satisfy Formality, Final Probability Incrementality, and Local Equivalence are non-decreasing functions of the posterior probability $p(\mathrm{H}|\mathrm{E})$.

Theorem 2.1 (Confirmation as Firmness, Schippers 2016) *Formality, Final Probability Incrementality and Local Equivalence hold if and only if there is a non-decreasing function* $g : [0,1] \to \mathbb{R}$ *such that for any* $\mathrm{H}, \mathrm{E} \in \mathcal{L}$ *and any* $p \in \mathfrak{P}$, $c(\mathrm{H}, \mathrm{E}) = g(p(\mathrm{H}|\mathrm{E}))$.

On the account of confirmation as firmness, scientific hypotheses count as well-confirmed whenever they are sufficiently probable, that is, when $p(\mathrm{H}|\mathrm{E})$ exceeds a certain (possibly context-relative) threshold. This also corresponds to Carnap's concept of probability$_1$ or "degree of confirmation" in his system of inductive logic (Carnap, 1950).

All confirmation measures that satisfy the three above conditions are **ordinally equivalent**, that is, they can be mapped onto each other by

means of a non-decreasing function. In particular, their confirmation rankings agree: if there are two functions g and g' that satisfy Theorem 2.1, with associated confirmation measures c and c', then $c(\mathrm{H},\mathrm{E}) \geq c(H',E')$ if and only if $c'(\mathrm{H},\mathrm{E}) \geq c'(H',E')$. Since confirmation as firmness is non-decreasing in $p(\mathrm{H}|\mathrm{E})$, it follows from the Qualitative-Quantitative Bridge Principle that E confirms H if and only if $p(\mathrm{H}|\mathrm{E}) \geq t$ for some—possibly context-dependent—$t \in [0,1]$.

The account of confirmation as firmness dispels some problems that have plagued their predecessors, and in particular qualitative accounts of confirmation. Among them is the idea of **hypothetico-deductive confirmation**, (where hypotheses are confirmed if they predict a phenomenon), which is explicated by a deductive entailment relation between hypothesis and evidence. This approach looks very natural: it aligns with Popper's view of scientific reasoning consisting of conjectures and refutations, and also with William Whewell's earlier view that

> our hypotheses ought to *foretel* phenomena which have not yet been observed ... the truth and accuracy of these predictions were a proof that the hypothesis was valuable and, at least to a great extent, true. (Whewell, 1847, 62–63)

However, H-D confirmation directly runs into the **paradox of tacking by conjunctions**: If E confirms H (because $\mathrm{H} \models \mathrm{E}$), then E confirms also $\mathrm{H} \wedge \mathrm{X}$, for an arbitrary hypothesis X, even if it stems from a completely different domain of science and is completely irrelevant for E. This is clearly too permissive since confirmation is allowed to spread in an uncontrolled way. The tacking paradox is therefore regarded as a major blow for the hypothetico-deductive approach to confirmation, notwithstanding recent solution attempts (Schurz, 1991; Gemes, 1993, 1998; Sprenger, 2011, 2013a).

The Bayesian account of confirmation as firmness naturally dissolves the tacking paradox. For any irrelevant X, it will be the case that $p(\mathrm{H} \wedge \mathrm{X}|\mathrm{E}) \leq p(\mathrm{H}|\mathrm{E})$. Theorem 2.1 then tells us that there exists a non-decreasing function g that maps the conditional probability of a hypothesis to its degree of confirmation. Hence, we can infer

$$c(\mathrm{H} \wedge \mathrm{X},\mathrm{E}) = g(p(\mathrm{H} \wedge \mathrm{X}|\mathrm{E})) \leq g(p(\mathrm{H}|\mathrm{E})) = c(\mathrm{H},\mathrm{E}), \tag{2.2}$$

demonstrating that the conjunction is confirmed to a lower degree than the original hypothesis H (especially so for an unlikely, far-fetched proposition

X). Confirmation as firmness gives the intuitively correct response to the tacking by conjunction paradox. It does not deny that H ∧ X is confirmed as well—after all, H is still obviously relevant for E—but the paradox is mitigated by decreasing the amount of confirmation.

On the other hand, confirmation as firmness does not always agree with the use of that concept in scientific reasoning. To be sure, relative to the totality of observed evidence, we would call a theory well-confirmed if and only if it is sufficiently probable, conditional on the evidence. But often, scientists are interested in whether a certain experiment supports or corroborates a hypothesis—independent of whether we find it probable that the hypothesis is true. It is essential for confirmatory evidence to provide a good reason for believing a theory, even if the theory is, all things considered, unlikely.

For instance, in the first years after Einstein invented the General Theory of Relativity (GTR), many scientists did not have a particularly high degree of belief in GTR because of its counterintuitive nature. However, it was agreed upon that GTR was well-confirmed by its predictive and explanatory successes, such as the bending of starlight by the sun and the explanation of the Mercury perihelion shift (Earman, 1992). The account of confirmation as firmness fails to capture this intuition. The same holds true for statistical analysis of experiments in modern science, where the dominant frequentist paradigm does not allow for assigning a probability to the tested hypothesis. Instead, the confirmatory strength of the evidence is evaluated on the basis of whether the results are statistically significant and give reason to reject the tested hypothesis in favor of an alternative. Moreover, on confirmation as firmness, E could confirm H even if it *lowers* the probability of H, as long as $p(\mathrm{H}|\mathrm{E})$ is still large enough. But few people would call an experiment where the results undermine H a confirmation of H.

In a now classical debate in philosophy of science, Karl R. Popper (1954, 2002) raised these points against Carnap: degree of confirmation cannot be (posterior) probability. As a reaction, Carnap distinguished two concepts of confirmation in the second edition (1962) of "Logical Foundations of Probability": confirmation as firmness and **confirmation as increase in firmness**. This brings us to the following natural definition that provides a more precise condition for the relation between qualitative confirmation judgments and probabilistic relevance:

Confirmation as Increase in Firmness For any propositions H, E $\in \mathcal{L}$ with probability measure $p \in \mathfrak{P}$,

1. Evidence E **confirms/supports** hypothesis H if and only if $p(\mathrm{H}|\mathrm{E}) > p(\mathrm{H})$.
2. Evidence E **disconfirms/undermines** hypothesis H if and only if $p(\mathrm{H}|\mathrm{E}) < p(\mathrm{H})$.
3. Evidence E is **neutral** with respect to H if and only if $p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H})$.

In other words, E confirms H if and only if E raises our degree of belief in H. Such explications of confirmation are also called **statistical relevance** accounts of confirmation because the neutral point is determined by the statistical independence of H and E. They measure the **evidential support** that H receives from E. The increase in firmness explication of confirmation receives empirical support from findings by Tentori et al. (2007a): ordinary people use the concept of confirmation in a way which can be dissociated from posterior probability and that is strongly correlated with measures of evidential support. In the remaining variations, we will standardly use increase in firmness, or evidential support, when modeling the confirmation of scientific hypotheses and theories.

Confirmation as increase in firmness has interesting relations to qualitative accounts of confirmation and the paradoxes we have encountered. For instance, hypothetico-deductive confirmation emerges as a special case: if H entails E and $p(\mathrm{E}) < 1$, then $p(\mathrm{E}|\mathrm{H}) = 1$ and by Bayes' Theorem, $p(\mathrm{H}|\mathrm{E}) > p(\mathrm{H})$. We will also show that confirmation as increase in firmness can address the tacking by conjunction paradox. But first, we will demonstrate how confirmation as increase in firmness handles the longstanding **paradox of the ravens**.

Let $\mathrm{H} = \forall x : Rx \rightarrow Bx$ stand for the hypothesis that all ravens are black. H is equivalent to the hypothesis $\mathrm{H}' = \forall x : \neg Bx \rightarrow \neg Rx$ that no non-black object is a raven. It is highly intuitive that logically equivalent hypotheses are confirmed or disconfirmed to the same degree; nothing in a formal theory of confirmation should depend on the particular formulation of the hypothesis. Hence, anything that confirms H also confirms H' and vice versa (Nicod, 1961). It is also intuitive that universal conditionals such as "all ravens are black" are confirmed by their instances, i.e., black ravens. However, as Hempel (1945a, 1945b) observed, the conjunction of

	W_1	W_2
Black ravens	100	1,000
Non-black ravens	0	1
Other birds	1,000,000	1,000,000

Table 2.2: I.J. Good's (1967) counterexample to the paradox of the ravens.

both principles leads to paradoxical results. A black raven is an instance of H and confirms the raven hypothesis. A white shoe is an instance of H' and confirms the hypothesis that non-black objects are not ravens. But because of the aforementioned equivalence condition, the white shoe also confirms the hypothesis that all ravens are black! This result is known as the paradox of the ravens, or alternatively, as Hempel's paradox.

The account of confirmation as firmness allows us to spot what is wrong with instance confirmation and thereby resolves the paradox. While that intuition is certainly valid for *some* background assumptions, it is not valid for all possible situations. I.J. Good (1967) constructed a simple counterexample: Assume that there are only two possible worlds, W_1 and W_2, whose properties are described by Table 2.2.

In this scenario, the raven hypothesis H is true whenever W_1 is the case, and false whenever W_2 is the case. Moreover, the observation of a black raven is evidence that W_2 is the case and therefore evidence that not all ravens are black:

$$p(Ra.Ba|W_1) = \frac{100}{1,000,100} \quad < \quad \frac{1,000}{1,001,001} = p(Ra.Ba|W_2).$$

By an application of Bayes' Theorem, we infer $p(W_1|Ra.Ba) < p(W_1)$ and $p(\mathrm{H}|Ra.Ba) < p(\mathrm{H})$. This shows that universal conditionals are not always confirmed by their instances. We see how the explication of confirmation as increase in firmness corrects our pre-theoretic intuitions regarding the theory-evidence relation. The raven paradox may thus be resolved by rejecting one of its assumptions, namely the idea that instances always confirm a hypothesis.

The raven paradox is threefold, however: apart from the (qualitative) question whether or not the observation of a white shoe confirms the raven hypothesis, there is also the **comparative paradox**: does the observation of a black raven confirm the raven hypothesis to a higher degree than the observation of a white shoe? Fitelson and Hawthorne (2011) show in their Theorem 2 that this is indeed the case if plausible assumptions on the

real world are made: $p(\mathrm{H}|Ra.Ba) < p(\mathrm{H}|\neg Ra.\neg Ba)$. By Final Probability Incrementality, this implies that $Ra.Ba$ confirms H more than $\neg Ra.\neg Ba$ does. This shows, ultimately, why we consider a black raven to be more important evidence for the raven hypothesis than a white shoe. In the light of these results, the paradox loses its bite.

Confirmation as increase in firmness also addresses another notorious paradox, Nelson Goodman's (1955) **new riddle of induction**. The name notwithstanding, it is not meant as a general charge on inductive reasoning, but on a particularly plausible view of inductive inference: namely that one and the same evidence cannot confirm two hypotheses whose predictions contradict each other.

As Goodman shows, this principle disagrees with what most people would classify as a justified inductive inference. Consider, for example, the following case:

> Observation: emerald e_1 is green.
> Observation: emerald e_2 is green.
> ...
>
> ---
>
> Generalization: All emeralds are green.

This seems to be a perfect example of a valid inductive inference. Now define the predicate "grue", which applies to all green objects if they were observed for the first time prior to time $t =$ "now", and to all blue objects if they are observed later. (This is just a description of the extension of the predicate—no object is supposed to change color.) The following inductive inference satisfies the same logical scheme as the previous one:

> Observation: emerald e_1 is grue.
> Observation: emerald e_2 is grue.
> ...
>
> ---
>
> Generalization: All emeralds are grue.

In spite of the gerrymandered nature of the "grue" predicate, the inference is sound: it satisfies the basic scheme of enumerative induction, and the premises are undoubtedly true. But then, it is paradoxical that two valid inductive inferences support flatly opposite conclusions. The first generalization predicts that emeralds observed in the future are green, the

second generalization predicts them to be blue. How do we escape from this dilemma?

One may propose that in virtue of its gerrymandered nature, the predicate "grue" should not enter inductive inferences. Goodman notes, however, that it is perfectly possible to re-define the standard predicates "green" and "blue" in terms of "grue" and its conjugate predicate "bleen" (=blue if observed prior to t, else green). Hence, any preference for the "natural" predicates and the "natural" inductive inference seems to be arbitrary, or at least conditional on the choice of a specific language. So another move is required.

The explication of confirmation as increase in firmness immediately comes up with an answer: both hypotheses (the "green" and the "grue" hypothesis) should count as confirmed. We need to abandon the idea that evidence cannot confirm incompatible hypotheses. If they share content that is confirmed by the current experiment, they are both supported by it. For example, Einstein's work on the photoelectric effect raised our degree of belief in the hypothesis that electromagnetic radiation can be divided into a finite number of quanta, and thereby also in different versions of quantum theory—e.g., those that were compatible with relativity theory and those that weren't.

How do confirmation judgments inform our predictions for future observations? Generalizing Goodman's green/grue argument, it seems that *any* prediction for the color of the next observed emerald seems equally reasonable. However, from a Bayesian point of view, this is only true if all hypotheses (that is, the green, grue, gred, etc. hypotheses) are equally probable at the time of observation. In practice, this will usually not be the case: some hypotheses have higher plausibility than others. Prior probabilities act as a counterweight to Goodman's paradox and guide our predictions when the observations cannot distinguish between two hypotheses. And of course, the grue hypothesis is much more implausible than the green hypothesis. Note that this choice cannot be based on Bayesian reasoning: they have to come from theoretical principles, past track record, coherence with other parts of science, and the vague reasoning faculty that we call scientific judgment. Bayesian confirmation theory explains how to amalgamate prior degree of belief with observed evidence, but it does not tell you which prior degrees of belief are reasonable. In this sense, Goodman shows a general problem for formal reasoning about confirmation

and evidence: there is no viable premise-independent theory of inductive support (see also Norton, 2016).

The three showcases above—the tacking by conjunction paradox, the paradox of the ravens and Goodman's new riddle of induction—make clear that Bayesian Confirmation Theory can successfully address long-standing puzzles in inductive reasoning. However, there is one question we have evaded so far, and now we shall turn to it: how can we measure confirmation as increase in firmness, or alternatively, how should evidential support be quantified?

The Plurality of Bayesian Confirmation Measures

For scientists who want to report the results of their experiments, quantifying the strength of the observed evidence is an urgent and challenging question. It is also crucial for giving a Bayesian answer to the Duhem-Quine problem (Duhem, 1914). If an experiment fails and we ask ourselves which hypothesis to reject, the degree of (dis)confirmation of the involved hypotheses can be used to evaluate their standing. Unlike purely qualitative accounts of confirmation, a measure of evidential support can indicate which hypothesis we should discard. For this reason, the search for a proper confirmation measure is more than a technical exercise: it is of a vital importance for distributing praise and blame between different hypotheses that bear on an observation. Such assessments may also be sensitive to the used measure, highlighting the need for characterizing their mathematical properties and comparing them on a normative basis (Eells and Fitelson, 2000, 2002; Fitelson, 1999, 2001a,b).

Table 2.3 gives a survey of measures that are frequently discussed in the literature. We have normalized them such that for each measure $c(\mathrm{H},\mathrm{E})$, confirmation amounts to $c(\mathrm{H},\mathrm{E}) > 0$, neutrality to $c(\mathrm{H},\mathrm{E}) = 0$ and disconfirmation to $c(\mathrm{H},\mathrm{E}) < 0$. This allows for a better comparison of the measures and their properties.

Evidently, these measures all have quite distinct logical and epistemological properties. It makes thus sense to apply the methodology that we used for confirmation as firmness, and to characterize them in terms of representation theorems where, as before, Formality and Final Probability Incrementality will serve as minimal reasonable constraints on any measure of evidential support. Notably, Final Probability Incremental-

Difference Measure	$d(\mathrm{H},\mathrm{E}) = p(\mathrm{H}\vert\mathrm{E}) - p(\mathrm{H})$
Log-Ratio Measure	$r(\mathrm{H},\mathrm{E}) = \log \frac{p(\mathrm{H}\vert\mathrm{E})}{p(\mathrm{H})}$
Log-Likelihood Measure	$l(\mathrm{H},\mathrm{E}) = \log \frac{p(\mathrm{E}\vert\mathrm{H})}{p(\mathrm{E}\vert\neg\mathrm{H})}$
Kemeny-Oppenheim Measure	$k(\mathrm{H},\mathrm{E}) = \frac{p(\mathrm{E}\vert\mathrm{H})-p(\mathrm{E}\vert\neg\mathrm{H})}{p(\mathrm{E}\vert\mathrm{H})+p(\mathrm{E}\vert\neg\mathrm{H})}$
Generalized Entailment Measure	$z(\mathrm{H},\mathrm{E}) = \begin{cases} \frac{p(\mathrm{H}\vert\mathrm{E})-p(\mathrm{H})}{1-p(\mathrm{H})} & \text{if } p(\mathrm{H}\vert\mathrm{E}) \geq p(\mathrm{H}) \\ \frac{p(\mathrm{H}\vert\mathrm{E})-p(\mathrm{H})}{p(\mathrm{H})} & \text{if } p(\mathrm{H}\vert\mathrm{E}) < p(\mathrm{H}) \end{cases}$
Christensen-Joyce Measure	$s(\mathrm{H},\mathrm{E}) = p(\mathrm{H}\vert\mathrm{E}) - p(\mathrm{H}\vert\neg\mathrm{E})$
Carnap's Relevance Measure	$c'(\mathrm{H},\mathrm{E}) = p(\mathrm{H}\wedge\mathrm{E}) - p(\mathrm{H})p(\mathrm{E})$
Rips Measure	$r'(\mathrm{H},\mathrm{E}) = 1 - \frac{p(\neg\mathrm{H}\vert\mathrm{E})}{p(\neg\mathrm{H})}$

Table 2.3: A list of popular measures of evidential support.

ity already rules out two of the measures in the list, namely Carnap's relevance measure $c'(\mathrm{H},\mathrm{E}) = p(\mathrm{H}\wedge\mathrm{E}) - p(\mathrm{H})p(\mathrm{E})$ and the Christensen-Joyce measure $s(\mathrm{H},\mathrm{E}) = p(\mathrm{H}\vert\mathrm{E}) - p(\mathrm{H}\vert\neg\mathrm{E})$ Christensen (1999). Carnap's relevance measure is also problematic because it relies on the symmetry $c(\mathrm{H},\mathrm{E}) = c(\mathrm{E},\mathrm{H})$, in other words, E confirms H as much as H confirms E. Many intuitive confirmation judgments violate this equality. For example, knowing that a specific card, in the deck, is the ace of spades confirms the hypothesis that this card is a spade much stronger than the other way round. The same problem affects the (log-)ratio measure $r(\mathrm{H},\mathrm{E})$ (Eells and Fitelson, 2002).

We will not discuss all representation results, but we present, *pars pro toto*, four specific conditions that figure in different representation theorems and that resurface at later points of the book, too. The first condition is the

Law of Likelihood

$$c(\mathrm{H},\mathrm{E}) > c(\mathrm{H}',\mathrm{E}) \qquad \text{if and only if} \qquad p(\mathrm{E}\vert\mathrm{H}) > p(\mathrm{E}\vert\mathrm{H}').$$

This condition has a long history of discussion in philosophy and statistics. The idea is that E favors H over H′ if and only if the likelihood of H on E is greater than the likelihood of H′ on E (Hacking, 1965; Edwards, 1972; Royall, 1997; Sober, 2008). In other words, E is more expected under H than under H′. Law of Likelihood is also at the basis of the **likelihoodist theory**

of confirmation, which regards confirmation as a comparative relation between two competing hypotheses and refuses to make a straightforward judgment on how much E confirms H.

The second condition demands that conditioning on E′ does not affect the confirmation relation between H and E′, as long as E′ is sufficiently independent from E and H:

Modularity If $E \perp\!\!\!\perp E' \,|\, H$ (that is, $p(E|\pm H \wedge E') = p(E|\pm H)$), then $c(H, E) = c_{|E'}(H, E)$ where $c_{|E'}$ denotes confirmation relative to the probability distribution conditional on E′.

That is, if E′ does not affect the likelihoods that H and ¬H have on E, then conditioning on E—now supposedly irrelevant evidence—does not alter the degree of confirmation (Heckerman, 1988; Crupi et al., 2013). The intuition behind Modularity is that probabilistically irrelevant information should not alter a judgment of degree of confirmation.

A third condition concerns the question of how the confirmation of hypothesis H by evidence E relates to the confirmation of the disjunction $H \vee H'$ by the same evidence. The idea is that the logical weakening of a hypothesis contributes to the confirmation of the compound if and only if the added disjunct is confirmed by the evidence.

Disjunction of Alternative Hypotheses Assume that H and H′ are inconsistent with each other. Then, $c(H, E) > c(H \vee H', E)$ if and only if E confirms H′ as well (that is, $p(H'|E) > p(H')$). Analogous conditions hold for $c(H, E) = c(H \vee H', E)$ and $c(H, E) < c(H \vee H', E)$.

Finally, the fourth condition is inspired by the analogy between deductive and inductive logic: confirmation is viewed as a generalization of logical entailment to uncertain reasoning (Crupi et al., 2007; Crupi and Tentori, 2013). Degree of confirmation should therefore display the symmetry that contraposition expresses for logical entailment: if E confirms H, then ¬H also confirms ¬E, and the two degrees of confirmation are the same. Similarly, disconfirmation is, like logical inconsistency, modeled as a symmetrical relation

Contraposition/Commutativity If E confirms H, then $c(H, E) = c(\neg E, \neg H)$; and if E disconfirms H, then $c(H, E) = c(E, H)$.

Combined with Formality and Final Probability Incrementality, each of these four principles gives rise to a representation theorem that singles out

a particular measure (for the theorems and proofs, see Heckerman, 1988; Crupi et al., 2013; Crupi, 2013):

Theorem 2.2 (Representation Theorems for Confirmation Measures) *1. If Formality, Final Probability Incrementality and Law of Likelihood hold, then there is a non-decreasing function g such that $c(\mathrm{H},\mathrm{E}) = g(r(\mathrm{H},\mathrm{E}))$.*

2. *If Formality, Final Probability Incrementality and Modularity hold, then there are non-decreasing functions g and g' such that $c(\mathrm{H},\mathrm{E}) = g(l(\mathrm{H},\mathrm{E}))$ and $c(\mathrm{H},\mathrm{E}) = g'(k(\mathrm{H},\mathrm{E}))$. Note that k and l are ordinally equivalent.*

3. *If Formality, Final Probability Incrementality and Disjunction of Alternative Hypotheses hold, then there is a non-decreasing function g such that $c(\mathrm{H},\mathrm{E}) = g(d(\mathrm{H},\mathrm{E}))$.*

4. *If Formality, Final Probability Incrementality and Commutativity hold, then there is a non-decreasing function g such that $c(\mathrm{H},\mathrm{E}) = g(z(\mathrm{H},\mathrm{E}))$.*

It should also be noted that the **Bayes factor**, a popular measure of evidential support in Bayesian statistics (Kass and Raftery, 1995; Goodman, 1999b), falls under the scope of the theorem. For mutually exclusive hypotheses H_0 and H_1 and evidence E, the Bayes factor in favor of H_0 is defined as

$$B_{01}(\mathrm{E}) := \frac{p(\mathrm{H}_0|\mathrm{E})}{p(\mathrm{H}_1|\mathrm{E})} \cdot \frac{p(\mathrm{H}_1)}{p(\mathrm{H}_0)} = \frac{p(\mathrm{E}|\mathrm{H}_0)}{p(\mathrm{E}|\mathrm{H}_1)}$$

It is not difficult to see that this quantity is ordinally equivalent to the log-likelihood measure l and the Kemeny-Oppenheim measure k (Kemeny and Oppenheim, 1952) when H_0 and H_1 exhaust the space of hypotheses: just substitute H and $\neg\mathrm{H}$ for H_0 and H_1.

To underine that the difference between the various confirmation measures has substantial philosophical ramifications, let us go back to the problem of irrelevant conjunctions. If we analyze this problem in terms of the ratio measure r, then we obtain, assuming $\mathrm{H} \vdash \mathrm{E}$, that for an "irrelevant" conjunct H′,

$$\begin{aligned} r(\mathrm{H}\wedge\mathrm{H}',\mathrm{E}) &= p(\mathrm{H}\wedge\mathrm{H}'|\mathrm{E})/p(\mathrm{H}\wedge\mathrm{H}') = p(\mathrm{E}|\mathrm{H}\wedge\mathrm{H}')/p(\mathrm{E}) \\ &= 1/p(\mathrm{E}) = p(\mathrm{E}|\mathrm{H})/p(\mathrm{E}) \\ &= r(\mathrm{H},\mathrm{E}) \end{aligned}$$

such that the irrelevant conjunction is supported to the same degree as the original hypothesis. This consequence is certainly unacceptable as a

judgment of evidential support since H′ could literally be any hypothesis unrelated to the evidence, e.g., "the star Sirius is a giant light bulb". In addition, the result does not only hold for the special case of deductive entailment: it holds whenever the likelihoods of H and H ∧ H′ on E are the same, that is, $p(\mathrm{E}|\mathrm{H} \wedge \mathrm{H}') = p(\mathrm{E}|\mathrm{H})$.

The other measures fare better in this respect: whenever $p(\mathrm{E}|\mathrm{H} \wedge \mathrm{H}') = p(\mathrm{E}|\mathrm{H})$, all other measures in Theorem 2.2 reach the conclusion that $c(\mathrm{H} \wedge \mathrm{H}', \mathrm{E}) < c(\mathrm{H}, \mathrm{E})$ (Hawthorne and Fitelson, 2004). In this way, we can see how Bayesian Confirmation Theory improves on H-D confirmation and other qualitative accounts of confirmation: the paradox is acknowledged, but at the same time, it is demonstrated how the paradox can be mitigated.

Discussion

Which of the remaining confirmation measures should be preferred? This is a difficult question that probably cannot be resolved on purely theoretical grounds. The adequacy conditions in the representation theorems have quite divergent motivations, and a straightforward comparison is unlikely to lead to conclusive results. For example, it has been shown that no confirmation measure satisfies the following two conditions: (i) degree of confirmation is maximal if E implies H; (ii) the a priori informativity (cashed out in terms of predictive content and improbability) of a hypothesis contributes to degree of confirmation (Brössel, 2013, 389–390). See also Huber (2005). Both conditions are intuitively plausible: (i) captures the idea that degree of confirmation generalizes logical entailment, (ii) the idea that hypotheses with informative predictions should be rewarded. But we have to choose, and our choice will depend on what we value in scientific reasoning.

The idea that there is "the one true measure of confirmation" (Milne, 1996) is therefore problematic. We may abandon such a confirmational monism in favor of pluralism (Fitelson, 1999, 2001b): we accept that there are different senses of degree of confirmation that correspond to different explications. For example, d strikes us as a natural explication of increase in subjective confidence, z generalizes deductive entailment, and l and k measure the discriminatory force of the evidence regarding H and ¬H.

Ultimately, the choice between the measures may also depend on em-

pirical findings. Crupi et al. (2007) and Tentori et al. (2007a) compare different confirmation measures in an experiment where white and black balls are drawn from an urn and the participants must assess the confirmation of different hypotheses about the composition of balls in the urn. Their results favor the z-measure, followed by the l-measure, whereas the difference measure d is at the bottom of the list. In a similar vein, a recent experiment by Colombo et al. (2016a) has pointed out that judgments of confirmation are enhanced by the prior plausibility of a hypothesis if the probabilistic relevance relations are held constant. This phenomenon, consistent with the findings of Crupi et al. (2007), is also called the *Matthew effect*: "For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath." (Matthew 25:29). For Bayesian confirmation measures, this means that measures which do not assign a *ceteris paribus* bonus to logically stronger and more informative hypotheses are probably more in line with our empirical confirmation judgments (Festa, 2012; Roche, 2014). If there is hope for confirmational monism, it might come from empirical research on confirmation judgments, showing that participants share the motivation behind a specific measure.

We have seen that Bayesian Confirmation Theory yields many interesting results in philosophy of science. But it is also a research paradigm that connects well to scientific disciplines: Bayesian reasoning has sparked interest among experimental psychologists and connects well to cutting-edge research on human cognition (e.g., Oaksford and Chater, 2000; Doya et al., 2007; Douven, 2016). There is a large number of interdisciplinary papers on probabilistic reasoning, where both cognitive scientists and philosophers have been involved (e.g., Tentori et al., 2007b; Crupi et al., 2008; Zhao et al., 2012). But also on the theoretical side, there is ample room for future research. Questions that are just about to be explored include an analysis of confirmation measures in information-theoretic terms (Crupi and Tentori, 2014) and the use of confirmation measures for analyzing the diagnostic value of a medical test (Crupi et al., 2009). Especially the latter question, which deals with designing medical tests that lead to a high amount of (dis)confirmation upon revealing the results, strikes us as an exciting combination of Bayesian philosophy of science with clinical practice.

In spite of all these cross-disciplinary connections, one criticism of

Bayesian Confirmation Theory has been levelled again and again: that it misrepresents actual scientific reasoning. As evidence for this claim, the Problem of Old Evidence is often cited (Glymour, 1980; Brössel and Huber, 2015). The next variation responds to these worries.

Variation 3: The Problem of Old Evidence

The Problem of Old Evidence—modeling how already known evidence confirms a scientific theory—is one of the most troubling and persistent challenges for Bayesian Confirmation Theory. The most famous case of confirmation by old evidence might be the Mercury perihelion shift (Glymour, 1980; Earman, 1992). For a long time, this phenomenon could not be fully explained by Newtonian mechanics or any other reputable physical theory. Then, Einstein realized that his General Theory of Relativity (GTR) accounted for the perihelion shift. According to most physicists, this discovery conferred a substantial degree of confirmation on GTR, perhaps even more than some pieces of novel evidence, such as Eddington's 1919 solar eclipse observations. Also in other scientific disciplines, newly introduced theories are commonly assessed with respect to their success at explaining away observational anomalies. Think, for example, of the assessment of global climate models against a track record of historical data, or of economic theories that try to explain anomalies in decision-making under uncertainty, such as the Allais or Ellsberg paradoxes.

We can extract a general scheme from these examples. A phenomenon E is unexplained by the available scientific theories. At some point, it is discovered that theory T accounts for E. E is old evidence: at the time when this relationship is developed, the scientist is already certain or close to certain that the phenomenon E is real. Indeed, in the GTR example, astronomers were collecting data on the Mercury perihelion shift for many decades. Nevertheless, E apparently confirms T because it resolves a well-known and persistent observational anomaly. How does this fit into the Bayesian view of confirmation?

The relevant sense of confirmation in this example is not firmness, but increase in firmness: confirming evidence raises a rational agent's confi-

dence in the theory. E confirms T if and only if $p(\mathrm{T}|\mathrm{E}) > p(\mathrm{T})$. These two probabilities are related by means of Bayes' Theorem:

$$p(\mathrm{T}|\mathrm{E}) = p(\mathrm{T})\frac{p(\mathrm{E}|\mathrm{T})}{p(\mathrm{E})}$$

When E is an old evidence and already known to the scientist, her degree of belief in E is maximal: $p(\mathrm{E}) = 1$. With that assumption, it follows that the probability of T conditional on E cannot be greater than the unconditional probability:

$$p(\mathrm{T}|\mathrm{E}) = p(\mathrm{T}) \cdot p(\mathrm{E}|\mathrm{T})/p(\mathrm{E}) = p(\mathrm{T}) \cdot p(\mathrm{E}|\mathrm{T}) \leq p(\mathrm{T}) \tag{3.1}$$

Hence, E does not confirm T in the sense of increasing the firmness of our belief in T. The very idea of confirmation by old evidence, or equivalently, **confirmation by accounting for well-known observational anomalies**, seems impossible to describe the Bayesian belief kinematics. This is the Problem of Old Evidence (POE). The problem may also be phrased differently, as exposing that the Bayesian cannot account how the **discovery of explanatory relations between theory and evidence** increases the epistemic standing of the theory. Notably, the problem does not allow for an easy fix by making assumptions on $p(\mathrm{T})$ and the likelihoods $p(\mathrm{E}|\pm\mathrm{T})$: as long as $0 < p(\mathrm{T}) < 1$, the law of total probability implies that $p(\mathrm{E}|\mathrm{T}) = p(\mathrm{E}|\neg\mathrm{T}) = 1$ if $p(\mathrm{E}) = 1$. Hence, T fails to be confirmed by E.

The POE has different aspects, as worked out by Ellery Eells (1985, 1990). First, there is the *static* (Eells: "ahistorical") POE: belief changes induced by discovering T or an explanatory relationship between T and E have already taken place. Still we would like to say that E is evidentially relevant for T: when faced with a decision between T and a competitor T', E is a good reason for preferring T over T'. But there is also the *dynamic* (Eells: "historical") POE: it refers to the moment in time where T and its relation to E are discovered. Why does the discovery that T accounts for E raise our confidence in T? How can the discovery of an explanatory success be confirmationally relevant? (see also Romeijn and Wenmackers, 2016). In other words, the dynamic POE deals with the question of confirmation relative to actual degrees of belief, which is not necessarily the case for the static POE.

This variation develops a new solution proposals for both the static and the dynamic POE. Section 3.1 comments on solutions of the dynamic POE

proposed by Garber (1983), Jeffrey (1983), Niiniluoto (1983) and Earman (1992). On these accounts, confirmation occurs through conditionalizing on the proposition that T implies E. Section 3.2 presents an improvement on this approach. Section 3.3 analyses the static POE while Section 3.4 explains our solution of that problem. We conclude with a brief discussion in Section 3.5 and by giving the proofs of our results in Section 3.6.

The Dynamic Problem of Old Evidence: The GJN Approach

The dynamic Problem of Old Evidence is concerned with how learning a deductive or an explanatory consequence of a theory can raise our confidence in that theory. An example from history may help to get this clear. In classical examples, such as explaining the Mercury perihelion shift, the newly invented theory (here: GTR) was initially not known to entail the old evidence. It took Einstein some time to find out that T entailed E (Brush, 1989; Earman, 1992). By learning the relationship $X = T \vdash E$, Einstein increased his confidence in T since such a strong consilience of theory and data could not be expected beforehand. Thus, the inequality

$$p(T|X, E) > p(T|E) \tag{3.2}$$

seems to be a plausible representation of Einstein's degrees of belief before and after making the discovery that GTR explained the perihelion shift of Mercury. Consequently, the relevant piece of evidence is not E itself, but the learning of a specific relation between theory and evidence, namely that T implies, accounts for or explains E.

However, such belief change is hard to model in a Bayesian framework. A Bayesian reasoner is assumed to be logically omniscient and the logical fact $X = T \vdash E$ should always have been known. Hence, X cannot be properly *learned* by a Bayesian: it is, and has always been, part of her background beliefs.

To solve this problem, several philosophers have relaxed the assumption of logical omniscience and enriched the set of propositions about which agents have degrees of belief. New atomic sentences of the form $X = T \vdash E$ are added (Garber, 1983; Jeffrey, 1983; Niiniluoto, 1983), such that Bayesian Confirmation Theory can account for our cognitive limitations in deductive reasoning. Then, it can be shown that under suitable assumptions, conditionalizing on X confirms T.

The first models along these lines have been developed by Daniel Garber, Richard Jeffrey and Ilkka Niiniluoto in a group of papers which all appeared in 1983. Henceforth, we will refer to the family of their solution proposals as the GJN solutions. In order to properly compare our own solution proposals to the state-of-the-art, and to assess their innovative value, we will briefly recap the achievements of the GJN models and elaborate their limitations and weaknesses.

All GJN models aim to show that conditionalizing on the proposition X increases the posterior probability of T. Eells (1990, 211) distinguishes three steps in this endeavor: First, parting with the logical omniscience assumption and developing a formal framework for imperfect Bayesian reasoning. Second, describing which kind of relation obtains between T and E. Third, showing that learning this relation increases the probability of T. While the GJN models neglect the second step, probably in due anticipation of the diversity of logical and explanatory relations in science, they are quite explicit on Step 1 and Step 3.

Garber's model focuses on Step 1 and on learning logical truths and explanatory relations in a Bayesian framework (Garber, 1983). For instance, learning logical/mathematical truths can be quite insightful and lead to great progress in science. The famous, incredibly complex proof of Fermat's Last Theorem may be a good example . Garber therefore enriches the underlying language L in a way that the proposition of the meta-language X is one of the *atomic* propositions of the extended language L'.

Garber also demands that the agent recognize some elementary relations in which X stands to other elements of L':

$$p(\mathrm{E}|\mathrm{T},\mathrm{X}) = 1 \qquad\qquad p(\mathrm{T},\mathrm{E},\mathrm{X}) = p(\mathrm{T},\mathrm{X}). \tag{3.3}$$

These constraints are an equivalent of modus ponens for a logic of degree of belief: conditional on T and X, the agent should be certain that E. In other words, if an agent takes T and X for granted, then she also believes E to maximal degree. Knowledge of such elementary inference schemes sounds eminently sensible when we are trying to model the boundedly rational reasoning of a scientist. Garber then proves the following theorem: there is at least one probability function on L' such that *every* non-trivial atomic sentence of the form X gets a value strictly between 0 and 1. Thus, one can coherently have a genuinely uncertain attitude about all propositions in the logical universe, including tautologies. Finally Garber shows

that there are infinitely many probability functions such that $p(\mathrm{E}) = 1$ and $p(\mathrm{T}|\mathrm{X},\mathrm{E}) > p(\mathrm{T}|\mathrm{E})$. A similar point is, though with less formal detail and rigor, made by Niiniluoto (1983).

While Garber's efforts are admirable, they only address the first step of solving the dynamic POE: he provides an existence proof for a solution to the POE, but he does not show that learning X confirms T for most plausible probability distributions over E, T and X. Also Niiniluoto (1983) only sketches a solution idea without filling in the details. This lacuna is closed by Richard Jeffrey (1983), who published his solution in the same volume where Garber's paper appeared.

Jeffrey considers the meta-proposition X as an object of subjective uncertainty, but he keeps formalities down to the standard level of Bayesian Confirmation Theory. Then he makes the following assumptions, using the notational convention $\mathrm{X}' := \mathrm{T} \vdash \neg\mathrm{E}$:

(α) $p(\mathrm{E}) = 1$.

(β) $p(\mathrm{T})$, $p(\mathrm{X})$, $p(\mathrm{X}') \in (0,1)$.

(γ) $p(\mathrm{X},\mathrm{X}') = 0$.

(δ) $p(\mathrm{T}|\mathrm{X} \vee \mathrm{X}') \geq p(\mathrm{T})$.

(η) $p(\mathrm{T},\neg\mathrm{E},\mathrm{X}') = p(\mathrm{T},\mathrm{X}')$.

From these assumptions, Jeffrey derives $p(\mathrm{T}|\mathrm{X},\mathrm{E}) > p(\mathrm{T},\mathrm{E})$, that is, the solution to the dynamic POE.

The strength of Jeffrey's solution crucially depends on how well we can motivate condition (δ). The other conditions are plausible: (α) is just the standard presumption that at the time where confirmation takes place, E is already known to the agent. (β) demands that we may not be certain about the truth of T or $\mathrm{T} \vdash \pm\mathrm{E}$ beforehand, in line with the typical description of the POE. (γ) requires that T do not entail E and $\neg$E at the same time. Finally, (η) is a Modus Ponens condition similar to (3.3): the joint degree of belief in T, $\neg$E and X′ is equal to the joint degree of belief in T and X′, demanding that the agent recognize that the latter two propositions entail $\neg$E.

Hence, (δ) really carries the burden of Jeffrey's argument. This condition has some odd technical consequences, as pointed out by Earman (1992, 127). For instance, with plausible additional assumptions, we can

derive $p(\mathrm{T}|\mathrm{X}) \geq 2p(\mathrm{T})$ which implies that the prior degree of belief $p(\mathrm{T})$ must have been smaller than .5. Jeffrey's solution of the dynamic POE does not apply to theories that were already quite probable, and this is an awkward feature.

That said, the real problem with (δ) is not technical, but philosophical. Jeffrey (1983, 148–149) supports (δ) by mentioning that Newton was, upon formulating his theory of gravitation G, convinced that it would bear on the phenomena he was interested in, namely the mechanism governing the tides. Although Newton did not know whether G would entail the phenomena associated to the tides or be inconsistent with them, he used his knowledge that G would bear on the tides as a reason for accepting it as a working hypothesis.

To our mind, this reconstruction conflates an *evidential* virtue of a theory with a *methodological* one. Theories of which we know that they make precise predictions on an interesting subject matter are worthy of further elaboration and pursuit, even if the content of their predictions is not yet known. This is basically a Popperian rationale for scientific inquiry: go for theories that have high empirical content, that make precise predictions, and develop them further. They are the ones that will finally help us to solve urgent scientific problems. Newton may have followed this methodological rule when discovering that his theory of gravitation would have some implications for the tides phenomena. Making such pragmatic acceptances, however, does not entail a commitment to the thesis that the plausiblity of a theory increases with its empirical content. Actually, Popper (2002, 268–269) thought the other way round: theories with high empirical content rule out more states of the world and will have low probability! This is just because they take, in the virtue of making many predictions, a higher risk of being falsified. Indeed, it is hard to understand why increasing the empirical content of T provides an argument that T is more likely to be true. Increasing the class of potential falsifiers of T should not increase its plausibility. Jeffrey's condition (δ) is therefore ill-grounded and at the very least too controversial to act as a premise in a solution of the POE.

Earman (1992, 128–129) considers two alternative derivations of $p(\mathrm{T}|\mathrm{X},\mathrm{E}) > p(\mathrm{T}|\mathrm{E})$ where assumptions different from Jeffrey's (δ) carry the burden of the argument. One of them is the inequality

(ϕ) $p(\mathrm{T}|\mathrm{X}) > p(\mathrm{T}|\neg\mathrm{X}, \neg\mathrm{X}')$.

but it is questionable whether this suffices to circumvent the above objections. What Earman demands here is very close to what is supposed to be shown: that learning $T \vdash E$ is more favorable to T than learning that T gives no definite prediction for the occurrence of E or $\neg E$. In the light of the above arguments against (δ) and in the absence of independent arguments in favor of (ϕ), this condition just seems to beg the question.

The second alternative derivation of $p(T|X) > p(T)$ relies on the equality

(ψ) $p(X \vee X') = 1$.

However, as Earman admits himself, this condition is too strong: it amounts to demanding that upon formulating T, the scientist was certain that it either implied E or $\neg E$. In practice, such relationships are rather discovered gradually. As Earman continues, discussing the case of GTR:

> the historical evidence goes against this supposition: [...] Einstein's published paper on the perihelion anomaly contained an incomplete explanation, since, as he himself noted, he had no proof that the solution of the field equations [...] was the unique solution for the relevant set of boundary conditions (Earman, 1992, 129)

Taking stock, we conclude that Garber, Jeffrey, Niiniluoto and Earman make interesting proposals for solving the dynamic Problem of Old Evidence, but that their solutions are either incomplete or based on highly problematic assumptions. We will now show how their approach to the dynamic POE can be improved.

Solving the Dynamic Problem of Old Evidence: Alternative Explanations

A problem with the traditional GJN approaches is that they require constraints on degrees of belief (e.g., Jeffrey's (δ) or Earman's (ψ)) that are either implausibly strong or too close to the desired confirmation-theoretic conclusion $p(T|X,E) > p(T|E)$ itself. To remedy this defect, we propose to take into account whether alternatives to T adequately explain E. Let the propositions X and Y be defined as follows:

- $X \stackrel{\text{def}}{=} T$ *adequately explains* (or accounts for) E.

- Y $\stackrel{\text{def}}{=}$ *some alternative* to T (=T') *adequately explains* (or accounts for) E.

Now, consider the following four ordinal constraints on the degrees of belief of a rational Bayesian agent:

$$p(\mathrm{T}|\mathrm{X}, \neg\mathrm{Y}) > p(\mathrm{T}|\neg\mathrm{X}, \neg\mathrm{Y}) \tag{3.4}$$

$$p(\mathrm{T}|\mathrm{X}, \neg\mathrm{Y}) > p(\mathrm{T}|\neg\mathrm{X}, \mathrm{Y}) \tag{3.5}$$

$$p(\mathrm{T}|\mathrm{X}, \mathrm{Y}) > p(\mathrm{T}|\neg\mathrm{X}, \mathrm{Y}) \tag{3.6}$$

$$p(\mathrm{T}|\mathrm{X}, \mathrm{Y}) \geq p(\mathrm{T}|\neg\mathrm{X}, \neg\mathrm{Y}) \tag{3.7}$$

Let's examine each of these four constraints in turn, assuming that E is either a certainty or very probable. Suppose that $\neg$Y is the case and that no alternative to T adequately explains E. Then, (3.4) asserts that T is more probable given X (=T explains E) than given $\neg$X (=T does not explain E). Judgments of evidential relevance translate into judgments of evidential support, if there is no alternative to explain E.

(3.5) is an even less controversial variant of the same proposition. If T is the only explanation of E, T is more probable than if it does not explain E and there is at least one good alternative. In other words, (3.4) and (3.5) say that T's being the only adequate explanation of E confers a greater probability on T than any possibility which implies that T does not adequately explain E. These two constraints strike us as pretty uncontroversial.

Constraint (3.6) also seems very plausible. If there are alternatives to T that adequately explain E, T is more plausible if it explains E than if it doesn't. This constraint mirrors the reasoning in (3.4) for the case that there are alternatives to T.

The fourth and final inequality (3.7) says that T is at least as probable, given the supposition that both T and some alternative scientific theory adequately explain E (i.e., given X $\wedge$ Y) as it is given the supposition that no scientific theory adequately explains E (i.e., given $\neg$X $\wedge$ $\neg$Y). It might even be compelling to rank $p(\mathrm{T}|\mathrm{X}, \mathrm{Y})$ strictly higher in one's comparative confidence ranking than $p(\mathrm{T}|\neg\mathrm{X}, \neg\mathrm{Y})$. After all, X $\wedge$ Y implies that T adequately explains old evidence E, whereas $\neg$X $\wedge$ $\neg$Y implies that T does not adequately explain E. On the other hand, one might also reasonably maintain that both suppositions place T and its alternatives on a par with respect to explaining E, and so they shouldn't confer different probabilities on T. Both of these positions are compatible with (3.7). The only thing

(3.7) rules out is the claim that T is more probable given E's inexplicability ($\neg X \wedge \neg Y$) than it is given E's multiple explicability by both T and some alternative to T ($X \wedge Y$). As such, (3.7) also seems eminently reasonable.

Now, the desired conclusion (3.1) follows from (3.4)–(3.7). To be precise, we can prove the following general result (see also Fitelson and Hartmann, 2016)

Theorem 3.1 *For propositions* $T, E, X, Y \in \mathcal{L}$ *with probability measure* $p \in \mathfrak{P}$, *let* $0 < p(T) < 1$. *Then conditions* (3.4)–(3.7) *jointly entail* $p(T|X) > p(T)$.

Of course, this result also applies to the case where we have already conditionalized on E and $p(E) = 1$:

Corollary 3.1 *For propositions* $T, E, X, Y \in \mathcal{L}$ *with probability measure* $p \in \mathfrak{P}$, *let* $0 < p(T) < 1$. *Then the analogues of conditions* (3.4)–(3.7) *for the probability distribution* $p(\cdot|E)$ *jointly entail* $p(T|X, E) > p(T|E)$.

This approach has the following three distinct advantages over the traditional GJN approaches.

(i) Our approach does not require the assumption that $p(E) = 1$. It may be true that our constraints (3.4)–(3.7) are most plausible given the background assumption that E is known with certainty. But, we think (3.4)–(3.7) retain enough of their plausibility, given only the weaker assumption that E is known with near certainty (i.e., $p(E) \approx 1$).

(ii) Our approach only rests on the ordinal constraints (3.4)–(3.7), not on judgments of degrees of confirmation.

(iii) Our approach is not restricted to cases in which T (and/or alternatives T') explain E in a deductive-nomological way. That is, our approach covers all cases in which scientists come to learn that their theory adequately explains E, not only those cases in which scientists learn that their theory entails E (or explains E deductive-nomologically).

A slight disadvantage of this approach is that conditions (3.4)–(3.7) are themselves phrased as confirmation judgments. That is, the solution of the Problem of Old Evidence (whether X confirms T) depends on which truth-functional combinations of X and Y confirm T. Such judgments may

be considered to be too close to what is supposed to be shown. We think that our assumptions are plausible enough to withstand this criticism, but we also present an alternative solution where the condtions are phrased in terms of the *likelihoods* of T and X on E. But first, we move on to the static POE.

The Static Problem of Old Evidence: A Counterfactual Perspective

The static Problem of Old Evidence is concerned with describing why old evidence E is evidentially relevant for theory T (Eells, 1985). The relation of evidential relevance is supposed to be independent of the moment when the evidence was observed, when a relationship between theory and evidence was discovered, and so on. It corresponds to the question "Why is E at all—in the present as well as in the future—a reason for preferring T over its competitors?" By definition, this question is hard to answer in the Bayesian framework, which is in the first place a theory of confirmation as *change* in degree of belief.

Christensen (1999) contends that the choice of a particular confirmation measure may help us to resolve the static POE. Take the measure $s^*(\mathrm{T},\mathrm{E}) = p(\mathrm{T}|\mathrm{E}) - p(\mathrm{T}|\neg\mathrm{E})$. If T entails E, as in the GTR example, then $\neg$E also entails $\neg$T, which implies $p(\mathrm{T}|\neg\mathrm{E}) = 0$ and $s^*(\mathrm{T},\mathrm{E}) = p(\mathrm{T}|\mathrm{E}) > 0$. Hence E confirms E. According to s^*, old evidence E can substantially confirm theory T whereas the degree of confirmation is zero for measures that compare the prior and posterior probability of T, such as $d(\mathrm{T},\mathrm{E}) = p(\mathrm{T}|\mathrm{E}) - p(\mathrm{T})$ or $r(\mathrm{T},\mathrm{E}) = \log(p(\mathrm{T}|\mathrm{E})/p(\mathrm{T}))$. Choosing the "right" confirmation measure therefore resolves the POE.

This approach strikes us as problematic. First, it is questionable whether s^* is a good explicatum for degree of confirmation. In Variation 2, we have argued that s^* fails to satisfy important adequacy criteria for degree of confirmation, such as Final Probability Incrementality. Also in general, the measure sensitivity of Christensen's proposal is somewhat awkward: the challenge posed by the POE consists in showing that E raises the probability of T. Inequality (3.1) does not target the degree of confirmation conferred by old evidence. Therefore, a solution that depends on the choice of a particular confirmation measure is less general than we desire.

Second, Christensen's move has its merits for cases where $p(\mathrm{E})$ is close to, but not entirely equal to one. But in the classical POE where $p(\mathrm{E}) = 1$, $p(\mathrm{T}|\neg\mathrm{E})$ may not have a clear-cut definition since $p(\mathrm{T}|\neg\mathrm{E}) = p(\mathrm{T} \wedge \neg\mathrm{E})/p(\neg\mathrm{E})$ involves a division by zero. We could solve this problem by evaluating $p(\mathrm{T}|\neg\mathrm{E})$ not via the Ratio Analysis of conditional probability, but as a counterfactual degree of belief: suppose that $\neg\mathrm{E}$ were the case, how likely would T be? But then, Christensen's solution proposal is more than an appeal to a particular confirmation measure: it requires a specific approach to conditional degree of belief which needs to be spelled out in more detail.

Such an attempt is made by Colin Howson (1984, 1985, 1991). He gives up the Bayesian explication of confirmation as positive probabilistic relevance relative to actual degrees of belief. Rather, he suggests to evaluate the confirmation relation with respect to a counterfactual degree of belief function where E is not taken for granted:

> [T]he Bayesian assesses the contemporary support E gives T by how much the agent would change his odds on T *were he now* to come to know E [...] In other words, the theory is explicitly a theory of dispositional properties of the agent's belief-structure [...]. (Howson, 1984, 246, original emphasis)

According to this account, conditional probabilities such as $p(\mathrm{E}|\mathrm{T})$ and $p(\mathrm{E}|\neg\mathrm{T})$ should not be understood as our actual degree of belief in E supposing T or $\neg$T: this would be equal to one since E is already known and the equation

$$1 = p(\mathrm{E}) = p(\mathrm{E}|\mathrm{T})p(\mathrm{T}) + p(\mathrm{E}|\neg\mathrm{T})p(\neg\mathrm{T})$$

would imply that also $p(\mathrm{E}|\mathrm{T}) = p(\mathrm{E}|\neg\mathrm{T}) = 1$. Hence, $p(\cdot|\mathrm{T})$ must be a different belief function: it describes those degrees of belief that we would have in E if we knew nothing about E and T were the case. If T is a statistical hypothesis, then we just add T hypothetically to our background knowledge and calculate the probability of E conditional on this assumption. For example, we could evaluate the probability of two heads and three tails in five i.i.d. tosses of a fair coin (T: $\theta = 0.5$) and a biased coin (T': $\theta = 0.6$) and infer $p(\mathrm{E}|\mathrm{T}) \approx 0.31$ and $p(\mathrm{E}|\mathrm{T}') \approx 0.23$. Similarly, in the GTR example, we could conclude that $p(\mathrm{E}|\mathrm{T}) = 1$ because GTR implies the Mercury perihelion shift, whereas $p(\mathrm{E}|\mathrm{T}') \ll 1$ for Newtonian

mechanics and other theories that do not make definite predictions about E. In general, the probability distributions $p(\cdot|\mathrm{T})$ and $p(\cdot|\mathrm{T}')$ describe our conditional degrees of belief in pieces of evidence, supposing that T or T′. In such a setting, we can meaningfully compare $p(\mathrm{E}|\mathrm{T})$ and $p(\mathrm{E}|\mathrm{T}')$ which is not possible if we interpret these probabilities as our actual degrees of belief in E, knowing that T (e.g., via Ratio Analysis).

How does this formalism translate into confirmation judgments? If p is an "impartial" prior probability distribution, that is, $p(\mathrm{T}) = p(\mathrm{T}')$, then we infer that $p(\mathrm{T}|\mathrm{E}) > p(\mathrm{T}'|\mathrm{E})$ if and only if $p(\mathrm{E}|\mathrm{T}) > p(\mathrm{E}|\mathrm{T}')$ (proof omitted). Final Probability Incrementality then implies that E confirms T more than it confirms T′. That is, our judgment on the conditional probability of E given T, relative to minimal background knowledge, translates into a judgment of evidential support if the priors do not favor one of the hypotheses. In our interpretation, the static POE abstracts away from background knowledge at a particular point in history, and therefore, the counterfactual approach to conditional probability is an adequate tool to tackle it. We will know show how learning the proposition $\mathrm{X} = \mathrm{T} \vdash \mathrm{E}$ raises the probability of T relative to a counterfactual probability function as described above. That is, we solve the dynamic POE in a framework that is typical of the static POE.

Solving the Hybrid Problem of Old Evidence: Learning Explanatory Relationships

The aim of this section consists in showing that learning explanatory or deductive relationships between theory and evidence can raise our degree of belief in the theory. In other words, learning $\mathrm{X} \stackrel{\text{def}}{=}$ (T adequately explains E) raises the subjective probability of T relative to a probability function where E is taken for granted. This looks like a formulation of the dynamic POE, but things are more subtle: we are not interested in whether X confirmed T for the scientist who discovered X (and relative to her degrees of belief), but in whether X should confirm T for all scientists in the community, irrespective of their actual degrees of belief. This second question is related to the static POE where scientific confirmation is independent of anybody's subjective degrees of belief at a particular time. Hence, our question in this section—the confirmatory impact of explanatory discoveries—needs to be phrased relative to a counterfactual

probability function, like in the static POE. That's why we would like to call it the hybrid POE.

What kind of probability function should be chosen? In our view, $p(\cdot|\pm \mathrm{T})$ should represent the degrees of belief of a scientist who has a sound understanding of theoretical principles and their impact on observational data, conditional on the assumption that T or $\neg$T is the case (see also Earman, 1992, 134). Such degrees of belief are required for making routine judgments in assessing evidence and reviewing journal articles: How probable would the actual evidence E be if T were true? How probable would E be if T were false? When T and $\neg$T are two definite statistical hypotheses, like in Howson's coin toss example, such judgments are immediately given by the corresponding sampling distribution. But even in more general contexts, such judgments may be straightforward, or a matter of consensus in the scientific community.

We now formulate constraints on an agent's conditional degrees of belief in the hybrid POE. The first condition characterizes the elementary inferential relations between E, T and X:

[1] $p(\mathrm{E}|\mathrm{T},\mathrm{X}) = 1$

If T is true and T entails E, then E can be regarded as a certainty. In this scenario, X codifies a strict deductive relation between T and E. Later, we will relax this condition in order to cover more general explannatory dependencies.

To motivate the second constraint, note that learning the predictions of a refuted hypothesis is irrelevant to our assessment of the *plausibility* of the predicted events. For instance, the astrological theory on which Nostradamus based his predictions is in all probability wrong. Upon learning the content of his predictions (e.g., the third World War starting in 2048), we should neither raise nor lower our credence in the events that his theory predicted to happen. This motivates the equation $p(\mathrm{E}|\neg\mathrm{T},\mathrm{X}) = p(\mathrm{E}|\neg\mathrm{T})$, or written differently, $p(\mathrm{E}|\neg\mathrm{T},\mathrm{X}) = p(\mathrm{E}|\neg\mathrm{T},\neg\mathrm{X})$. Again, these degrees of belief ought to be interpreted in the neo-Ramseyian, counterfactual sense: supposing that T has been disproved, would learning something about the predictions of T affect our confidence in the occurrence of E? Plausibly not since T has ceased to be relevant for empirical forecasts. Hence we demand that if $\neg$T is already known, then learning X or $\neg$X does not change the probability of E. However, E should still be *possible* if T were false. Hence:

[2] $p(\mathrm{E}|\neg\mathrm{T},\mathrm{X}) = p(\mathrm{E}|\neg\mathrm{T},\neg\mathrm{X}) > 0$

Finally, we have the following inequality:

[3]
$$p(\mathrm{E}|\mathrm{T},\neg\mathrm{X}) < \frac{1-p(\mathrm{X}|\neg\mathrm{T})}{1-p(\mathrm{X}|\mathrm{T})}\,\frac{p(\mathrm{X}|\mathrm{T})}{p(\mathrm{X}|\neg\mathrm{T})}$$

This condition demands that the value of $p(\mathrm{E}|\mathrm{T},\neg\mathrm{X})$ be smaller than the threshold on the right hand side. When X and T are positively relevant to each other or probabilistically independent, [3] is trivially satisfied since in that case, $p(\mathrm{X}|\mathrm{T}) \geq p(\mathrm{X}|\neg\mathrm{T})$, implying that the right hand side of [3] is greater or equal than one. But also if X and T are negatively relevant to each other, [3] is plausibly satisfied. When the mutual negative impact of X and T is not too strong, the two quotients in [3] are close to 1, and the inequality will be satisfied as long as $p(\mathrm{E}|\neg\mathrm{X},\mathrm{T})$ is not too close to 1 itself. Given that T is assumed to be true, but that by $\neg$X, it does not fully account for E, E should be far from certain for a rational Bayesian agent. Here it is essential that the conditional probabilities are interpreted in the counterfactual sense. Otherwise, we would always obtain $p(\mathrm{E}|\cdot) = 1$ for old evidence E, regardless of which proposition stands to the right of the vertical dash. In the (plausible) case of independence of T and X, this would contradict [3] ($p(\mathrm{E}|\mathrm{T},\neg\mathrm{X}) < 1$).

Together with the unproblematic assumption that neither T nor $\neg$T is a certainty beforehand ($0 < p(\mathrm{T}) < 1$), these three conditions are jointly sufficient to prove that X confirms T relative to E.

Theorem 3.2 *Let* T, X, and E *be three propositions of* $\mathcal{L}$ *with probability measure* $p \in \mathfrak{P}$ *and* $0 < p(\mathrm{T}) < 1$. *Let the following three conditions be satisfied:*

[1]
$$p(\mathrm{E}|\mathrm{T},\mathrm{X}) = 1;$$

[2]
$$p(\mathrm{E}|\neg\mathrm{T},\mathrm{X}) = p(\mathrm{E}|\neg\mathrm{T},\neg\mathrm{X}) > 0;$$

[3]
$$p(\mathrm{E}|\mathrm{T},\neg\mathrm{X}) < \frac{1-p(\mathrm{X}|\neg\mathrm{T})}{1-p(\mathrm{X}|\mathrm{T})}\,\frac{p(\mathrm{X}|\mathrm{T})}{p(\mathrm{X}|\neg\mathrm{T})}$$

Then, X *confirms* T *relative to (old evidence)* E*; that is,* $p(\mathrm{T}|\mathrm{E},\mathrm{X}) > p(\mathrm{T}|\mathrm{E})$.

In other words, if E is taken for granted, learning X raises the conditional degree of belief in T if conditions [1]-[3], whose adequacy we have justified above, are accepted. Or in other words: if we knew little or nothing about the observational history of a discipline, and we were informed that E, then discovering X would raise our confidence in T. This seems to be a perfectly reasonable sense in which X is evidence for T, relative to E.

This theorem solves the hybrid POE based on a combination of strategies. The main idea stems from the GJN models—the confirming event is the discovery that T accounts for/explains E—, but the relevant constraints are spelled out in terms of conditional degrees of belief which are interpreted in a counterfactual sense, like in the static POE. Then, with the help of Bayes' Theorems, the constraints transfer to bounds on the conditional probability of T given E and X.

In many cases of scientific reasoning, the condition $p(\mathrm{E}|\mathrm{T},\mathrm{X}) = 1$ may be too strong. It may apply well to the Mercury perihelion shift, which is deductively implied by GTR, but it may fail to cover cases where T accounts for E in a less rigorous manner (Earman, 1992, 121)—see also Fitelson (2015). If we allow for a weaker interpretation of X, e.g., as providing some explanatory mechanism, then we are faced with the possibility that even if we are certain that T is true, and that T explains E, the conditional degree of belief in E may not be a certainty. $p(\mathrm{E}|\mathrm{T}) < 1$ could even make sense if the relationships between T and E are deductive: the proof of X could so complex that the involved scientists have some doubts about its soundness and refrain from assigning it maximal degree of belief. Again, Fermat's Last Theorem may be a plausible intuition pump.

For covering this case, we prove another theorem which covers the case of $p(\mathrm{E}|\mathrm{T},\mathrm{X}) = 1 - \varepsilon$ for some small $\varepsilon > 0$.

Theorem 3.3 *Let* T, X, and E *be three propositions of* $\mathcal{L}$ *with probability measure* $p \in \mathfrak{P}$ *and* $0 < p(\mathrm{T}) < 1$. *Let the following three conditions be satisfied:*

[1']

$$p(\mathrm{E}|\mathrm{T},\mathrm{X}) = 1 - \varepsilon \text{ for some } 0 < \varepsilon < 1;$$

[2']

$$p(\mathrm{E}|\neg\mathrm{T},\mathrm{X}) = p(\mathrm{E}|\neg\mathrm{T},\neg\mathrm{X}) > 0$$

[3']

$$p(\mathrm{E}|\mathrm{T},\neg\mathrm{X}) < (1-\varepsilon)\,\frac{1-p(\mathrm{X}|\neg\mathrm{T})}{1-p(\mathrm{X}|\mathrm{T})}\,\frac{p(\mathrm{X}|\mathrm{T})}{p(\mathrm{X}|\neg\mathrm{T})}$$

Then, X *confirms* T *relative to (old evidence)* E*; that is,* $p(\mathrm{T}|\mathrm{E},\mathrm{X}) > p(\mathrm{T}|\mathrm{E})$.

The motivations and justifications for the above assumptions are the same like in Theorem 3.2. [1′] just accounts for lack of full certainty about the old evidence, and [2′] is identical to [2]. Moreover, condition [3] of Theorem 3.2 can, with the same line of reasoning, be extended to condition [3′] in Theorem 3.3. [3′] sharpens [3] by a factor of $1 - \varepsilon$, but leaves the qualitative argument for [3] intact. As long as $p(\mathrm{E}|\mathrm{T},\neg\mathrm{X})$ and $p(\mathrm{E}|\mathrm{T},\mathrm{X})$ decrease by roughly the same margin, the result of Theorem 3.2 transfers to Theorem 3.3.

Thus, we can extend the novel solution of POE to the case of residual uncertainty about the old evidence E—a case that is highly relevant for case studies in the history of science. If we compare this solution of the POE to Jeffrey's and Earman's proposals, we note that our assumptions [1], [2] and [3] are silent on whether Jeffrey's (δ)—or Earman's (ϕ) and (ψ), for that matter—is true or false. For a proof with the help of Branden Fitelson's PrSAT package for Mathematica (Fitelson, 2008), see Sprenger (2015). Hence we can discard Jeffrey's dubious assumption (δ) that increasing empirical content makes a theory more plausible, without jeopardizing our own results.

We have thus provided a solution of the POE that successfully tackles a hybrid version of the POE. Notably, our solution makes less demanding assumptions than Jeffrey and Earman. Conceptually, however, this solution is anchored in the static POE and in the use of counterfactual (rather than actual) probability functions. We now discuss the repercussions of our results on the general debate about POE, and the role of Bayesian Confirmation Theory in scientific reasoning.

Discussion

This variation has analyzed Bayesian attempts to solve the Problem of Old Evidence (POE), and it has proposed two new solutions. We have started with a distinction between the static and the dynamic aspect of the problem. Simplifying a bit, we can say that the static POE deals with the question of providing an account of conditional probability where $p(\mathrm{E}|\mathrm{T}) > p(\mathrm{E}|\neg\mathrm{T})$ for old evidence E, demonstrating the evidential relevance of E for T. The dynamic problem, on the other hand, deals with the challenge to provide reasonable constraints on p such that

$p(\mathrm{T}|\mathrm{X},\mathrm{E}) > p(\mathrm{T}|\mathrm{E})$, with X denoting the proposition that T implies or explains E ($\mathrm{T} \vdash \mathrm{E}$).

We first presented our criticism of existing solutions to the dynamic POE in the footsteps of Garber, Jeffrey and Niiniluoto (GJN). Our first model of the dynamic POE was based on judgments when T is confirmed by the presence and absence of alternative hypotheses that could account for old evidence E. The second model was based on constraints on the conditional degrees of belief $p(\mathrm{E}|\pm\mathrm{T},\pm\mathrm{X})$.

To avoid that these degrees of belief are equal to one because E is old evidence, we interpreted p as a properly counterfactual degree of belief function and not as describing our actual degrees of belief. Indeed, in scientific practice, we typically interpret $p(\mathrm{E}|\mathrm{T})$ and $p(\mathrm{E}|\neg\mathrm{T})$ as *principled* statements about the predictive import of $\pm$T on E, without referring to our complete observational record. Such judgments are part and parcel of scientific reasoning, e.g., in statistical inference, where theories T, T', etc. impose definite probability distributions on the observations, and our degrees of belief $p(\mathrm{E}|\mathrm{T})$, $p(\mathrm{E}|\mathrm{T}')$, etc. follow suit. However, the standard account of conditional degree of belief (Ratio Analysis) does not have this property. We therefore suggested that the counterfactual account of conditional degree of belief presented in the introduction can contribute to solving the (static and dynamic) POE.

It is also worth mentioning that our treatment of the POE allows for a distinction between theories that have been constructed to explain the old evidence E and those that explain E surprisingly (like Einstein's GTR). In the first case, we would not speak about proper confirmation. Indeed, if we accommodate the parameter values of a general theory T such that it explains the old evidence E, whatever this evidence turns out to be, then X is actually a certainty: $p(\mathrm{X}) = 1$. This is because T has been designed to explain E. As a consequence, $p(\mathrm{T}|\mathrm{E},\mathrm{X}) = p(\mathrm{T}|\mathrm{E})$ and X fails to confirm T. Whereas in the case of a *surprising* discovery of an explanatory relation between T and E, $p(\mathrm{X}) < 1$. The degree of confirmation that X confers on T gets the bigger the more surprising X is—in line with our intuitive judgment that surprising explanations have special cognitive value. In general, there seem to be strong parallels between the POE and the prediction vs. accommodation debate in philosophy of science (e.g., Hitchcock and Sober, 2004). Future research could investigate this relationship in more detail, and also come up with case studies about scientific reasoning with

old evidence, enabling a better evaluation of our solution proposals. Other research projects that spring to mind are an integration of the POE with explanatory reasoning in science ($\rightarrow$ Variation 7) and providing a solution of the POE in terms of learning conditionals. After all, the dynamic POE can be described as learning the (strict) conditional that if T, then also E. We can use our account of learning conditionals from Variation 1 in order to describe conditions when learning $\mathrm{T} \vdash \mathrm{E}$ raises the probability of T.

Finally, a general, but popular critique of Bayesian approaches to the POE is inspired by the view that the POE poses a principled and insoluble problem for Bayesian Confirmation Theory. For instance, Glymour writes at the end of his discussion of the POE:

> [...] our judgment of the relevance of evidence to theory depends on the perception of a structural connection between the two, and [...] degree of belief is, at best, epiphenomenal. In the determination of the bearing of evidence on theory there seem to be mechanisms and strategems that have no apparent connections with degrees of belief. (Glymour, 1980, 92-93)

What Glymour argues here is not so much that a specific formal aspect of the Bayesian apparatus (e.g., logical omniscience) prevents it from solving the POE, but that these shortcomings are a symptom of a more general inadequacy of Bayesian Confirmation Theory: the inability to capture *structural relations* between evidence and theory. This criticism should not be misunderstood as claiming that confirmation has to be conceived of as an objective relation that is independent of contextual knowledge or contingent background assumptions. Rather, it suggests that solutions to the dynamic POE mistake an increase in degree of belief for a structural relation between T and E. But what makes E relevant for T is not the increase in degree of belief $p(\mathrm{T}|\mathrm{E}) > p(\mathrm{T})$, but the entailment relation between T and E. Hence Glymour's verdict that Bayesian Confirmation Theory gives "epiphenomenal" results.

To our mind, this criticism commits two oversights. First, solutions to the static POE answer Glymour's challenge by showing how the concept of evidential support can be interpreted in an way that is not bound to belief updating at a particular point in time. We have tried to provide such an account in Section 3.3. Second, the criticism is too fundamental to be a source of genuine concern: it is not specific to the (dynamic) POE

or one of its solutions, but it attacks the entire Bayesian explication of confirmation as increase in firmness. However, as we have seen in the previous variation, Bayesian Confirmation Theory can point to a lot of success stories: resolving the tacking by conjunction paradoxes, the raven paradox, the new riddle of induction, and so on. What we have shown here is that confirmation by old evidence might be added to this list. The next variation moves on to an argument that we already touched upon in Section 3.2: failure to find adequate alternatives confirms a theory.

Proofs of the Theorems

Proof of Theorem 3.1: Let $\mathfrak{a} \stackrel{\text{def}}{=} p(\mathrm{T}|\mathrm{X} \wedge \neg \mathrm{Y})$, $\mathfrak{b} \stackrel{\text{def}}{=} p(\mathrm{T}|\mathrm{X} \wedge \mathrm{Y})$, $\mathfrak{c} \stackrel{\text{def}}{=} p(\mathrm{T}|\neg \mathrm{X} \wedge \neg \mathrm{Y})$, $\mathfrak{d} \stackrel{\text{def}}{=} p(\mathrm{T}|\neg \mathrm{X} \wedge \mathrm{Y})$, $x \stackrel{\text{def}}{=} p(\neg \mathrm{Y}|\mathrm{X})$, and $y \stackrel{\text{def}}{=} p(\neg \mathrm{Y}|\neg \mathrm{X})$. Given these assignments, (3.4)–(3.7) translate as follows.

$$\mathfrak{a} > \mathfrak{c} \qquad\qquad \mathfrak{a} > \mathfrak{d}$$
$$\mathfrak{b} > \mathfrak{d} \qquad\qquad \mathfrak{b} \geq \mathfrak{c}$$

Suppose that $\mathfrak{a} \in (0, 1]$, $\mathfrak{d} \in [0, 1)$, and $\mathfrak{b}, \mathfrak{c}, x, y \in (0, 1)$. [1] Then, (3.4)–(3.7) jointly entail

$$\mathfrak{a}x + \mathfrak{b}(1 - x) > \mathfrak{c}y + \mathfrak{d}(1 - y).$$

And, by the law of total probability, we have:

$$\begin{aligned} p(\mathrm{T}|\mathrm{X}) &= \mathfrak{a}x + \mathfrak{b}(1 - x) \\ p(\mathrm{T}|\neg \mathrm{X}) &= \mathfrak{c}y + \mathfrak{d}(1 - y) \end{aligned}$$

Thus, (3.4)–(3.7) jointly entail $p(\mathrm{T}|\mathrm{X}) > p(\mathrm{T}|\neg \mathrm{X})$, which entails $p(\mathrm{T}|\mathrm{X}) > p(\mathrm{T})$. □

Proof of Theorem 3.2: First, we define

$$e_1 = p(\mathrm{E}|\mathrm{T}, \mathrm{X}) \qquad e_2 = p(\mathrm{E}|\neg \mathrm{T}, \mathrm{X})$$
$$e_3 = p(\mathrm{E}|\mathrm{T}, \neg \mathrm{X}) \qquad e_4 = p(\mathrm{E}|\neg \mathrm{T}, \neg \mathrm{X})$$
$$t = p(\mathrm{T}) \qquad r = p(\mathrm{X}|\mathrm{T})$$
$$t' = p(\mathrm{T}|\mathrm{X}) \qquad \bar{r} = p(\mathrm{X}|\neg \mathrm{T})$$

By making use of [1] ($e_1 = 1$), [2] ($e_2 = e_4 > 0$), and the Extension Theorem $p(\mathrm{X}|\mathrm{Z}) = p(\mathrm{X}|\mathrm{Y}, \mathrm{Z})p(\mathrm{Y}|\mathrm{Z}) + p(\mathrm{X}|\neg \mathrm{Y}, \mathrm{Z})p(\neg \mathrm{Y}|\mathrm{Z})$, we can quickly verify the identities

$$p(\mathrm{E}|\mathrm{T}) = p(\mathrm{E}|\mathrm{T}, \mathrm{X})p(\mathrm{X}|\mathrm{T}) + p(\mathrm{E}|\mathrm{T}, \neg \mathrm{X})p(\neg \mathrm{X}|\mathrm{T})$$

[1] The only two conditional credences that may reasonably take extremal values here are $\mathfrak{a}$ and $\mathfrak{d}$. If T is the only theory that adequately explains E, then it may be reasonable to assign T maximal credence. And, if some alternative to T (e.g., T′) is the only theory that adequately explains E, then it may be reasonable to assign minimal credence to T. This is why we allow $\mathfrak{a} \in (0, 1]$ and $\mathfrak{d} \in [0, 1)$. The other conditional credences involved in our theorem (*i.e.*, $\mathfrak{b}$, $\mathfrak{c}$, x, y) should, in general, take non-extreme values.

$$
\begin{aligned}
&= r + e_3\,(1-r)\\
p(\mathrm{E}|\neg\mathrm{T}) &= p(\mathrm{E}|\neg\mathrm{T},\mathrm{X})p(\mathrm{X}|\neg\mathrm{T}) + p(\mathrm{E}|\neg\mathrm{T},\neg\mathrm{X})p(\neg\mathrm{X}|\neg\mathrm{T})\\
&= e_2\bar{r} + e_4(1-\bar{r})\\
&= e_2
\end{aligned}
$$

that will be useful later. Second, we note that by Bayes' Theorem and assumption [1],

$$
\begin{aligned}
p(\mathrm{T}|\mathrm{E},\mathrm{X}) &= p(\mathrm{T}|\mathrm{X})\frac{p(\mathrm{E}|\mathrm{T},\mathrm{X})}{p(\mathrm{E}|\mathrm{X})}\\
&= \left(1 + \frac{p(\neg\mathrm{T}|\mathrm{X})}{p(\mathrm{T}|\mathrm{X})}\frac{p(\mathrm{E}|\neg\mathrm{T},\mathrm{X})}{p(\mathrm{E}|\mathrm{T},\mathrm{X})}\right)^{-1}\\
&= \left(1 + \frac{1-t'}{t'}\cdot e_2\right)^{-1} \qquad (3.8)
\end{aligned}
$$

Third, we observe that by [1], [2] and the above identities for $p(\mathrm{E}|\mathrm{T})$ and $p(\mathrm{E}|\neg\mathrm{T})$,

$$
\begin{aligned}
p(\mathrm{T}|\mathrm{E}) &= p(\mathrm{T})\frac{p(\mathrm{E}|\mathrm{T})}{p(\mathrm{E})}\\
&= \left(1 + \frac{p(\neg\mathrm{T})}{p(\mathrm{T})}\frac{p(\mathrm{E}|\neg\mathrm{T})}{p(\mathrm{E}|\mathrm{T})}\right)^{-1}\\
&= \left(1 + \frac{1-t}{t}\frac{e_2}{r + e_3\,(1-r)}\right)^{-1} \qquad (3.9)
\end{aligned}
$$

We also note by [3] that $e_3 < \frac{1-\bar{r}}{\bar{r}}\,\frac{r}{1-r}$. Note that it is implicit in condition [3] that $1 > p(\mathrm{X}|\mathrm{T}), p(\mathrm{X}|\neg\mathrm{T}) > 0$ since otherwise, the expression would either be undefined (divison by zero), or $p(\mathrm{E}|\mathrm{T},\neg\mathrm{X})$ would have to be smaller than zero, which is impossible.

This allows us to derive

$$
\begin{aligned}
r + e_3\,(1-r) &< r + \frac{1-\bar{r}}{\bar{r}}\,\frac{r}{1-r}\,(1-r)\\
&= r\cdot\left(1 + \frac{1-\bar{r}}{\bar{r}}\right)\\
&= \frac{r}{\bar{r}}
\end{aligned}
$$

and consequently also

$$
\frac{1}{r + e_3\,(1-r)} > \frac{\bar{r}}{r} \qquad (3.10)
$$

Moreover, note the equality

$$\frac{1-t'}{t'} = \frac{p(\neg \mathrm{T}|\mathrm{X})}{p(\mathrm{T}|\mathrm{X})} = \frac{p(\mathrm{X}|\neg T)}{p(\mathrm{X}|T)} \cdot \frac{p(\neg \mathrm{T})}{p(\mathrm{T})} = \frac{\bar{r}}{r} \cdot \frac{1-t}{t} \tag{3.11}$$

All this implies that

$$\begin{aligned}
\frac{p(\mathrm{T}|\mathrm{E},\mathrm{X})}{p(\mathrm{T}|\mathrm{E})} &\overset{(5.12),(3.9)}{=} \left(1 + \frac{1-t'}{t'} \cdot e_2\right)^{-1} \cdot \left(1 + \frac{1-t}{t} \frac{e_2}{r + e_3\,(1-r)}\right) \\
&\overset{(3.10)}{>} \left(1 + \frac{1-t'}{t'} \cdot e_2\right)^{-1} \cdot \left(1 + \frac{1-t}{t} \cdot \frac{\bar{r}}{r} e_2\right) \\
&\overset{(3.11)}{=} \left(1 + \frac{1-t'}{t'} \cdot e_2\right)^{-1} \cdot \left(1 + \frac{1-t'}{t'} e_2\right) \\
&= 1,
\end{aligned}$$

completing the proof. The second line has also used that $e_2 > 0$, as ensured by [2], and that $t, r, \bar{r} \in (0,1)$ also implies $t' \in (0,1)$. □

Proof of Theorem 3.3: By means of performing the same steps as in the proof of Theorem 3.2, we can easily verify the equalities

$$\begin{aligned}
p(\mathrm{T}|\mathrm{E},\mathrm{X}) &= \left(1 + \frac{p(\neg \mathrm{T}|\mathrm{X})}{p(\mathrm{T}|\mathrm{X})} \frac{p(\mathrm{E}|\neg \mathrm{T},\mathrm{X})}{p(\mathrm{E}|\mathrm{T},\mathrm{X})}\right)^{-1} \\
&= \left(1 + \frac{1-t'}{t'} \cdot \frac{e_2}{1-\varepsilon}\right)^{-1} \\
p(\mathrm{T}|\mathrm{E}) &= \left(1 + \frac{p(\neg \mathrm{T})}{p(\mathrm{T})} \frac{p(\mathrm{E}|\neg \mathrm{T})}{p(\mathrm{E}|\mathrm{T})}\right)^{-1} \\
&= \left(1 + \frac{1-t}{t} \frac{e_2 \bar{r} + e_4(1-\bar{r})}{(1-\varepsilon)r + e_3\,(1-r)}\right)^{-1} \\
&= \left(1 + \frac{1-t}{t} \frac{e_2}{(1-\varepsilon)r + e_3\,(1-r)}\right)^{-1}
\end{aligned}$$

where we have made use of [1′] and [2′]=[2]. We also note that [3′] implies

$$\begin{aligned}
(1-\varepsilon)r + e_3\,(1-r) &< (1-\varepsilon)r + (1-\varepsilon)\,\frac{1-\bar{r}}{\bar{r}}\,\frac{r}{1-r}\,(1-r) \\
&= (1-\varepsilon) \cdot r \cdot \left(1 + \frac{1-\bar{r}}{\bar{r}}\right)
\end{aligned}$$

$$= \quad (1-\varepsilon)\,\frac{r}{\bar{r}}$$

and therefore also

$$\frac{1}{(1-\varepsilon)r + e_3\,(1-r)} > \frac{\bar{r}}{(1-\varepsilon)\,r} \tag{3.12}$$

This brings us to the final calculation:

$$\begin{aligned}
\frac{p(\mathrm{T}|\mathrm{E},\mathrm{X})}{p(\mathrm{T}|\mathrm{E})} &= \left(1 + \frac{1-t}{t}\frac{1}{(1-\varepsilon)r + e_3\,(1-r)}\,e_2\right) \cdot \left(1 + \frac{1-t'}{t'} \cdot \frac{e_2}{1-\varepsilon}\right)^{-1} \\
&> \left(1 + \frac{1-t}{t}\frac{\bar{r}}{(1-\varepsilon)\,r}e_2\right) \cdot \left(1 + \frac{1-t}{t} \cdot \frac{\bar{r}}{r} \cdot \frac{e_2}{1-\varepsilon}\right)^{-1} \\
&= 1,
\end{aligned}$$

where we have simultaneously applied Equations (3.11) and (3.12) in the second line. This completes the proof.

□

Variation 4: The No Alternatives Argument

In the previous chapters, we have described how empirical observations confirm or disconfirm a scientific hypothesis by means of probabilistic relevance. For example, the observation of a black raven raises the probability of the hypothesis that all ravens are black, and certain clicks in a particle detector make us more confident in the existence of the top quark. However, there are situations where empirical evidence is unattainable over long periods of time. Such situations arise with particular force in contemporary high energy physics, where the characteristic empirical signatures of theories like Grand Unified Theories or string theory must be expected to lie many orders of magnitude beyond the reach of present day experimental technology. They are also entirely common in scientific fields such as palaeontology or anthropology, where scientists must rely on the scarce and haphazard empirical evidence they happen to find in the ground. Interestingly, scientists are at times quite confident regarding the adequacy of their theories even when empirical evidence is largely or entirely absent. In such cases, they base their trust on what we call **non-empirical evidence**: evidence for T that neither falls into the (broadly construed) intended domain of T nor is logically or probabilistically related to T. Such evidence can, for example, consist in observations about the research process leading up to the construction of T, or the standing of T in the research community. The wording "non-empirical" is not supposed to express a rationalist or idealist concept of theory confirmation.

From an empiricist point of view, arguments relying on non-empirical evidence may be regarded as mere speculation: they neither contribute to actual theory confirmation nor do they have objective scientific weight. We challenge this claim by exploring the following case: scientists develop considerable trust in a theory T because despite considerable efforts, no

alternatives to T have been found that meet crucial theoretical and empirical constraints. We call this argument the **No Alternatives Argument (NAA)** and set up a Bayesian model to prove its validity. The name of the argument stems from its crucial premise (scientists have not yet found a suitable alternative to T); it does not draw the conclusion that there are no alternatives to T. If valid, the NAA would demonstrate the possibility of non-empirical theory confirmation.

A variant of NAA is actually often used in politics and is well-known under the acronym **TINA: "There is no alternative"**. The former British prime minister Margaret Thatcher was famous for promoting her politics of privatization and economic liberalization by means of this slogan. The present German chancellor, Angela Merkel, has also used TINA to defend her political course in the euro crisis (and more recently, in the refugee crisis) as *alternativlos*, that is, without alternatives: "If the euro falls, Europe will fall, too". An investigation of the NAA will therefore not only shed light on patterns of scientific reasoning and the possibility of non-empirical theory confirmation, but also elucidate the validity of argument patterns that are frequently used in the political debate. But above all, it complements and completes the investigation of confirmatory arguments in science that we have begun in the first three variations.

The setup of this variation is very simple: Section 4.1 introduces a formal model of the NAA and makes plausible assumptions for the epistemic ramifications of non-empirical evidence. By contrast, Section 4.2 presents our main results, discusses their significance and explores an application to Inference to the Best Explanation. For more details, see Dawid et al. (2015). Section 4.3 reports the proofs.

Modeling the No Alternatives Argument

We would like to investigate whether observing a lack of alternatives to T confirms the empirical adequacy of T. Two disclaimers to begin with: if two theories make the same predictions, then we consider them to be identical. Moreover, we would like to sidestep debates about scientific realism (→ Variation 5) and focus on the empirical adequacy of T rather than on its truth.

On a Bayesian account of confirmation, the subjective degree of belief in T has to be raised by the lack of alternatives. But how is this possible?

After all, a lack of alternatives is neither deductively nor probabilistically implied by T. It does not even fall into the intended domain of T. Does this observation then qualify as (non-empirical) evidence in an *argument from ignorance*, such as: if T were not empirically adequate, then we would have disproved it before (Walton, 1995; Hahn and Oaksford, 2007; Sober, 2009)? More generally, how is the Bayesian addressing the problem that there may always be unconceived alternatives to T which may explain the available evidence as well as T, or even better (Stanford, 2006)?

The most plausible way to solve this problem is to deploy a two step process. First, we find a statement that predicts the failure to find alternatives to T. Then, we show that this statement provides evidential support for the empirical adequacy of T. In the case of NAA, our non-empirical evidence $\neg \mathrm{F_A}$ consists in the fact that scientists have not found any alternatives to a specific solution of a research problem, despite looking for them with considerable energy and for a long time. Obviously, the **number of satisfactory alternatives to T** matters here. A small number of available alternatives renders $\neg \mathrm{F_A}$ more likely than a large number of alternatives: in the latter case, one might expect that scientists would have already found one of them.

The number of scientific theories which can account for a certain set of data is in turn relevant for the degree of belief in the empirical adequacy of T. The more alternatives exist, the less is it likely that a particular theory is empirically adequate. In other words, the observation that scientists have not yet found an alternative to T indicates that there are not too many alternatives to T, and thus figures as an argument for T. A lower number of possible scientific theories that accounts for a certain set of empirical data increases the degree of belief that our actual theory is adequate—see also the *argument from no choice* (Dawid, 2006, 2009).

Based on this reasoning, we introduce a random variable A measuring the number of alternatives to T and taking values in the natural numbers. $\mathrm{A_k} := \{A = k\}$ expresses the proposition that there are k adequate and distinct alternatives which satisfy a set of theoretical constraints $\mathcal{C}$, are consistent with the existing data $\mathcal{D}$, and give distinguishable predictions for the outcome of some set $\mathcal{E}$ of future experiments. We will later show that, via its effect on the $\mathrm{A_k}$, the non-empirical evidence $\neg \mathrm{F_A}$ confirms the empirical adequacy of T under plausible conditions.

Inferences about the number of alternatives to a theory T naturally de-

pend on what counts as a genuine alternative. This is, in turn, sensitive to the specific scientific context. Therefore we leave the individuation problem to the scientific community which typically has the best grip on what should count as a distinct theory. Moreover, for the No Alternatives Argument, we only require the premise that the number of alternatives to T *possibly be finite*. In other words, we are not certain a priori that there are infinitely many alternatives.

In order to motivate this assumption, we assume that different theories provide different solutions to a given research problem. That is, theories which only differ in a detail, such as the precise value of a parameter or the existence of a physically meaningless dummy variable, do not count as different theories. For example, the simple Higgs model in particle physics is treated as *one* theory, although the Higgs particle could have different mass values. Generally, if it were enough to slightly modify the value of a certain parameter in order to arrive at a new theory, then coming up with new theories would be an easy and not very creative task. Inventing a novel mechanism or telling a new story of why a certain phenomenon occurred is much harder.

Scientists often formulate no alternatives arguments at the level of general conceptual principles while allowing for a large spectrum of specific realizations of those principles. For example, since the 1980s particle physicists strongly supported a no alternatives argument with respect to the Higgs mechanism. That is, they believed that no alternatives to a gauge theory that was spontaneously broken by a Higgs sector of scalar fields could account for the available empirical data. Physicists strongly believed, based on a NAA, that the Higgs sector would be observed at the LHC experiment but did not have particular trust in any of the specific models of the Higgs sector. In this case, the NAA clearly was placed at the level of physical principles rather than specific models.

Following this line of reasoning, we reconstruct NAA based on the notion that there exists a specific but unknown number k of possible scientific theories. As stated above, these theories have to satisfy constraints $\mathcal{C}$, explain data $\mathcal{D}$ and predict the outcomes of the experiments $\mathcal{E}$. We will then show that failure to find an alternative to T raises the probability of T being empirically adequate and thus confirms T.

To do so, we introduce the binary propositional variables H and F_A. As before, H takes the values

H Theory T is empirically adequate.

¬H Theory T is not empirically adequate.

and F_A takes the values

$\mathrm{F_A}$ The scientific community has found an alternative to T that fulfills $\mathcal{C}$, explains $\mathcal{D}$ and predicts the outcomes of $\mathcal{E}$.

$\neg\mathrm{F_A}$ The scientific community has not yet found an alternative to T that fulfills $\mathcal{C}$, explains $\mathcal{D}$ and predicts the outcomes of $\mathcal{E}$.

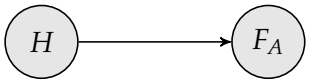

Figure 4.1: The Bayesian Network representation of the two-propositions scenario.

We would now like to explore under which conditions $\neg\mathrm{F_A}$ confirms H, that is, when

$$p(\mathrm{H}|\neg\mathrm{F_A}) > p(\mathrm{H})\,. \tag{4.1}$$

Figure 4.1 suggests a direct influence of H on F_A. But since a direct influence is blocked by the non-empirical nature of F_A, we introduce a third variable A which mediates the connection between H and F_A. A has values in the natural numbers, and $\mathrm{A_k}$ corresponds to the proposition that there are exactly k hypotheses that fulfil $\mathcal{C}$, explain $\mathcal{D}$ and predict the outcomes of $\mathcal{E}$.

We should also note that the value of F_A—that scientists find/do not find an alternative to T—does not only depend on the number of available alternatives, but also on the difficulty of the problem, the cleverness of the scientists, or the available computational, experimental, and mathematical resources. Call the variable that captures these complementary factors D, and let it take values in the natural numbers, with $\mathrm{D_j} := \{D = \mathrm{j}\}$ and $d_j := p(\mathrm{D_j})$. The higher the values of D, the more difficult the problem. For the purpose of our argument, it is not necessary to assign a precise operational meaning to the various levels of D—see condition **A3** later on. It is clear that D has no direct influence on A and H (or vice versa), but that it matters for F_A and that this influence has to be represented in our Bayesian Network.

We now list five plausible assumptions that we need for showing the validity of the No Alternatives Argument.

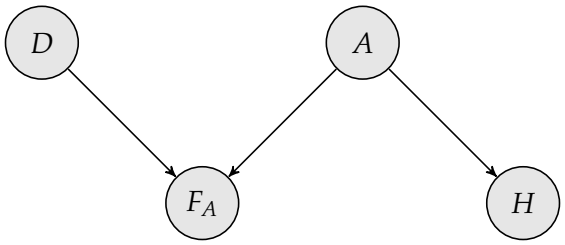

Figure 4.2: The Bayesian Network representation of the four-propositions scenario.

A1 The variable H is conditionally independent of F_A given A:

$$H \perp\!\!\!\perp F_A | A \tag{4.2}$$

Hence, learning that the scientific community has or has not found an alternative to T does not alter our belief in the empirical adequacy of T if we already know the value of A (e.g., that there are exactly k viable alternatives).

A2 The variable D is (unconditionally) independent of A:

$$D \perp\!\!\!\perp A \tag{4.3}$$

Recall that D represents the aggregate of those context-sensitive factors that affect whether scientists find an alternative to T, but that are not related to the number of suitable alternatives. In other words, D and A are orthogonal to each other by construction.

These are our most important assumptions, and we consider them to be eminently sensible. Figure 4.2 shows the corresponding Bayesian Network. To complete it, we have to specify the prior distribution over D and A and the conditional distributions over F_A and T, given the values of their parents. This is done in the following three assumptions.

A3 The conditional probabilities

$$f_{kj} := p(\neg \mathrm{F_A} | \mathrm{A_k}, \mathrm{D_j}) \tag{4.4}$$

are non-increasing in k for all $j \in \mathbb{N}$ and non-decreasing in j for all $k \in \mathbb{N}$.

The (weak) monotonicity in the first argument reflects the intuition that for fixed difficulty of a problem, a higher number of available

alternatives increases the chance of finding one of them. In other words, the more reasonable alternatives to T are around, the less likely it is that scientists fail to find one. The (weak) monotonicity in the second argument reflects the intuition that increasing difficulty of a problem does not increase the likelihood of finding an alternative to T, provided that the number of alternatives to T is fixed.

A4 The conditional probabilities

$$t_k := p(\mathrm{H}|\mathrm{A_k}) \tag{4.5}$$

are non-increasing in k.

This assumption reflects the intuition that an increase in the number of alternative theories does not make it more likely that scientists have already identified an empirically adequate theory.

A5 There is at least one pair (i, k) with $i < k$ for which (i) $a_i\, a_k > 0$ where $a_k := p(\mathrm{A_k})$, (ii) $f_{ij} > f_{kj}$ for some $j \in \mathbb{N}$, and (iii) $t_i > t_k$.

This assumption demands that at least two of the a_i be greater than zero, and it strengthens A3 and A4 by demanding that the f_{ij} and t_i not be constant in i.

Results and Discussion

The previous section has set up a formal model of the NMA in a Bayesian Network (see Figure 4.2) and made five assumptions on how the variables in that network hang together (see A1-A5). With these assumptions in hand, we can now show the following main result:

Theorem 4.1 (Validity of the NAA) *If A takes values in the natural numbers $\mathbb{N}$ and assumptions* **A1** *to* **A5** *hold, then $\neg\mathrm{F_A}$ confirms* H, *that is, $p(\mathrm{H}|\neg\mathrm{F_A}) > p(\mathrm{H})$.*

We have therefore shown that $\neg\mathrm{F_A}$ confirms the empirical adequacy of T under rather weak and plausible assumptions.

In line with the introduction of A in section 4.1, we have assumed that A only takes values in the natural numbers. This might be seen as evading the skeptical argument that there may be infinitely many (theoretically adequate, empirically successful, ...) alternatives to T. Therefore we now

extend the theorem by explicitly allowing for the possibility $\mathrm{A}_\infty := \{A = \infty\}$, and we modify our assumptions accordingly. In particular, we observe that **A5** entails $p(A_\infty) < 1$, define $f_{\infty j} := p(\neg \mathrm{F_A} | \mathrm{A}_\infty, \mathrm{D_j})$, $t_\infty := p(\mathrm{H}|\mathrm{A}_\infty)$ and demand that

$$f_{ij} \geq f_{\infty j} \; \forall i, j \in \mathbb{N} \qquad f_{\infty i} \leq f_{\infty j} \; \forall i, j \in \mathbb{N} \text{ with } i < j \tag{4.6}$$

$$t_i \geq t_\infty \; \forall i \in \mathbb{N}\,. \tag{4.7}$$

These requirements naturally extend assumptions **A3** and **A4** to the case of infinitely many alternatives. Then, we obtain the following generalization of the NAA:

Theorem 4.2 (Validity of the NAA, infinitely many alternatives) *If A takes values in $\mathbb{N} \cup \{\infty\}$ and assumptions* **A1** *to* **A5** *hold together with their extensions* (4.6) *and* (4.7)*, then $\neg \mathrm{F_A}$ confirms* H*, that is, $p(\mathrm{H}|\neg \mathrm{F_A}) > p(\mathrm{H})$.*

In other words, even if we concede to the skeptic that there may be infinitely many alternatives to T, she must still acknowledge the validity of NAA as long as her degrees of belief satisfy $p(\mathrm{A}_\infty) < 1$. This is, to our mind, a quite substantial and surprising result. For a long time, philosophy of science has focused on logical and probabilistic relations between theory and evidence and neglected other forms of theory confirmation. However, the above theorem demonstrates that non-empirical evidence (in our specific sense of the word) can raise our confidence in the empirical adequacy of a theory.

Note that only a dogmatic skeptic who insists on $p(\mathrm{A}_\infty) = 1$ can deny the validity of NAA. Theorem 4.2 convinces anyone who does not want to commit herself with respect to the probability of (in)finitely many genuine alternatives to T. (Recall that theories do not count as distinct when they are just different realizations of the same mechanism or principle.) Convincing such a fair and non-committal skeptic is, to our mind, much more important than convincing a dogmatic who just denies our premises by insisting on $p(\mathrm{A}_\infty) = 1$. That assumption might only be warranted if we were interested in the *truth* of T rather than its empirical adequacy.

We have seen that the NAA can be used in support of a proposed theory. The question remains, however, whether the resulting support is of significant strength and whether using the NAA in a specific situation is justified. To facilitate matters, we conduct this analysis for the finite case (Theorem 4.1); the infinite case is analogous.

The Bayesian Network representation of NAA in Figure 4.2 suggests that the NAA cannot easily obtain confirmatory significance without supportive reasoning. According to Figure 4.2, $\neg \mathrm{F_A}$ may confirm an instance of D—limitations to the scientists' abilities to solve difficult problems—as well as an instance of A, such as limitations to the number of possible theories. It is then easy to see that for all $l \in \mathbb{N}$,

$$p(\mathrm{D_l}|\neg \mathrm{F_A}) = \frac{p(\mathrm{D_l}, \neg \mathrm{F_A})}{p(\neg \mathrm{F_A})} = \frac{d_l \cdot \sum_k a_k \, f_{kl}}{\sum_{j,k} d_j \, a_k \, f_{kj}}$$

which may be greater than $p(\mathrm{D_l})$ for plausible assignments of parameter values. To successfully apply NAA, one has to amend the *qualitative* claim shown above with a *comparative* claim, namely that $\neg \mathrm{F_A}$ confirms T more than the claim $\{D > K\}$ ("the problem is just very difficult") for some threshold K. But such a statement is sensitive to the specific parameter assignments as well as to the chosen confirmation measure—and therefore hard to prove in general. Applied to the political context (the TINA variant of NAA), this result means that the failure to find viable alternatives to a particular policy does indeed confirm that the chosen policy may be the best one, in the sense of confirmation as increase in firmness. But without additional assumptions, it would be invalid to conclude that the probability has increased substantially, let alone that we should now be *confident* (e.g., with a degree of belief greater than 1/2) that the chosen policy is the best one. Be this as it may, we would like to stress that even without such a comparative assessment, the *validity* of the NAA, that it raises the probability of T being empirically adequate, is a surprising and substantial philosophical result.

The NAA some interesting philosophical perspectives. In particular, *Inference to the Best Explanation* (Lipton, 2004; Douven, 2011) can, to a certain extent, be explicated in terms of NAA. The fact that no other genuinely satisfactory explanation has been found corresponds to the failure to find alternatives in the NAA pattern. Then, the structure of the argument is similar; only the interpretation changes from "empirically adequate" to "best/only satisfactory explanation". The relevant propositional variables then read as follows:

H The hypothesis T is the only satisfactory explanation of phenomenon $\mathcal{E}$.

$\neg$H The hypothesis T is the only satisfactory explanation of phenomenon $\mathcal{E}$.

F_A The scientific community has found an alternative to T that explains $\mathcal{E}$.

$\neg F_A$ The scientific community has not yet found an alternative to T that explains $\mathcal{E}$.

A_k then denotes the number of alternatives that explain $\mathcal{E}$. It is not difficult to motivate analogues of A1-A5 for this interpretation of our propositional variables, and to derive that $\neg F_A$ confirms H. We conjecture that some Inferences to the Best Explanations in science are actually NAA's in disguise: they take the failure of attempts to find an alternative as a reason to infer the truth or empirical adequacy of the only available hypothesis.

Second, the reasoning scheme of the NAA is similar to eliminative induction in the style of Francis Bacon, or more recently, Arthur Conan Doyle's figure Sherlock Holmes: "when you have eliminated the impossible, whatever remains, however improbable, must be the truth". Could we use the NAA as a case study for creating deeper links between Bayesian and eliminative inference (Earman, 1992; Forber, 2011)?

Third, our theoretical analysis should be complemented by more case studies in science: from string theory as a classical application of the NAA (Dawid, 2006, 2009), and also from other disciplines where empirical evidence is scarce, such as palaeontology, archaelogy or anthropology.

Fourth and last, the relationship between NAA and the TINA argument in public policy should be investigated more closely. Can the endorsement of a defended policy really be defended with a type of NAA? Or does "failure to find viable alternatives" mean something very different in the political context, invalidating the application of the NAA in that domain?

In the next variation, we will address the issue of scientific realism which looms behind the NAA and related reasoning schemes. In particular, we will show how the model underlying NAA, with an explicit probability distribution over the number of alternatives to T, can be used to develop a sophisticated Bayesian version of the No Miracles Argument (NMA).

Proof of the Theorems

Proof of Theorem 4.1: $\neg \mathrm{F_A}$ confirms H if and only if $P(\mathrm{H}|\neg \mathrm{F_A}) - P(\mathrm{H}) > 0$, that is, if and only if

$$\Delta := P(\mathrm{H}, \neg \mathrm{F_A}) - P(\mathrm{H})P(\neg \mathrm{F_A}) > 0.$$

We now apply the theory of Bayesian Networks to the structure depicted in Figure 4.2, using assumptions **A1** ($H \perp\!\!\!\perp F_A|A$) and **A2** ($D \perp\!\!\!\perp A$):

$$\begin{aligned}
P(\neg \mathrm{F_A}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P(\neg \mathrm{F_A}|\mathrm{A_i}, \mathrm{D_j})P(\mathrm{A_i}, \mathrm{D_j}) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, a_i\, f_{ij} \\
P(\mathrm{H}) &= \sum_{k=0}^{\infty} P(\mathrm{H}|\mathrm{A_k})P(\mathrm{A_k}) = \sum_{k=0}^{\infty} t_k\, a_k \\
P(\mathrm{H}, \neg \mathrm{F_A}) &= \sum_{i=0}^{\infty} P(\neg \mathrm{F_A}, \mathrm{H}|\mathrm{A_i})P(\mathrm{A_i}) = \sum_{i=0}^{\infty} a_i\, P(\neg \mathrm{F_A}|\mathrm{A_i})\, P(\mathrm{H}|\mathrm{A_i}) \\
&= \sum_{i=0}^{\infty} a_i\, t_i \left(\sum_{j=0}^{\infty} P(\neg \mathrm{F_A}|\mathrm{A_i}, \mathrm{D_j})\, P(\mathrm{D_j}|\mathrm{A_i}) \right) \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, a_i\, t_i\, f_{ij}
\end{aligned}$$

Hence, we obtain, using $\sum_{k\in\mathbb{N}} a_k = 1$,

$$\begin{aligned}
\Delta &= \left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, a_i\, t_i\, f_{ij} \right) - \left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, a_i\, f_{ij} \right) \left(\sum_{k=0}^{\infty} a_k\, t_k \right) \\
&= \left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, a_i\, t_i\, f_{ij} \right) \left(\sum_{k=0}^{\infty} a_k \right) - \left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, a_i\, f_{ij} \right) \left(\sum_{k=0}^{\infty} t_k\, a_k \right) \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \left(d_j\, a_i\, a_k\, t_i\, f_{ij} - d_j\, a_i\, a_k\, t_k\, f_{ij} \right) \\
&= \sum_{j=0}^{\infty} d_j \sum_{i=0}^{\infty}\sum_{k\neq i=0}^{\infty} a_i\, a_k\, f_{ij}\, (t_i - t_k) \\
&= \sum_{j=0}^{\infty} d_j \sum_{i=0}^{\infty}\sum_{k>i} \left(a_i\, a_k\, f_{ij}\, (t_i - t_k) + a_k\, a_i\, f_{kj}\, (t_k - t_i) \right) \\
&= \sum_{j=0}^{\infty} d_j \sum_{i=0}^{\infty}\sum_{k>i} a_i\, a_k\, (f_{ij}\, (t_i - t_k) + f_{kj}\, (t_k - t_i)) \\
&= \sum_{j=0}^{\infty} d_j \sum_{i=0}^{\infty}\sum_{k>i} a_i\, a_k\, (t_i - t_k)\, (f_{ij} - f_{kj})
\end{aligned}$$

$$> \quad 0$$

because of **A3**-**A5** taken together: **A3** entails that the difference $(f_{ij} - f_{kj})$ is non-negative, **A4** does the same for the $(t_i - t_k)$, and **A5** entails that these differences are strictly positive for at least one pair (i, k). Hence, the entire double sum is strictly positive. □

Proof of Theorem 4.2: We perform essentially the same calculations as in the proof of Theorem 4.1 and additionally include the possibility $\{A = \infty\}$.[2] Defining $f_{\infty j} := P(\neg \mathrm{F_A} | \mathrm{D_j A_\infty})$ leads us to the equalities

$$\begin{aligned} P(\neg \mathrm{F_A}) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} d_j \, a_i \, f_{ij} + \sum_{j=0}^{\infty} d_j \, a_\infty \, f_{\infty j} \\ P(\mathrm{H}) &= \sum_{k=0}^{\infty} t_k \, a_k + t_\infty \, a_\infty \\ P(\neg \mathrm{F_A}, \mathrm{H}) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} d_j \, a_i \, t_i \, f_{ij} + \sum_{j=0}^{\infty} d_j \, t_\infty \, a_\infty \, f_{\infty j} \end{aligned}$$

from which it follows, using $\lim_{K \to \infty} \sum_{k=1}^{K} a_k = 1 - a_\infty$, that

$$\begin{aligned} P(\neg \mathrm{F_A}) P(\mathrm{H}) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} d_j \, t_k \, a_i \, a_k \, f_{ij} + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} d_j \, t_\infty \, a_i \, a_\infty \, f_{ij} \\ &\quad + \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} d_j \, t_k \, a_k \, a_\infty \, f_{\infty j} + \sum_{j=0}^{\infty} d_j \, t_\infty \, a_\infty^2 \, f_{\infty j} \\ P(\neg \mathrm{F_A}, \mathrm{H}) &= \frac{1}{1 - a_\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} d_j \, t_i \, a_i \, a_k f_{ij} + \sum_{j=0}^{\infty} d_j \, t_\infty \, a_\infty \, f_{\infty j} \end{aligned}$$

With the definition

$$\Delta := \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} d_j \, t_i \, a_i \, a_k f_{ij} - \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} d_j \, t_k \, a_i \, a_k \, f_{ij}$$

we observe that $\Delta > 0$, as shown above in the proof of Theorem 4.1 (the parameter values satisfy the relevant conditions **A3-A5**). Noting that **A5** requires $a_\infty < 1$, it follows that

$$P(\neg \mathrm{F_A}, \mathrm{H}) - P(\mathrm{H}) \, P(\neg \mathrm{F_A})$$

[2] The notation suggests that ∞ is already included in the summation index, but the infinity sign on top of the sum is just the shortcut for the limit of the sequence of all natural numbers. Thus, the case $A = \infty$ has to be treated separately.

$$
\begin{aligned}
&= \Delta + \frac{a_\infty}{1-a_\infty} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} d_j\, t_i\, a_i\, a_k f_{ij} + \sum_{j=0}^{\infty} d_j\, t_\infty\, a_\infty\, (1-a_\infty)\, f_{\infty j} \\
&\quad - \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, t_\infty a_i\, a_\infty\, f_{ij} - \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_j\, t_i\, a_i\, a_\infty\, f_{\infty j} \\
&= \Delta + \frac{a_\infty}{1-a_\infty} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} d_j\, t_i\, a_i\, a_k f_{ij} + \frac{a_\infty}{1-a_\infty} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} d_j\, t_\infty\, a_i\, a_k\, f_{\infty j} \\
&\quad - \frac{a_\infty}{1-a_\infty} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} d_j\, t_\infty a_i\, a_k\, f_{ij} - \frac{a_\infty}{1-a_\infty} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} d_j\, t_i\, a_i\, a_k\, f_{\infty j} \\
&= \Delta + \frac{a_\infty}{1-a_\infty} \sum_{j=0}^{\infty} d_j \sum_{i=0}^{\infty}\sum_{k>i} a_i\, a_k (t_i f_{ij} + t_\infty f_{\infty j} - t_\infty f_{ij} - t_i f_{\infty j}) \\
&= \Delta + \frac{a_\infty}{1-a_\infty} \sum_{j=0}^{\infty} d_j \sum_{i=0}^{\infty}\sum_{k>i} a_i\, a_k (t_i - t_\infty)(f_{ij} - f_{\infty j}) \\
&> 0
\end{aligned}
$$

since the extensions of **A3** and **A4** imply $f_{ij} \geq f_{\infty j}$ and $t_i \geq t_\infty$ (equations (4.6) and (4.7)), independent of the values of i and j. □

Variation 5: Scientific Realism and the No Miracles Argument

The debate between scientific realists and anti-realists is one of the classics of philosophy of science, comparable to a soccer match between Brazil and Argentina. Realism comes in different varieties, e.g., metaphysical, semantic and epistemological realism (see Chakravartty, 2011, for a survey). The most ambitious and most contested form of realism is the epistemological thesis that we are justified to believe in the truth of our best scientific theories, and that they constitute knowledge of the external world (Boyd, 1983; Psillos, 1999, 2009). In this view, the existence of a mind-independent world (metaphysical realism) and the reference of theoretical terms to mind-independent entities (semantic realism) is usually presupposed—the real question concerns the epistemic status of our best scientific theories.

In this variation, we demonstrate how Bayesian methods can be used for clarifying and sharpening the debate between realists and anti-realists. It consists of three parts. In the first part, we explain the **No Miracles Argument (NMA)** and examine the well-known objection that the realist commits the base rate fallacy (Howson, 2000; Magnus and Callender, 2004). The second and third part respond to this issue. In the second part, we demonstrate how **observed stability of scientific theories** can make a case for scientific realism and we argue that the validity of the NMA crucially depends on the specific context of a scientific discipline. In the third part, we show how shifting from individual theories to a series of theories in a scientific discipline may alleviate the base rate fallacy. Both arguments are **refinements of the NMA within a Bayesian model**. Note that none of our arguments makes a sufficient case for scientific realism: the scope of the probabilistic NMA does not extend beyond claims to empirical adequacy. However, the NMA is necessary for parrying vital threats to the

realist view, such as Laudan's argument from Pessimistic Meta-Induction.

We do not argue that Bayesian philosophy of science should be aligned with a realist stance. Bayesian philosophy of science is a methodological approach to investigating substantial philosophical questions, not a specific answer to them. Rather, we show how Bayesian methods can be used to clearly articulate the realist argument, to investigate its validity and to determine its scope.

The Probabilistic No Miracles Argument

A major player in the debate about scientific realism is the **No Miracles Argument (NMA)**. It contends that the truth of our best scientific theories is the only hypothesis that does not make the astonishing predictive, retrodictive and explanatory success of science a mystery (Putnam, 1975, 73). If our best scientific theories did not correctly describe the world, why should we expect them to be successful at all? The truth of our best theories is an excellent, and perhaps the only explanation of their success. Therefore, we should accept the realist hypothesis: our best scientific theories are true and constitute knowledge of the world.

It is not entirely clear whether the NMA is an empirical or a super-empirical argument. As an argument from past and present success of our best scientific theories to their truth, it involves two major steps: the step from observed success to justified belief in empirical adequacy, and the step from justified belief in empirical adequacy to justified belief in truth (see Figure 5.1). The first of them is an empirical inference, the second most probably not: ordinary empirical evidence cannot distinguish between different theoretical structures that yield the same observable consequences.

Much philosophical discussion has been devoted to the second step of the NMA (e.g., Psillos, 1999; Lipton, 2004; Stanford, 2006), which seems in greater need of a philosophical defense. After all, the realist has to address the problem of underdetermination of theory by evidence. But also the first step of the NMA is far from trivial, and strengthening it against criticism is vital for the scientific realist. For instance, Laudan (1981) has argued that there have been lots of successful, but non-referring (and empirically inadequate) scientific theories. If Laudan were right, then the entire NMA would break down, even if objections to the second step

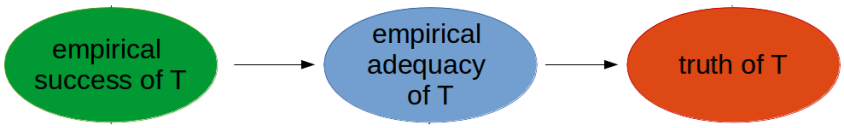

Figure 5.1: The structure of the NMA as a two-step argument from the empirical success of T to its truth. We conceptualize the NMA as an argument for the first inference in this figure, that is, for an inference from empirical success of T to its empirical adequacy.

could be answered successfully.

Such arguments do not only threaten full scientific realists, but also structural realists (Worrall, 1989) and some varieties of anti-realism that make substantial epistemic commitments. One of them is Bas van Fraassen's constructive empiricism (van Fraassen, 1980; Monton and Mohler, 2012). Proponents of this view deny that we have reasons to believe that our best scientific theories are literally true. However, they affirm that we are justified to believe in their observable parts. Thus they are also affected by criticism and defense of the first step of the NMA.

Hence, the first step of the NMA does not draw a sharp divide between realists and anti-realists. Rather, the debate takes place between those who derive epistemic commitments from the success of science, and those who deny them. This variation is devoted to exploring the question of whether such epistemic commitments are justified. For convenience, we stick to the traditional terminology and refer to the first group as "realists" and to the second group as "anti-realists". We begin with a Bayesian analysis of a simple form of the NMA.

We first apply the NMA to a particular scientific theory T which is predictively and explanatorily successful in a certain scientific domain. Since we only investigate arguments for the empirical adequacy of T, we introduce a propositional variable H—the hypothesis that T is empirically adequate. See Figure 5.2 for a simple Bayesian network representation of the dependence between H and the propositional variable S that represents the empirical success of T.

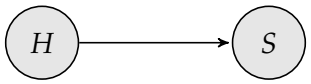

Figure 5.2: The Bayesian Network representation of the impact of H—the empirical adequacy of theory T—on the empirical success of T, denoted by S.

Expressed as a Bayesian argument, the simple NMA then runs as follows: S is much more probable if T is empirically adequate than if it is not. This can be expressed by the following two assumptions:

$\mathcal{A}_1$ $s := p(\mathrm{S}|\mathrm{H})$ is large.

$\mathcal{A}_2$ $s' := p(\mathrm{S}|\neg\mathrm{H}) < k \ll 1$.

From Bayes' Theorem, we can then infer

$$p(\mathrm{H}|\mathrm{S}) > p(\mathrm{H})$$

In other words, S confirms H: our degree of belief in the empirical adequacy of T is increased if T is successful.

Anti-realists object to the above argument that the inequality $p(\mathrm{H}|\mathrm{S}) > p(\mathrm{H})$ falls short of establishing the first step of the NMA. We are primarily interested in whether H is sufficiently probable given S, not in whether S raises our degree of belief in H. After all, the increase in probability could be negligibly small. The result $p(\mathrm{H}|\mathrm{S}) > p(\mathrm{H})$ does not establish that $p(\mathrm{H}|\mathrm{S}) > K$ for a critical threshold K, e.g., $K = 1/2$. We already know this distinction between posterior probability and incremental confirmation from Variation 2, under the name of confirmation as firmness vs. confirmation as increase in firmness.

More specifically, it has been argued that the NMA commits the **base rate fallacy** (Howson, 2000; Magnus and Callender, 2004). This is an unwarranted type of inference that frequently occurs in medicine. Consider a highly sensitive medical test which yields a positive result. On the other hand, the medical condition in question is very rare, that is, the base rate of the disease is very low. In such a case, the posterior probability of the patient having the disease, given the test, will still be quite low. Nonetheless, medical practitioners tend to disregard the low base rate and to infer that the patient really has the disease in question (e.g., Goodman, 1999a).

This objection can be elucidated by a brief inspection of Bayes' Theorem. Our quantity of interest is the posterior probability $p(\mathrm{H}|\mathrm{S})$, our

confidence in H given S. This quantity can be written as

$$\begin{aligned} p(\mathrm{H}|\mathrm{S}) &= \frac{p(\mathrm{H})\; p(\mathrm{S}|\mathrm{H})}{p(\mathrm{S})} \\ &= \left(1 + \frac{1-p(\mathrm{H})}{p(\mathrm{H})}\,\frac{p(\mathrm{S}|\neg\mathrm{H})}{p(\mathrm{S}|\mathrm{H})}\right)^{-1} \end{aligned} \tag{5.1}$$

which shows that $p(\mathrm{H}|\mathrm{S})$ is not only an increasing function in $p(\mathrm{S}|\mathrm{H})$ and a decreasing function in $p(\mathrm{S}|\neg\mathrm{H})$: its value crucially depends on the base rate or prior plausibility of H, $p(\mathrm{H})$ (=the prior plausibility of H).

Anti-realists claim that NMA is built on a base rate fallacy: from the high value of $p(\mathrm{H}|\mathrm{S})$ ("the empirical adequacy of T explains its success") and the low value of $p(\mathrm{S}|\neg\mathrm{H})$ ("success of T would be a miracle if T were not empirically adequate"), justified belief in H ("T is empirically adequate") is inferred. The probabilistic model of the NMA demonstrates that we need additional assumptions about $p(\mathrm{H})$ to warrant this inference. In the absence of such assumptions, the NMA does not entitle us to accept T as empirically adequate.

What do these considerations show? First, they expose that the NMA, reconstructed as a probabilistic inference to the posterior probability of H, is essentially subjective. After all, any weight of evidence in favor of H can be counterbalanced by a sufficiently skeptical prior, that is, a sufficiently low value assigned to $p(\mathrm{H})$. The realist needs to provide convincing reasons why $p(\mathrm{H})$ should not be arbitrarily close to zero, and such reasons will typically presuppose realist inclinations. This is a problem for those realists who claim that the NMA is an intersubjectively compelling argument in favor of scientific realism. Howson (2013, 211) concludes that due to the dependence on unconstrained prior degrees of belief, the NMA is, "as a supposedly objective argument, [...] dead in the water". See also Howson (2000, ch. 3), Lipton (2004, 196–198), and Chakravartty (2011).

What is more, a low value for our prior degree of belief in H may be more rational than a high value. Take, for example, Larry Laudan's argument from Pessimistic Meta-Induction (PMI): "I believe that for every highly successful theory in the past of science which we now believe to be a genuinely referring theory, one could find half a dozen successful theories which we now regard as substantially non-referring" (Laudan, 1981, 35). Why should our currently best theory T_n = T not suffer the same fate as it predecessors $\mathrm{T}_1, \ldots, \mathrm{T}_{n-1}$ which proved to be empirically

inadequate although they were once the best scientific theory?

PMI affects the values of $p(\mathrm{S}|\neg\mathrm{H})$ and $p(\mathrm{H})$ as follows: On the one hand, history teaches us that there have often been false theories that explained the data well (and were superseded later). In other words, empirically inadequate theories can be highly successful and $p(\mathrm{S}|\neg\mathrm{H})$ need not be low. On the other hand, PMI advises a low degree of belief that T is empirically adequate since in the past, comparable theories turned out to be false.

To substantiate these concerns, we conduct a numerical analysis of the probabilistic NMA. For the sake of simplicity, let us sharpen $\mathcal{A}_1$ to $s = p(\mathrm{S}|\mathrm{H}) = 1$: if theory T is empirically adequate, then it is also successful. Furthermore, define $s' := p(\mathrm{S}|\neg\mathrm{H})$ and let $h := p(\mathrm{H})$ be the prior probability of H. We now ask the question: for which values of s' and h is the posterior probability of H, $p(\mathrm{H}|\mathrm{S})$, greater than $1/2$? That is, when would it be more plausible to believe that T is empirically adequate than to deny it? Satisfying this condition is arguably a minimal requirement for the claim that the success of T entitles us to justified belief in its empirical adequacy.

By using Bayes' Theorem, we can easily calculate when the inequality $p(\mathrm{H}|\mathrm{S}) > 1/2$ is satisfied. Equation (5.1) brings us to the inequality

$$\frac{1}{2} < \left(1 + s' \frac{1-h}{h}\right)^{-1}$$

which can be written as

$$s' < \frac{h}{1-h} \tag{5.2}$$

See Figure 5.3 for a graphical illustration.

However, inequality (5.2) is not easy to satisfy. As mentioned above, false theories and models often make accurate predictions and perform well on other cognitive values (see Frigg and Hartmann, 2012, for an overview). Classical examples that are still used today involve Newtonian mechanics, the Lotka-Volterra model from population biology (e.g., Weisberg, 2007) and Rational Choice Theory. Hence, the value of $s' = p(\mathrm{S}|\neg\mathrm{H})$ should not be too low. But if we choose, for example, $s' = 1/4$, then we would require $p(\mathrm{H}) \in [1/3, 1]$ to satsify inequality (5.2) and to make the NMA work! In other words, the NMA only works for theories which are already likely to be empirically adequate. What is more, for a mildly skeptical prior such as $p(\mathrm{H}) = 0.05$, the value of s' would have to be in

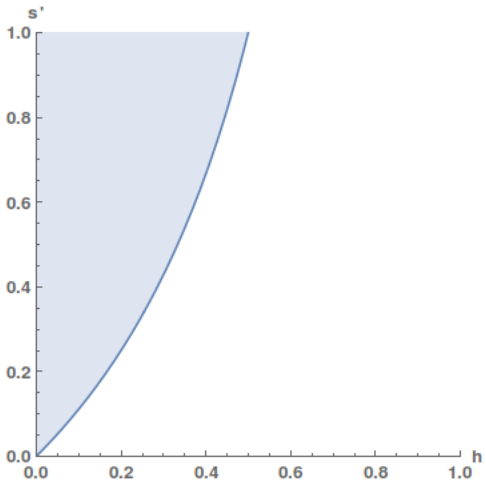

Figure 5.3: The scope of the No Miracles Argument, represented graphically. $p(\mathrm{H}|\mathrm{S}) > 1/2$ is the case in the white area below the line.

the range $[0, 0.053]$. This amounts to making the assumption that only the empirical adequacy (or truth) of a scientific theory can explain its success. But this is essentially a realist premise which the anti-realist would refuse to accept. She could point to the existence of unconceived alternatives (Stanford, 2006, ch. 6), the explanatory successes of false theories, etc. In other words: the simple probabilistic model of the NMA demonstrates that (1) to the extent that the NMA is valid, its premises presuppose realist inclinations; (2) to the extent that the NMA builds on premises that are neutral between the realist and the anti-realist, it fails to be valid.

Are things thus hopeless for the realist who wants to convince the anti-realist that the NMA is a good argument? Does "all realistic hope of resuscitating the [no miracles] argument [fail]", as Howson (2013, 211) writes? Perhaps not necessarily so. So far, the probabilistic NMA only took into account the predictive and explanatory success of T. Now we also consider the stability of scientific theories as evidence for scientific realism. This move is related to the No Alternatives Argument (NAA, $\rightarrow$ Variation 4), and in fact, our probabilistic model will be inspired by NAA-type reasoning.

Extending the No Miracles Argument to Stable Scientific Theories

Recently, Ludwig Fahrbach (2009, 2011) has argued that the stability of major scientific theories in the second half of the 20th century provides a strong argument in favor of scientific realism. In this section, we show how observing **theoretical stability in a scientific discipline** could give a boost to the probabilistic NMA.

Fahrbach's argument is mainly based on scientometric data. He observes an exponential growth of scientific activity in the 20th century, with a doubling of scientific output every 20 years (Meadows, 1974). He also notes that at least 80% of all scientific work has been done since the year 1950 and observes that our best scientific theories (e.g., the periodic table of elements, optical and acoustic theories, the theory of evolution, etc.) were stable during that period of time. That is, they did not undergo rejection or major conceptual change. On the other hand, Laudan's examples in favor of PMI stem from the early periods of science, e.g., the caloric theory of heat, the ether theory in physics, or the humoral theory in medicine.

For giving a fair assessment of PMI, we have to take into account the amount of scientific work done in a particular period. This implies, for example, that the period 1800–1820 should receive much less weight than the period 1950–1970. According to Fahrbach, PMI fails because most "theory changes occurred during the time of the first 5% of all scientific work ever done by scientists" (Fahrbach, 2011, 149). If PMI were valid, we should have observed more substantial theory changes or scientific revolutions in the recent past. However, although the theories of modern science often encounter difficulties, revolutionary turnovers do not (or only very rarely) happen. According to Fahrbach, PMI stands refuted—or at the very least, it is not more rational than an optimistic meta-induction.

Certainly, Fahrbach's model is quite simplified. For example, the number of published papers in a discipline need not be a reliable indicator of the probative value of a scientific theory. However, we are not interested in whether Fahrbach sketches an accurate picture of 20th century science. Rather, we will use a Bayesian framework for showing that such observations can *in principle* support the realist thesis. More precisely, we explore if observations of long-term stability expand the scope of the NMA. To this end, we refine our probabilistic model from the previous section.

As before, the propositional variable H expresses the empirical adequacy of theory T, and S denotes the predictive, retrodictive and explanatory success of T. The integer-valued random variable A expresses the number of satisfactory alternatives to T, and A_j is our shortcut for the proposition $A = j$. Like in the previous variation on the NAA, we demand that genuine alternatives satisfy a set of (context-dependent) theoretical constraints $\mathcal{C}$, be consistent with the currently available data $\mathcal{D}$, and give distinguishable predictions for the outcome of some set $\mathcal{E}$ of future experiments. In line with our focus on empirical adequacy rather than truth, we do not distinguish between empirically equivalent theories with different theoretical structures. Finally, major theory change in the domain of T is denoted by C, and absence of change and theoretical stability by ¬C. "Theory change" is understood in a broad sense, including scenarios where rivalling theories emerge and end up co-existing with T.

The dependency between these four propositional variables—A, C, H and S—is given by the Bayesian network in Figure 5.4. S, the success of theory T, only depends on the empirical adequacy of T, that is, on H. The probability of H depends on the number of distinct alternatives that are also consistent with the current data, etc. Finally, C, the probability of observing substantial theory change, depends on S and A: the empirical success of T and the number of available alternatives. To rule out preservation of a theory by a of series degenerative accommodating moves, the variable C should be evaluated over a longer period (e.g., 30–50 years).

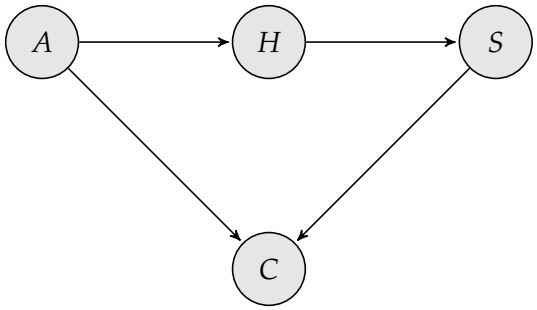

Figure 5.4: The Bayesian Network representation of the relation between variables A (the number of alternatives to T), H (empirical adequacy of theory T), S (success of T) and C (major theory change).

We now define a number of real-valued variables in order to facilitate calculations:

- Denote by $a_j := p(\mathrm{A_j})$ the probability that there are exactly j alternatives to T that satisfy the theoretical constraints $\mathcal{C}$, are consistent with current data $\mathcal{D}$ and give definite predictions for future experiments $\mathcal{E}$, etc.
- Denote by $h_j := p(\mathrm{H}|\mathrm{A_j})$ the probability that T is empirically adequate if there are exactly j alternatives to T.
- As before, denote by $s := p(\mathrm{S}|\mathrm{H})$ and $s' := p(\mathrm{S}|\neg\mathrm{H})$ the probability that T is successful if it is (not) empirically adequate.
- Denote by $c_j := p(\neg\mathrm{C}|\mathrm{A_j},\mathrm{S})$ the probability that no substantial theory change occurs if T is successful and there are exactly j alternatives to T.

Suppose that we now observe ¬C (no substantial theory change has occurred in the last decades) and S (theory T is successful). The Bayesian network structure allows for a simple calculation of the posterior probability of H.

Proposition 5.1 *The posterior probability of* H *given* ¬C *and* S *is given by*

$$p(\mathrm{H}|\neg\mathrm{CS}) = \frac{\sum_{j=0}^{\infty} a_j\, c_j\, s\, h_j}{\sum_{j=0}^{\infty} a_j\, c_j\, (s\, h_j + s'(1-h_j))} \tag{5.3}$$

We now make some assumptions on the values of these quantities.

B0 The variables A, C, H and S satisfy the (conditional) independencies in the Bayesian Network structure of Figure 5.4.

B1 If T is empirically adequate then it will be successful in the long run: $p(\mathrm{S}|\mathrm{H}) = 1$.

B2 The empirical adequacy of T is no more or less probable than the empirical adequacy of an alternative which satisfies the same set of theoretical and empirical constraints: $h_j := p(\mathrm{H}|\mathrm{A_j}) = 1/(j+1)$. In other words, there is no "actualist bias" in favor of T.

B3 The more satisfactory alternatives exist, the less likely is an extended period of theoretical stability. In other words, $c_j := p(\neg\mathrm{C}|\mathrm{A_j})$ is a decreasing function of j. For convenience, we choose $c_j := 1/(j+1)$. (This particular assignment will be relaxed later on.)

B4 Assume that T is our currently best theory and we happen to find a satisfactory alternative T′. Then, the probability of finding another alternative T″ is the same as the probability of finding T′ in the first place. Formally:

$$p(A > j | \bigvee_{k=j}^{\infty} \mathrm{A_k}) = p(A > j+1 | \bigvee_{k=j+1}^{\infty} \mathrm{A_k}) \; \forall j \geq 0. \tag{5.4}$$

In other words, Equation (5.4) expresses the idea that finding an alternative does not, in itself, raise or lower the probability of finding another alternative.

Note that B0–B4 are equally plausible for the realist and the anti-realist. In other words, no realist bias has been incorporated into the assumptions. We can now show the following proposition (proof in the appendix):

Proposition 5.2 *From Equation* (5.4) *it follows that* $a_j := a_0 \cdot (1 - a_0)^j$.

Together with this proposition, B0–B4 allow us to rewrite Equation (5.12) as follows:

$$p(\mathrm{H}|\neg\mathrm{CS}) = \frac{\sum_{j=0}^{\infty}(1-a_0)^j \frac{1}{(j+1)^2}}{\sum_{j=0}^{\infty}(1-a_0)^j \frac{1-s'j}{(j+1)^2}} \tag{5.5}$$

With the help of this formula, we can now rehearse the NMA once more and determine its scope, that is, those parameter values where $p(\mathrm{H}|\neg\mathrm{CS}) > 1/2$. The two relevant parameters are a_0, the prior probability that there are no satisfactory alternatives to T, and s', the probability that T is successful although not empirically adequate. Since an analytical solution of Equation (5.5) is not feasible, we conduct a numerical analysis. Results are plotted in Figure 5.5.

These results are very different from the ones in the previous section. With the hyperplane $z = 0.5$ dividing the graph into a region where T may be accepted and a region where this is not the case, we see that the scope of the NMA has increased substantially compared to Figure 5.3. For instance, $a_0 = p(\mathrm{H}) > 0.1$ suffices for a posterior probability greater than $1/2$, almost regardless of the value of s'. This is a striking difference to the previous analysis where way more optimistic values had to be assumed in order to make the NMA work.

So far, the analysis has been conducted in terms of absolute confirmation, that is, the posterior probability of H. We now complement it

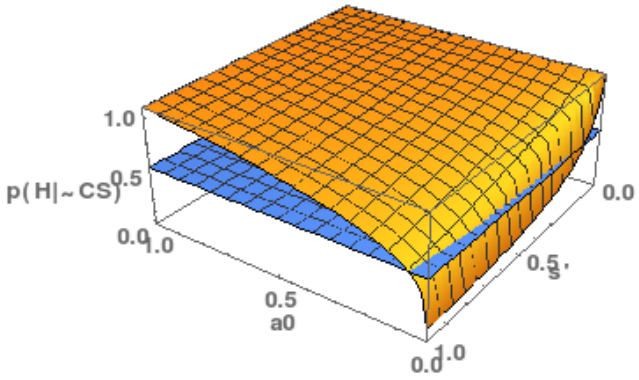

Figure 5.5: The scope of the No Miracles Argument in the revised formulation. The posterior probability of H, $p(\mathrm{H}|\neg\mathrm{CS})$, is plotted as a function of (1) the prior probability that T is empirically adequate (a_0); (2) the probability that T is successful if T is false ($s' = p(\mathrm{S}|\neg\mathrm{H})$). The hyperplane $z = 1/2$ is inserted in order to show for which parameter values $p(\mathrm{H}|\neg\mathrm{CS})$ is greater than $1/2$.

by an analysis in terms of confirmation as increase in firmness. That is, we calculate the evidential support that ¬CS confers on H. We use the log-likelihood measure $l(\mathrm{H}, \mathrm{E}) = \log_2 p(\mathrm{E}|\mathrm{H})/p(\mathrm{E}|\neg\mathrm{H})$ which has a good reputation in confirmation theory ($\rightarrow$ Variation 2) and a firm standing in scientific practice (e.g., Royall, 1997; Good, 2009). Also, it is a confirmation measure that describes the discriminative power of the evidence with respect to the realist and the anti-realist hypothesis and that is relatively insensitive to prior probabilities. The necessary calculations can be found in the final section of this variation.

In Figure 5.6, we have plotted the degree of confirmation as a function of the value of s', for three different values of a_0, namely 0.01, 0.05 and 0.1. As visible from the graph, the (logarithmic) degree of confirmation is substantial for all three cases, even for large values of s'. In particular, it is robust vis-à-vis the values of a_0 and s' and able to withstand the anti-realist argument that plagued the original version of the NMA. Note that if s' is small, as it will often be the case in practice, the logarithmic (!) degree of confirmation comes close to 10, which corresponds to a likelihood ratio of more than 1.000! And even if an anti-realist insists that $s' \approx .2$—not a very plausible assumption—, the likelihood ratio hovers in the range between 15 and 30. This finding accounts for the realist intuition that the stability of scientific theories over time, together with their empirical success, is

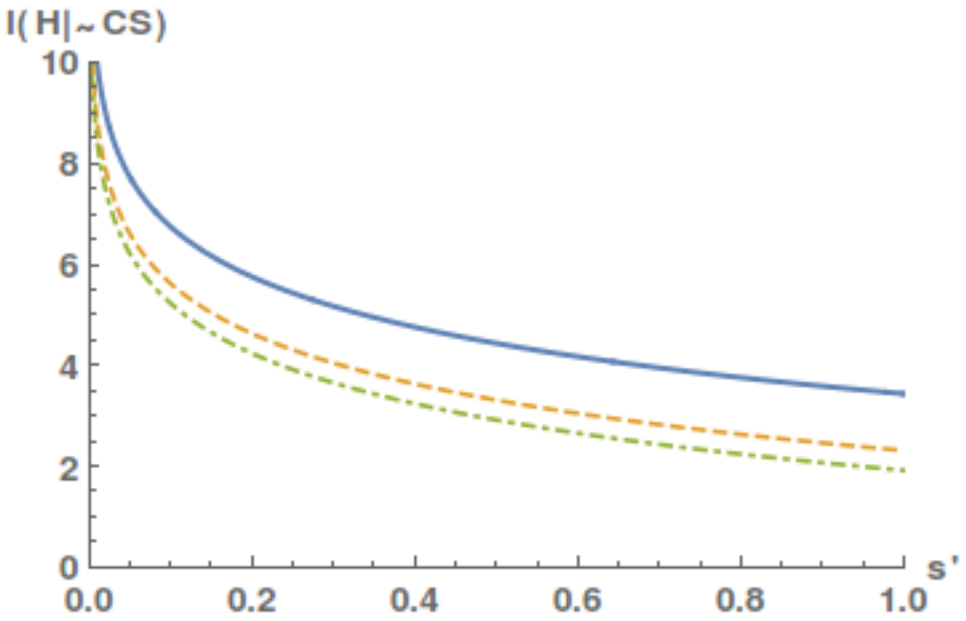

Figure 5.6: The degree of confirmation $l(\mathrm{H}, \neg\mathrm{CS}) = \log_2 p(\neg\mathrm{CS}|\mathrm{H})/p(\neg\mathrm{CS}|\neg\mathrm{H})$, for three different values of a_0. Full line: $a_0 = 0.01$. Dashed line: $a_0 = 0.05$. Dot-dashed line: $a_0 = 0.1$

strong evidence for their empirical adequacy.

Finally, we relax our assumptions B0–B4. Qualitatively, our results do not change if we replace B1 with the more cautious formulation $p(\mathrm{S}|\mathrm{H}) = 1 - \epsilon$. More interesting is a robustness analysis regarding the explication of B3. Arguably, the function $c_j := p(\neg\mathrm{C}|\mathrm{A_j},\mathrm{S}) = 1/(j+1)$ suggests that scientists are quite ready to give up on their currently best theory in favor of a good alternative. But as many have philosophers and historians of science have argued (e.g., Kuhn, 1977b), scientists may be more conservative and continue to work in the standard framework, even if good alternatives exist. Therefore we also analyze a different choice of the c_j, namely $c_j := e^{-\frac{1}{2}\left(\frac{x}{\alpha}\right)^2}$, where c_j falls more gently in j. This choice can then be plugged into Equation (5.12), yielding values of $p(\mathrm{H}|\neg\mathrm{CS})$ that are different from the ones in Equation (5.5).

The corresponding graph of $p(\mathrm{H}|\neg\mathrm{CS})$, as a function of a_0 and s', is presented in Figure 5.7. We have set $\alpha = 4$, corresponding to a high degree of reluctance to reject the currently best theory. Yet, the results match those from Figure 5.5: the scope of the NMA is much larger than in the simple version of the probabilistic NMA. Hence, our findings seem to

be robust toward different choices of c_j.

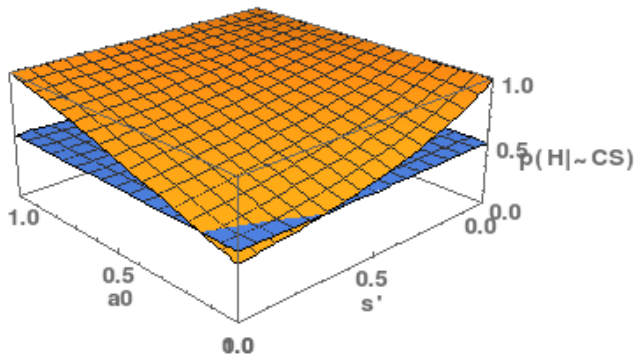

Figure 5.7: The scope of the No Miracles Argument in the revised formulation, with $c_j := e^{-\frac{1}{2}\left(\frac{x}{\alpha}\right)^2}$. The posterior probability of H, $p(\mathrm{H}|\neg\mathrm{CS})$, is plotted as a function of a_0 and s', like in Figure 5.5, and contrasted with the hyperplane $z = 1/2$.

All in all, our model shows that a probabilistic NMA need not be doomed. Its validity depends crucially on the disciplinary context where it operates in. What are our expectations regarding the invention of satisfactory alternatives to T? Has the discipline been in a long period of theoretical stability? And so on. Of course, our model makes simplifying assumptions, but unlike the assumptions in the original model, they do not carry a realist bias. This allows for a more nuanced and context-sensitive assessment of realist argument. The first step of the NMA is valid when theories are stable and the discipline allows for few potential explanations of observed phenomena. Anti-realist objections are supported by case studies where scientific theories have been volatile or one of our assumptions B0–B4 is implausible. The probabilistic reconstruction of the NMA can thus explain and guide the strategies that realists and anti-realists pursue when defending their positions.

We would like to stress that the context-sensitivity of the NMA is not a vice, but a virtue. It explains why realists and anti-realists often talk past each other, and it sketches a fruitful research program for future case studies. In particular, more research is needed into which areas of science have

been theoretically stable, and whether the kind of stability that Fahrbach cites is genuine or based on a superficial continuity that hides substantial meaning changes. We now proceed to another variation of the NMA: the frequency-based No Miracles Argument.

The Frequency-Based No Miracles Argument

In the above analysis, the empirical adequacy of a particular theory T explained why T is predictively successful. We shall call a NMA that tries to derive the empirical adequacy of theory T from its predictive success an **individual-theory-based NMA**. However, there is another way of conceptualizing the NMA. Following this second understanding, what is to be explained by the realist conjecture is not the empirical adequacy of a particular theory (e.g., the Standard Model of modern physics), but the *tendency* of theories in mature science to be empirically adequate. In this version, the NMA primarily relies on observed characteristics of science as a whole, or of a specific segment of science. Theories that are part of that segment, such as theories that are part of mature science or that are part of a specific mature research field, are expected to have a high rate of being empirically adequate. We will call a NMA based on the frequency of predictive success **frequency-based NMA**. This is actually the type of NMA that Hilary Putnam put forward in his famous first formulation of NMA:

> The positive argument for realism is that it is the only philosophy that does not make the success of science a miracle. That terms in mature science *typically* refer [...], that the theories in a mature science are *typically* true, that the same term can refer to the same thing even when it occurs in different theories—these statements are viewed by the scientific realist not as necessary truths but as the only scientific explanation of the *success of science* and hence as part of any adequate scientific description of science and its relations to its objects. (Putnam, 1975, our emphasis)

Note that Putnam speaks of the success of science rather than of the success of an individual scientific theory. He clearly understands the success of science as a general and observable phenomenon. Since he obviously

would not want to say that each and every scientific theory is always predictively successful, he asserts that we find a high success rate of scientific theories based on our observations of the history of (mature) science. He then infers from the success of science that mature scientific theories are typically approximately true, or at least empirically adequate.

Another early main exponent of the NMA, Richard Boyd (1981, 1983, 1984), is committed to frequency-based NMA as well. Boyd emphasizes that only what he calls the "predictive reliability of well-confirmed scientific theories" and the "reliability of scientific methodology in identifying predictively reliable theories" provides the basis for the NMA.

It is important to note that the frequency-based NMA is not adequately captured by Howson's reconstruction. An accurate Bayesian reconstruction of the frequency-based NMA must include updating under the observation of scientific successes and failures in the entire research field. This shall be done now. To start, we specify a scientific discipline or research field. We count all n_E theories in the field that have been empirically tested and determine the number n_S of theories that were predictively successful. We can thus state the following observation O:

O n_S out of n_E theories in the research field are predictively successful.

Let us assume that we are confronted with a new and so far empirically untested theory T in that research field. We want to extract the probability $p(\mathrm{S}|\mathrm{O})$ for the predictive success S of T given observation O. In order not to beg the question by assuming predictive success a priori, we assume a prior probability $p(\mathrm{S}) = \epsilon$ where ϵ can be an arbitrarily small number.

We then assume that each new theory that appears in the research field can be treated as a random pick with respect to predictive success. That is, we assume that there is a certain overall rate of predictively successful theories in the research field and, in the absence of further knowledge, the success chances of a new theory should be estimated according to our best estimate r of that success rate:

$$r = p(\mathrm{S}|\mathrm{O}) \tag{5.6}$$

r will be based on observation O. The most straightforward assessment of r is to use the long-run information about the frequency of success in a discipline and to identify r with

$$r_{freq} = \frac{n_S}{n_E}\,. \tag{5.7}$$

Moreover, we make two assumptions similar to the individual-theory-based NMA:

$\mathcal{A}_1^O$: $p(\mathrm{S}|\mathrm{H},\mathrm{O})$ is quite large.

$\mathcal{A}_2^O$: $p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}) < k \ll 1$

Note that realists assume that the empirical adequacy of T is the dominating element in explaining the theory's predictive success. If that is so, then S is roughly conditionally independent of O given H (=theory T is empirically adequate) and we have $p(\mathrm{S}|\mathrm{H},\mathrm{O}) \approx p(\mathrm{S}|\mathrm{H})$ and $p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}) \approx p(\mathrm{S}|\neg\mathrm{H})$. The conditions $\mathcal{A}_1^O$ and $\mathcal{A}_2^O$ then roughly correspond to $\mathcal{A}_1$ and $\mathcal{A}_2$.

We now come to the crucial point of our analysis: **accounting for observation O blocks the base-rate fallacy**. The base-rate fallacy in individual theory-based NMA consisted in disregarding the possibility of arbitrarily small priors $p(\mathrm{H})$. In the frequency-based NMA, however, the crucial probability is $p(\mathrm{H}|\mathrm{O})$ rather than $p(\mathrm{H})$. Updating the probability of S on observation O has an impact on $p(\mathrm{H}|\mathrm{O})$.

Proposition 5.3 *If conditions $\mathcal{A}_1^O$ and $\mathcal{A}_2^O$ are satisfied, then the following inequality holds:*

$$p(\mathrm{H}|\mathrm{O}) > r - k\,, \tag{5.8}$$

The frequency-based NMA takes it as a premise, as an observed fact about (parts of) mature science, that n_S/n_E is fairly large. Hence, r is fairly large and we can infer from Equation (5.8) that $p(\mathrm{H}|\mathrm{O})$ is also substantially greater than zero. Thus, the base-rate fallacy is avoided. Note that the first and crucial inference in the frequency-based NMA is made before accounting for the predictive success of T itself: it relates $p(\mathrm{S}|\mathrm{O})$ to $p(\mathrm{H}|\mathrm{O})$ by the law of total probability.

However, the final strength of NMA is expressed by the value $p(\mathrm{H}|\mathrm{S},\mathrm{O})$. In other words, the realist has to show that

$$p(\mathrm{H}|\mathrm{S},\mathrm{O}) > K, \tag{5.9}$$

where K is, as before, some reasonably high probability value. $K = 1/2$ may be viewed as a plausible condition for taking the NMA seriously. How does a condition on $p(\mathrm{H}|\mathrm{S},\mathrm{O})$ translate into a condition on $p(\mathrm{S}|\mathrm{O})$

and therefore on the observed success frequency r? First we observe the following result:

Proposition 5.4

$$p(\mathrm{H}|\mathrm{S},\mathrm{O}) = \frac{p(\mathrm{S}|\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{O})} \cdot \left(\frac{p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{H},\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \right)$$

Then, we observe that $p(\mathrm{H}|\mathrm{S},\mathrm{O})$ is decreasing in $p(\mathrm{S}|\mathrm{H},\mathrm{O})$ if $p(\mathrm{S}|\mathrm{O})$ and $p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})$ are held fixed. It thus makes sense to focus on the case where it is most difficult for the realist to make the frequency-based NMA work, namely the case $p(\mathrm{S}|\mathrm{H},\mathrm{O}) = 1$. Then, the following theorem describes a sufficient condition for $p(\mathrm{H}|\mathrm{S},\mathrm{O})$ to exceed the threshold K:

Theorem 5.1 *If conditions $\mathcal{A}_1^O$ and $\mathcal{A}_2^O$ are satisfied, if $p(\mathrm{S}|\mathrm{H},\mathrm{O}) = 1$, and the inequality below holds:*

$$p(\mathrm{S}|\mathrm{O}) > \frac{p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{1 - K + K \cdot p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \tag{5.10}$$

then inequality (5.9) *is satisfied as well: $p(\mathrm{H}|\mathrm{S},\mathrm{O}) > \mathrm{K}$.*

For $K = 1/2$, and using the base rate estimate $p(S|O) = n_S/n_E$, this is the case if and only if

$$n_S/n_E > 2\,\frac{p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{1 + p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}. \tag{5.11}$$

In particular, $2p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}) \approx n_S/n_E$ is sufficient for satisfying Equations (5.11) and (5.9). Thus, we don't need an impressively high rate of predictive success for a significant argument in favor of scientific realism. A defender of the NMA can avoid the base-rate fallacy by taking a global perspective on the success of science. We have also argue that this perspective is more faithful to the intentions of those who put the NMA forward in the first place—namely Hilary Putnam and Richard Boyd.

All this does not imply that the NMA is valid. A supporter of the frequency-based NMA must specify on which grounds she takes a high frequency of predictive success in science to be borne out by the data. And she must undertake the difficult task of justifying assumptions $\mathcal{A}_1^{\mathbf{O}}$ and (especially) $\mathcal{A}_2^{\mathbf{O}}$. But we have shown, contra Howson, that the NMA still has a fighting chance. For philosophers like ourselves, who are not

committed to a particular position in the debate between realists and anti-realists, this probabilistic reconstruction of the NMA offers the chance to understand the argumentative mechanics behind the realist intuition, to better appreciate the context-dependency of the NMA, and to critically evaluate the merits of realist and anti-realist standpoints.

Discussion

This variation has investigated scope and limits of the No Miracles Argument (NMA) when formalized as a probabilistic argument aiming at the empirical adequacy of a particular theory T. In the simple probabilistic model of the NMA, we have confirmed the diagnosis that it does not hold water as an objective argument (Howson, 2000, 2013): too much depends on the choice of the prior probability $p(\mathrm{H})$, assuming what is supposed to be shown. We have supported this diagnosis by a detailed analysis of the probabilistic mechanics of NMA.

Then, we have shown two possible ways out of the dilemma. First, we have investigated how the stability of a scientific theory over time may impact the probability that it is empirically adequate. We have shown that such observations can greatly increase the range of prior probabilities for which the NMA leads to acceptance of T. Second, we have demonstrated that Howson's objection can be mitigated if the base rate of predictively successful theories in a specific discipline is taken into account. This is also faithful to the line of argument of those scientific realists who think of the NMA as a global argument based on the high frequency of successful theories in science. In both cases, we have supplemented the classical NMA reasoning with novel and distinct kinds of evidence that can be embedded into a Bayesian framework. Using our models, the realist thesis (or at least the part leading up to empirical adequacy) can be defended with much weaker assumptions than in the simple version of the NMA.

Finally, we should mention the No Alternatives Argument (NAA): the claim that the continuous failure to find satisfactory alternatives to a theory provides evidence for it. In the previous variation, we have shown that under plausible assumptions, this observation indeed raises the probability that theory T is empirically adequate. The NAA can also be seen as a variation of the NMA: the empirical adequacy of T is the only explanation for why scientists have not yet found an alternative. Yet, we have also seen

that while this reasoning is valid in principle, it is usually not sufficient to push the probability of H (the hypothesis that T is empirically adequate) beyond a critical threshold—at least not without additional assumptions.

This analogy brings us to a project for future research. It would be exciting to investigate the parallel between the NAA and the NMA in greater detail, and to proceed to a general analysis of argument patterns that take non-empirical evidence (in our technical sense of the word) as their premise. Second, we have seen that the probabilistic versions of the NMA are highly sensitive to $p(\mathrm{S}|\neg\mathrm{H})$ and related quantities that express the probability of empirical success if T is not empirically adequate. The evaluation of this quantity is itself a point of contention between realists and anti-realists: after all, anti-realists often stress the explanatory success of false models whereas realists are usually committed to the thesis that only true models yield stable empirical success. An investigation of this question, supported by case studies, would be highly useful. Third, we need to examine whether theories have really been more stable during the 20th century than before since this is a crucial premise in the probabilistic individual-theory-based NMA. For tackling this research question, a combination of case studies and scientometric analysis (e.g., along the lines of Herfeld and Doehne, 2016) strikes us as a promising approach. Fourth, it would be good to explore whether the probabilistic NMA can be extended into an argument for the full realist position, that is, the view that T is true (and not only empirically adequate). At present, we do not see an obvious way of doing so—it seems that the argument would just run into the underdetermination problem—but we invite realists to take our formalism and to apply it to a full-fledged defense of the realist view.

It is noteworthy that all formalizations of the NMA stressed the scientific track record in the particular discipline to which T belongs. Instead of reading NMA as a "wholesale argument" for scientific realism that is valid across the board, we should understand it as a "retail argument" (Magnus and Callender, 2004), that is, as an argument that may be strong for some scientific theories and weak for others. While context-sensitive assumptions are required in our arguments, their relative weakness leaves open the possibility of a coherent, non-circular realist position in philosophy of science. It also makes the realist argument more sensitive to scientific practice which is, ultimately, something that all formal reconstructions of scientific reasoning should aim at. Together with the preceding variation,

this variation has shown how Bayesian reasoning can model and vindicate argument patterns that support the realist hypothesis, and how Bayesian models can contribute to a fair assessment of the debate between realists and anti-realists.

Proofs of the Theorems

Proof of Proposition 5.1:

$$\begin{aligned} p(\neg \mathrm{CSH}) &= \sum_A p(\mathrm{A}) p(\neg \mathrm{C}|\mathrm{AS}) p(\mathrm{S}|\mathrm{H}) p(\mathrm{H}|\mathrm{A}) \\ &= \sum_{j=0}^{\infty} a_j \, c_j \, s \, h_j \end{aligned}$$

$$\begin{aligned} p(\neg \mathrm{CS}) &= \sum_{A,H} p(\mathrm{A}) p(\neg \mathrm{C}|\mathrm{AS}) p(\mathrm{S}|\mathrm{H}) p(\mathrm{H}|\mathrm{A}) \\ &= \sum_A p(\mathrm{A}) p(\neg \mathrm{C}|\mathrm{AS}) p(\mathrm{S}|\mathrm{H}) p(\mathrm{H}|\mathrm{A}) + \sum_A p(\mathrm{A}) p(\neg \mathrm{C}|\mathrm{AS}) p(\mathrm{S}|\neg \mathrm{H}) p(\neg \mathrm{H}|\mathrm{A}) \\ &= \sum_{j=0}^{\infty} a_j \, c_j \, (s \, h_j + s'(1 - h_j)) \end{aligned}$$

With the help of Bayes' Theorem, these equations allows us to calculate the posterior probability of H conditional on C and S:

$$\begin{aligned} p(\mathrm{H}|\neg \mathrm{CS}) &= \frac{p(\neg \mathrm{CSH})}{p(\neg \mathrm{CS})} \\ &= \frac{\sum_{j=0}^{\infty} a_j \, c_j \, s \, h_j}{\sum_{j=0}^{\infty} a_j \, c_j \, (s \, h_j + s'(1 - h_j))} \end{aligned}$$ □

Proof of Proposition 5.2: Assumption B4 is equivalent to the following claim:

$$p(\mathrm{A_j}| \bigvee_{\mathrm{k=j}}^{\infty} (\mathrm{A_k})) = p(\mathrm{A_{j+1}}| \bigvee_{\mathrm{k=j+1}}^{\infty} (\mathrm{A_k})) \; \forall j \geq 0.$$

which entails that for all $j \geq 0$, we have

$$\frac{p(\mathrm{A_j})}{p(\bigvee_{\mathrm{k=j}}^{\infty}(\mathrm{A_k}))} = \frac{p(\mathrm{A_{j+1}})}{p(\bigvee_{\mathrm{k=j+1}}^{\infty}(\mathrm{A_k}))}.$$

This implies in turn

$$\begin{aligned} p(\mathrm{A_{j+1}}) &= p(\mathrm{A_j}) \, \frac{p(\bigvee_{\mathrm{k=j+1}}^{\infty}(\mathrm{A_k}))}{p(\bigvee_{\mathrm{k=j}}^{\infty}(\mathrm{A_k}))} \\ &= p(\mathrm{A_j}) \, \frac{1 - \sum_{k=0}^{j} p(\mathrm{A_k})}{1 - \sum_{k=0}^{j-1} p(\mathrm{A_k})} \end{aligned}$$

By a simple induction proof, we can now show

$$p(\mathrm{A_n}) = p(\mathrm{A_0}) \left(1 - \sum_{k=0}^{n-1} p(\mathrm{A_k})\right) \tag{5.12}$$

For $n = 1$, equation (5.12) follows immediately. Assuming that it holds for level n, we then obtain

$$\begin{aligned} p(\mathrm{A_{n+1}}) &= p(\mathrm{A_n}) \frac{1 - \sum_{k=0}^{n} p(\mathrm{A_k})}{1 - \sum_{k=0}^{n-1} p(\mathrm{A_k})} \\ &= p(\mathrm{A_0}) \left(1 - \sum_{k=0}^{n-1} p(\mathrm{A_k})\right) \frac{1 - \sum_{k=0}^{n} p(\mathrm{A_k})}{1 - \sum_{k=0}^{n-1} p(\mathrm{A_k})} \\ &= p(\mathrm{A_0}) \left(1 - \sum_{k=0}^{n} p(\mathrm{A_k})\right) \end{aligned}$$

where we have used the inductive premise in the second step. Finally, we use straight induction once more to show that

$$p(\mathrm{A_n}) = p(\mathrm{A_0})(1 - p(\mathrm{A_0}))^n \tag{5.13}$$

where the case $n = 0$ is trivial and the inductive step $n \to n + 1$ is proven as follows:

$$\begin{aligned} p(\mathrm{A_{n+1}}) &= p(\mathrm{A_0}) \left(1 - \sum_{k=0}^{n} p(\mathrm{A_k})\right) \\ &= p(\mathrm{A_0}) \left(1 - \sum_{k=0}^{n} p(\mathrm{A_0}) \left(1 - p(\mathrm{A_0})\right)^k\right) \\ &= p(\mathrm{A_0}) \left(1 - p(\mathrm{A_0}) \frac{1 - (1 - p(\mathrm{A_0}))^{n+1}}{1 - (1 - p(\mathrm{A_0}))}\right) \\ &= p(\mathrm{A_0})(1 - (1 - (1 - p(\mathrm{A_0}))^{n+1})) \\ &= p(\mathrm{A_0})(1 - p(\mathrm{A_0}))^{n+1} \end{aligned}$$

In the second line, we have applied the inductive premise to $p(\mathrm{A_k})$, and in the third line, we have used the well-known formula for the geometric series:

$$\sum_{k=0}^{n} q^k = \frac{1 - q^{n+1}}{1 - q}$$

This shows (5.13) and completes the proof of the proposition. □

Calculation of the degree of confirmation (Figure 5.6):

$$\begin{aligned} p(\neg\mathrm{CS}|\mathrm{H}) = \frac{p(\neg\mathrm{CSH})}{p(\mathrm{H})} &= \frac{\sum_A p(\mathrm{A})p(\neg\mathrm{C}|\mathrm{AS})p(\mathrm{S}|\mathrm{H})p(\mathrm{H}|\mathrm{A})}{\sum_A p(\mathrm{A})\, p(\mathrm{H}|\mathrm{A})} \\ &= \frac{\sum_{j=0}^{\infty} a_j\, c_j\, s\, h_j}{\sum_{j=0}^{\infty} a_j\, h_j} \\ &= \frac{\sum_{j=0}^{\infty}(1-a_0)^j \frac{1}{(1+j)^2}}{\sum_{j=0}^{\infty}(1-a_0)^j \frac{1}{1+j}} \end{aligned}$$

$$\begin{aligned} p(\neg\mathrm{CS}|\neg\mathrm{H}) = \frac{p(\neg\mathrm{CS}\neg\mathrm{H})}{p(\neg\mathrm{H})} &= \frac{\sum_A p(\mathrm{A})p(\neg\mathrm{C}|\mathrm{AS})p(\mathrm{S}|\neg\mathrm{H})p(\neg\mathrm{H}|\mathrm{A})}{\sum_A p(\mathrm{A})\, p(\neg\mathrm{H}|\mathrm{A})} \\ &= \frac{\sum_{j=0}^{\infty} a_j\, c_j\, s'\, (1-h_j)}{\sum_{j=0}^{\infty} a_j\, (1-h_j)} \\ &= \frac{\sum_{j=0}^{\infty}(1-a_0)^j \frac{s'j}{(1+j)^2}}{\sum_{j=0}^{\infty}(1-a_0)^j \frac{j}{1+j}} \end{aligned}$$

Proof of Proposition 5.3: We first apply the law of total probability and obtain

$$\begin{aligned} p(\mathrm{S}|\mathrm{O}) &= p(\mathrm{S}|\mathrm{H},\mathrm{O})p(\mathrm{H}|\mathrm{O}) + p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})p(\neg\mathrm{H},\mathrm{O}) \\ &= p(\mathrm{S}|\mathrm{H},\mathrm{O})p(\mathrm{H}|\mathrm{O}) + p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}) \cdot (1 - p(\mathrm{H},\mathrm{O})) \\ &= p(\mathrm{H},\mathrm{O}) \cdot (p(\mathrm{S}|\mathrm{H},\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})) + p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}) \end{aligned}$$

Hence,

$$p(\mathrm{H}|\mathrm{O}) = \frac{p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{H},\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}\,. \tag{5.14}$$

From Equation (5.14), we derive

$$\begin{aligned} p(\mathrm{H}|\mathrm{O}) &= \frac{p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{H},\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \\ &\geq \frac{p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} && (5.15) \\ &\geq p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}) && (5.16) \\ &> p(\mathrm{S}|\mathrm{O}) - k && (5.17) \end{aligned}$$

Here Equation (5.16) follows from assumptions $\mathcal{A}_1^O$ and $\mathcal{A}_2^O$ and Equation (5.17) follows from assumption $\mathcal{A}_2^O$. This leads directly to the desired result:

$$p(\mathrm{H}|\mathrm{O}) > p(\mathrm{S}|\mathrm{O}) - k = R - k \qquad \square$$

Proof of Proposition 5.4: From Bayes' Theorem and Equation (5.14), we infer

$$\begin{aligned} p(\mathrm{H}|\mathrm{S},\mathrm{O}) &= \frac{p(\mathrm{H}|\mathrm{O})p(\mathrm{S}|\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{O})} \\ &= \frac{p(\mathrm{S}|\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{O})} \cdot \left(\frac{p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{H},\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \right) \square \end{aligned} \tag{5.18}$$

Proof of Theorem 5.1: We have assumed that $p(\mathrm{S}|\mathrm{H},\mathrm{O}) = 1$. Inserting this equality into (5.18) gives us

$$p(\mathrm{H}|\mathrm{S},\mathrm{O}) = \frac{1}{p(\mathrm{S}|\mathrm{O})} \cdot \left(\frac{p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \right).$$

Hence, we can write the condition $p(\mathrm{H}|\mathrm{S},\mathrm{O}) > K$ as

$$\frac{1}{p(\mathrm{S}|\mathrm{O})} \cdot \left(\frac{p(\mathrm{S},\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \right) > K,$$

Rewriting this inequality a couple of times, we obtain

$$\begin{aligned} \frac{1}{p(\mathrm{S}|\mathrm{O})} \cdot \left(\frac{p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \right) &> K \\ \frac{1}{p(\mathrm{S}|\mathrm{O})} \cdot (p(\mathrm{S}|\mathrm{O}) - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})) &> K\,(1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})) \\ 1 - \frac{p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{O})} &> K\,(1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})) \\ 1 - K\,(1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})) &> \frac{p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{p(\mathrm{S}|\mathrm{O})} \\ p(\mathrm{S}|\mathrm{O}) \cdot (1 - K(1 - p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}))) &> p(\mathrm{S}|\neg\mathrm{H},\mathrm{O}) \\ p(\mathrm{S}|\mathrm{O}) &> \frac{p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})}{1 - K + K \cdot p(\mathrm{S}|\neg\mathrm{H},\mathrm{O})} \end{aligned}$$

This was exactly one of the assumptions of the theorem. Thus we can infer that $p(\mathrm{H}|\mathrm{S},\mathrm{O}) > K$. $\square$

Variation 6: Causal Effect

From Aristotle to the 21st century, causation is usually treated as a qualitative, all-or-nothing concept. Either C is a cause of E or it is not. However, sometimes we have to make more nuanced causal judgments that involve a quantitative dimension: C is a more effective cause of E than C′, the causal effect of C on E is twice as high as the effect of C′, etc. This is especially important for purposes of prediction and evaluating experimental findings (e.g., Rubin, 1974; Rosenbaum and Rubin, 1983; Pearl, 2001). For instance, a regulatory medical body like the US Food and Drug Administration (FDA) or the European Medical Agency (EMA) only admits a new drug to the market if there is a substantial causal effect on recovery rates. The effect of immigration on the crime rate shapes political views, creates prejudices and affects concrete policy decisions. The effect of the driver's speeding on a traffic accident influences the amount of compensation that the victim may receive. All these judgments tap onto the concept of *causal effect*, or equivalently, causal strength or graded causation.

While a huge amount of literature has been devoted to the qualitative question "When is C a cause of E?" (e.g., Hume, 1739; Suppes, 1970; Lewis, 1973; Mackie, 1974; Woodward, 2003), and the comparative question "Is C or C′ a more effective cause of E?" starts to get explored as well (e.g., Chockler and Halpern, 2004; Halpern and Hitchcock, 2016), the quantitative question "What is the causal effect of C on E?" is relatively neglected, given the huge scope of actual and potential applications in science. There are proposals from different disciplines, such as psychology (Cheng, 1997), computer science (Pearl, 2000), statistics (Good, 1961a,b) and philosophy (Eells, 1991), but apart from a survey paper by Fitelson and Hitchcock (2011), no attempt is made at a unified theory of causal effect.

Measures of causal effect can differ substantially. Consider a clinical trial where the effect of a new drug for treating migraine is compared to a control group that receives the standard treatment. The effects are

Group/Outcome	Effect	No Effect	Total Number
Treatment	A=30	B=90	A+B=120
Control	C=15	D=105	C+D=120

Table 6.1: The result of a clinical trial where the efficacy of a new migraine treatment is compared to a control group. How should the causal effect of the treatment be quantified?

measured on a binary scale: did the pain diminish significantly or not? Suppose the results are described by Table 6.1. In the epidemiological literature, several measures have been proposed to measure the size of such an effect (Davies et al., 1998; Deeks, 1998; Sistrom and Garvan, 2004; King et al., 2012):

Relative Risk (RR) The ratio of the observed relative frequencies of an effect in both groups.

$$\mathrm{RR} = \frac{A/(A+B)}{C/(C+D)}$$

Odds Ratio (OR) The ratio of the odds for an effect in both groups.

$$\mathrm{OR} = \frac{A/B}{C/D}$$

Absolute Risk Reduction (ARR) The difference between the relative frequencies of an effect in both groups.

$$\mathrm{ARR} = \frac{A}{A+B} - \frac{C}{C+D}$$

To give an example with the numbers from Table 6.1: The relative risk would be RR $= 2$, meaning that the treatment halves the frequency of pain in the affected population. The result looks similar for the odds ratio OR $= 2.33$, but the absolute risk reduction ARR $= 0.125$ tells a less enthusiastic story: only for 12,5% of the affected population, the new treatment makes a difference. This prompts the question of which measure should be preferred, and for which reasons (e.g., Stegenga, 2015; Sprenger and Stegenga, 2016). Related questions pop up in the psychological literature on causal induction (e.g., Cheng, 1997; Sloman and Lagnado, 2015): how can we quantify the power of a cause to produce an effect?

The challenge for a philosophical theory of causal effect is to characterize these measures and to weigh the reasons for preferring one of them

over another. In this variation, we develop axiomatic foundations for measures of causal effect between binary variables (e.g., propositions). First, we defend the choice of a framework where different measures can be embedded and compared: causal Bayes nets. Second, we derive representation theorems for various measures of causal effect, that is, theorems that characterize a measure of causal in terms of a set of adequacy conditions. Third, we compare and discuss these measures with a view towards applications: Under which conditions are they invariant? What are beneficial and what are problematic properties? To the extent that the proposed adequacy conditions are found compelling, the technical results have normative implications for the choice of a measure of causal effect. Indeed, we will make a case for a particular measure as opposed to working with a plurality of causal effect measures (see also Sprenger, 2016c).

Our approach is methodologically innovative in transferring methods from probabilistic accounts of confirmation and explanatory power to a probabilistic theory of causal effect (e.g., Schupbach and Sprenger, 2011; Crupi and Tentori, 2012, 2013; Crupi et al., 2013). Thereby, we also create a bridge between different areas of philosophy of science and formal epistemology, and between different parts of this book (→ Variation 2 and 7).

The remainder is structured as follows: Section 6.1 motivates the choice of causal Bayes nets as a framework for explicating causal effect. Then we provide a set of general adequacy conditions in Section 6.2 which are complemented by more specific conditions in Sections 6.3-6.6. These sections also contain the representation theorems. Section 6.7 presents a brief application in medical science while Section 6.8 discusses future research questions and concludes. Section 6.9 contains the proofs of the theorems.

The Framework: Causal Bayes Nets

When we reason about causes, we often think that they make a *difference* to their effects. Causes which do not matter for the occurrence of an effect are no real causes. This thought has already been articulated by David Hume (1711–1776) in his famous description of two causally related objects: "if the first object had not been, the second never had existed" (Hume 1748/77). This line of reasoning is developed in the counterfactual and probabilistic accounts of causation (Lewis, 1973, 1979; Reichenbach,

1956; Suppes, 1970; Cartwright, 1979). It is also exemplified in many cases of scientific inference, such as Randomized Controlled Trials (RCT). There, we would like to assess the efficacy of a drug and we divide the trial participants in two groups: one that receives the new drug, and one that receives the standard treatment, or a placebo. The causal efficacy of the new drug is then a function of the divergence between the results in the treatment and the control group, like in the example of Table 6.1.

With an eye on scientific applications, it is also clear that the envisioned account of causal effect should be graded rather than an "all-or-nothing" account. Probability is a natural tool which can step in here. After all, medical drugs typically raise the probability of recovery; almost none of them makes recovery certain. The same can be said about psychological experiments or economic policy decisions: interventions increase the frequency of a particular response, but they do not guarantee it.

Probabilistic measures of causal effect thus nicely square with an account of causal relevance where causes raise the probability of the effects.

On the **probabilistic account of causal relevance**, C is a cause of E if and only if a change in the value of C (e.g., C instead of $\neg$C) changes the probability that E occurs. This theory captures the basic intuition that causes must make a difference to their effects without necessitating them. For example, not every regular smoker will eventually suffer from lung cancer, but still, we would like to classify smoking as a cause of lung cancer. The account of causation as probability-raising gets this example right, and many other examples of scientific reasoning as well.

However, the probabilistic account struggles to distinguish genuinely causal relevance from mere statistical correlation, e.g., in a case where both variables are correlated as a result of a common cause X. For example, a high crime rate in certain neighborhoods of Dutch cities was found to be correlated with a high percentage of migrants living there. However, the correlation did not indicate real causation. It could be explained away by a common cause: the low socio-economic status of these neighborhoods (Jensma, 2014). Conditional on the various levels of average income, there was no correlation between the crime rate and the number of migrants in a neighborhood. The naïve probabilistic account gets this wrong and still judges the number of migrants to be a cause of the high crime rate. To solve this problem, it has been suggested that the putative cause has to raise the probability of the effect in all background contexts

(e.g., Cartwright, 1979). However, such a condition is very strict: causal relations vanish as soon as there is a single background context where the probability is lowered. For purposes of control and intervention, such a requirement is often impractical. It has also been criticized as failing to match our intuitions in causal reasoning (Dupré, 1984; Eells, 1991).

The **interventionist account of causation** (Spirtes et al., 2000; Pearl, 2000) provides an alternative to a purely probabilistic model of causation. It is relative to the choice of a causal model M: a directed acyclical graph (DAG) G, consisting of a set of vertices (=variables) and directed edges, and a probability distribution over the variables in G. The edges represent how the effect of an intervention transfers to other variables; conversely, lack of a direct connection between variables codifies a conditional independence. On the interventionist account, C is a cause of E if and only if an *intervention* on C causes a change of value in E, or changes the probability that E takes a certain value (Woodward, 2012).

But what is an intervention? An ideal intervention forces a variable C to take a certain value while breaking the influence that other variables may have on it. Pearl's notation for such an intervention is $do(C = x)$. Formally, this means "lifting C from the influence of the old functional mechanism and placing it under the influence of a new mechanism that sets the value $C = x$ while keeping all other mechanisms undisturbed" (Pearl, 2000, 70, notation changed). See also Spirtes et al. (2000). Imagine that we would like to study the effects of classroom light on whether students are awake or asleep. The intensity of classroom light depends on the settings of the audiovisual system. However, we may press the light switch manually, overruling the system settings, and then study the effects of our intervention on the students (e.g., they wake up from deep sleep). This way, we directly intervene on the light intensity and break the functional dependency on the preconfigured system settings.

The interventionist account naturally distinguishes genuinely causal relations between C and E from relations where both variables are correlated as a result of a common cause X. See Figure 6.1. When one intervenes on C, the causal arrow leading from X to C is broken and no effect on E occurs. On the probabilistic account, it is less straightforward to express this difference since C and E are positively correlated (e.g., Eells, 1991). While the probabilistic account describes causation in terms of statistical relevance, comparing $p(\mathrm{E}|\mathrm{C})$ and $p(\mathrm{E}|\neg\mathrm{C})$, the interventionist account fo-

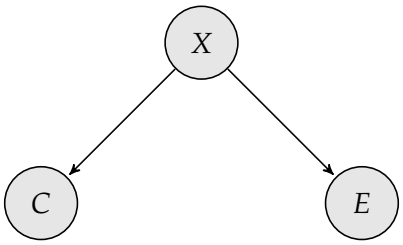

Figure 6.1: A typical common cause (conjunctive fork) structure. An intervention on C would disrupt the causal arrow leading into this variable from X and not have any effect on E.

cuses on probability of the effect conditional on an intervention on the cause, that is, $p(\mathrm{E}|do(\mathrm{C}))$.

In this variation, both perspectives are combined. We express causal effect as a function of $p(\mathrm{E}|do(\mathrm{C}))$ and $p(\mathrm{E}|do(\neg\mathrm{C}))$. This means that our account of causal strength supervenes on **causal models represented by causal Bayes nets**, already familiar from the introduction and the previous variations. In fact, we believe that causal Bayes nets are an intuitive and helpful tool in causal reasoning: first, it is easy to spot which interventions affect which variables; second, they resemble other tools for causal inference, such as neural networks or connectionist expert systems. Causal inference with Bayes nets, including measuring causal effect, can therefore be easily transferred to causal inference in the mind and brain sciences.

Table 6.2 translates various probabilistic measures of causal effect to the causal Bayes nets framework described above. The next sections will characterize and compare these measures.

Pearl (2000)	$\eta(\mathrm{C},\mathrm{E}) = p(\mathrm{E}\|do(\mathrm{C}))$
Suppes (1970)	$\eta(\mathrm{C},\mathrm{E}) = p(\mathrm{E}\|do(\mathrm{C})) - p(\mathrm{E})$
Eells (1991)	$\eta(\mathrm{C},\mathrm{E}) = p(\mathrm{E}\|do(\mathrm{C})) - p(\mathrm{E}\|do(\neg\mathrm{C}))$
"Galton" (covariation)	$\eta(\mathrm{C},\mathrm{E}) = 4p(do(\mathrm{C}))\,p(do(\neg\mathrm{C}))[p(\mathrm{E}\|do(\mathrm{C})) - p(\mathrm{E}\|do(\neg\mathrm{C}))]$
Lewis (1986)	$\eta(\mathrm{C},\mathrm{E}) = \frac{p(\mathrm{E}\|do(\mathrm{C}))}{p(\mathrm{E}\|do(\neg\mathrm{C}))}$
Cheng (1997)	$\eta(\mathrm{C},\mathrm{E}) = \frac{p(\mathrm{E}\|do(\mathrm{C})) - p(\mathrm{E}\|do(\neg\mathrm{C}))}{1 - p(\mathrm{E}\|do(\neg\mathrm{C}))}$
Good (1961a,b)	$\eta(\mathrm{C},\mathrm{E}) = \frac{1 - p(\mathrm{E}\|do(\neg\mathrm{C}))}{1 - p(\mathrm{E}\|do(\mathrm{C}))}$

Table 6.2: Some prominent measures of causal effect. We follow the labels of Fitelson and Hitchcock (2011).

General Adequacy Conditions

We aim at capturing the size of a causal effect between categorical, binary variables. To this end, let $\mathcal{L}$ be a propositional language with variables $C, E \in \mathcal{L}$. In agreement with the framework presented in the previous section, we demand that the causal effect between C and E depend on the causal model M, that is, a directed acyclical graph in which C and E are included, with a probability distribution over the variables. This leaves out external factors such as typicality, normative expectations and defaults, which are of theoretical significance and have been shown to affect causal judgments in experimental settings (Knobe and Fraser, 2008; Hitchcock and Knobe, 2009; Halpern and Hitchcock, 2016). While this implies that our model does not capture all aspects of judgments of causal effect, there are many applications (e.g., quantifying effect size in science) where it is desirable to eliminate normative considerations, and to derive causal effect from observed relative frequencies. Moreover, our approach quantifies causal effect with respect to a single background context, sidestepping a substantial discussion in the field of probabilistic causation (e.g., Cartwright, 1979; Dupré, 1984; Eells, 1991).

Formality For two binary variables C, E $\in \mathcal{L}$ and a causal model $M \in \mathcal{M}$, the causal effect of C on E, $\eta(\mathrm{C},\mathrm{E})$, is a continuous real-valued

function operating on a subset of $\mathcal{L}^2 \times \mathcal{M}$, namely the set

$$\mathcal{S} := \{\langle C, E, M\rangle \in \mathcal{L}^2 \times \mathcal{M} | \text{M contains } C \text{ and } E \text{ as variables}\} \quad (6.1)$$

In particular, the causal effect measure $\eta(\mathrm{C}, \mathrm{E})$ can be represented by a continuous function $f : [0,1]^3 \to \mathbb{R}$ such that

$$\eta(\mathrm{C}, \mathrm{E}) = f(p(\mathrm{C}), p(\mathrm{E}|do(\mathrm{C})), p(\mathrm{E}|do(\neg\mathrm{C})))$$

This means that $\eta(\mathrm{C}, \mathrm{E})$ can be expressed as a function of the base rate of the cause and the probability of E under the relevant interventions on the cause: $p(\mathrm{E}|do(\mathrm{C}))$ and $p(\mathrm{E}|do(\neg\mathrm{C}))$. This takes up the basic idea behind probabilistic relevance accounts of causation.

Formality is blind to mediator variables or multiple paths leading from C to E. Also this choice is conscious. The reason is that mediators are sometimes latent variables and not directly measurable. When we administer a medical drug C to cure headache E, there are numerous mediators in an appropriate causal model that includes C and E. However, the medical practitioner, who has to choose between different drugs, is mainly interested in the overall effect that C has on E (how often does the headache go away?), not in the details of the causal transmission within the human body. Therefore we keep the model simple and amalgamate the effects that C may have on E via different paths into one number (e.g., Dupré, 1984; Eells, 1991). This omission does not rule out a path-specific perspective. Investigating path-specific causal effect is an interesting topic and relevant for many cases of policy-making and attribution, but it stands orthogonal to our efforts. By appropriate conditionalization on other factors in the causal model, any measure of causal effect can be used for calculating path-specific effects and comparing them to the net effect (cf., Pearl, 2001).

While Formality sketches the ground on which the different measures compete, the following adequacy conditions describe how they should rank different cause/effect pairs. They describe different ways to think about measures of causal effect.

We start with the case of comparing two putative causes of an effect E. Suppose for example that we ask what is a stronger cause of headache (E): thinking hard about a difficult research problem (C_1) or going for a night of binge drinking (C_2)? In such cases, it is natural to answer that C_1 is more effective than C_2 if and only if C_1 makes E more expected than C_2.

In other words, a cause of an effect is stronger than another cause if it has a higher likelihood of producing the effect. Such a requirement is analogous to Final Probability Incrementality in Bayesian confirmation theory (Crupi, 2013; Crupi et al., 2013) that we have encountered in Variation 2.

Effect Production

$$\eta(\mathrm{C}_1, \mathrm{E}) > \eta(\mathrm{C}_2, \mathrm{E}) \qquad \text{if and only if} \qquad p(\mathrm{E}|do(\mathrm{C}_1)) > p(\mathrm{E}|do(\mathrm{C}_2))$$

It may be objected that Effect Production neglects the contrastive nature of measures of causal effect. They should measure the degree of causal dependence on C as opposed to ¬C, that is, the difference that an intervention on *C* makes for *E*. This aspect gets lost for measures of causal effect that satisfy Effect Production: what happens if $\neg\mathrm{C}_1$ or $\neg\mathrm{C}_2$ is the case does not matter for calculating causal effect.

If one follows this argument, one could replace Effect Production by anadequacy condition that focuses on the difference that two competing causes make for the target effect:

Difference-Making

$$\eta(\mathrm{C}_1, \mathrm{E}) > \eta(\mathrm{C}_2, \mathrm{E})$$

if and only if for a function $g : [0,1]^2 \to \mathbb{R}$ which is monotonically increasing in the first and monotonically decreasing in the second argument:

$$g(p(\mathrm{E}|do(\mathrm{C}_1)), p(\mathrm{E}|do(\neg\mathrm{C}_1))) > g(p(\mathrm{E}|do(\mathrm{C}_2)), p(\mathrm{E}|do(\neg\mathrm{C}_2)))$$

This condition demands in particular that the base rates of C_1 and C_2 should not matter for ranking their causal effect for an effect E. Instead, we only look at the degree to which intervening on C_1 and C_2 makes a difference for E. The monotonicity constraint on g expresses the intuitive condition that the more likely a cause C is to bring about an effect E, the more sizeable its causal effect, all other things being equal, and vice versa for ¬C.

Another important general property is a symmetry proposed by Fitelson and Hitchcock (2011): the degree to which C prevents E (=the degree to which C causes ¬E) is the negative of the degree to which C causes E:

Causation-Prevention Symmetry (CPS)

$$-\eta(\mathrm{C}, \mathrm{E}) = \eta(\mathrm{C}, \neg\mathrm{E})$$

In other words, when C is a strong preventive cause of E, it is just a weak cause of ¬E, and vice versa. CPS is more than a purely ordinal constraint: it assigns meaning to the exact numbers yielded by a measure of causal effect. Evidently, only measures of causal effect which take both positive and negative values can satisfy CPS, and we will often rescale candidate measures into a form that satisfies CPS. A purely ordinal version of CPS is the following, strictly weaker condition:

Weak Causation-Prevention Symmetry (WCPS) For two effects E_1 and E_2 which are screened off by a common cause C (i.e., $E_1 \perp\!\!\!\perp E_2$ given C and ¬ C),

$$\eta(C, E_1) = \eta(C, E_2) \quad \text{if and only if} \quad \eta(C, \neg E_1) = \eta(C, \neg E_2)$$

This condition demands that for two equally strong effects of a common cause, their negation is also prevented to an equal degree.

We now proceed to more specific adequacy conditions that charaterize an individual measure, or a class of measures that delivers the same rankings of causal effect. This latter property is called **ordinal equivalence of measures**; it is also familiar from Variation 2. Two measures η and η' are ordinally equivalent if and only if

$$\eta(C, E) > \eta(C', E') \quad \text{if and only if} \quad \eta'(C, E) > \eta'(C', E').$$

The point of the following sections is to bring out the characteristic properties of the various available measures, in order to create a basis for comparing, discussing and appraising them. In the end, we will also explain our personal preferences and draw some tentative conclusions regarding the question of whether we should work with a single causal effect measure, or a plurality of measures.

Causal Production and the Suppes-Pearl Measure

In this subsection, we derive an axiomatic characterization of the Pearl-Suppes measure $\eta(C, E) = p(E|do(C))$ (Suppes, 1970; Pearl, 2000). To this end, we introduce a condition which is motivated by the intuition that causes produce their effects. Consider Table 6.3 which we already know from Variation 2. Three teams in the Italian *Seria A*, AS Roma, FC Internazionale ("Inter"), and Juventus ("Juve") are still competing for the

Rank	Team	Points	Team	Points
	after 36 out of 38 rounds		after 37 out of 38 rounds	
1	Roma	78	Inter	79
2	Inter	76	Roma	78
3	Juve	74	Juve	74

Table 6.3: A motivating example for Conditional Equivalence. Top of the Seria A after 36 and 37 out of 38 rounds, respectively.

scudetto, the national soccer championship. On the penultimate match day, Inter beats Juve in the *Derby d'Italia* while Roma loses to another team. Call this conjunction of events C. Let E_1 = Inter will win the championship and E_2 = Roma will be the runner-up. Given C, E_1 and E_2 are logically equivalent. (Juve misses four and five points on both teams and cannot surpass them any more.) It is now very natural to claim that C causes E_1 and E_2 to an equal degree. This intuition is stated in the following condition:

Conditional Equivalence If E_1 and E_2 are logically equivalent given C, then $\eta(C, E_1) = \eta(C, E_2)$.

Taking this condition together with Formality and Effect Production, we can prove the following theorem:

Theorem 6.1 (Representation Theorem for η_{SP}) *All measures of causal effect that satisfy Formality, Effect Production and Conditional Equivalence are ordinally equivalent to*

$$\eta_{SP}(C, E) = p(E|do(C))$$

Pearl (2000, 70) calls $\eta_{SP}(C, E) = p(E|do(C))$ the "causal effect" of C on E. This measure fits quite well with cases of causal production where we are asked to rank causes of an event according to the degree that they produced E or were responsible for E. For instance, should a car accident (E) be attributed to driving a bit too fast (C_1) or to ignoring a red traffic light (C_2)? Although both causes describe the violation of a norm, one of them has a much higher tendency to cause an accident, and $p(E|do(C))$ seems to be a good guide for ranking the causes. This position is also defended in two recent papers that apply causal relevance to liability and legal reasoning (Kaiserman, 2016a,b). Note, however, that η_{SP} violates Difference-Making, and it does not distinguish between (positive) causation, causal prevention, and causal irrelevance.

We now proceed to representation theorems for measures which violate Effect Production and satisfy Difference-Making. Given Formality and Difference-Making, each of the adequacy conditions discussed below is sufficient to single out a measure of causal effect up to ordinal equivalence. In this specific sense, those conditions are therefore incompatible with each other.

The Multiplicativity Principle and the Difference Measure

How should causal effect combine on a single path, e.g., in the graph in Figure 6.2? If causal effect is conceived of as the intensity of a physical process linking cause and effect, overall causal effect should be a function of the causal effect between the individual links. But which function $g : \mathbb{R}^2 \to \mathbb{R}$ should be chosen such that $\eta(\mathrm{C}, \mathrm{E}) = g(\eta(\mathrm{C}, \mathrm{X}), \eta(\mathrm{X}, \mathrm{E}))$?

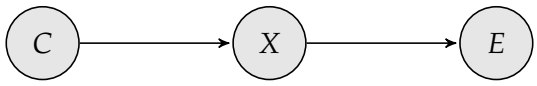

Figure 6.2: A DAG representing causation along a single path.

A couple of requirements suggest themselves. First of all, g should be symmetric: the order of mediators in a chain does not matter. Whether a weak link precedes a strong link, or vice versa, should not matter for overall causal effect. Second, it seems that the overall causal effect cannot be stronger than the weakest link in the chain: If C and X are almost independent, it does not matter how strongly X and E are correlated: the causal effect will be still weak. Similarly, if both links are weak, the overall link will be even weaker. On the other hand, if the link is maximally strong (e.g., $\eta(\mathrm{C}, \mathrm{X}) = 1$), then the strength of the entire chain will just be the strength of the rest of the chain. See also Good (1961a, 311–312).

A very simple function that satisfies all these requirements is multiplication. Thus, we obtain the following principle:

Multiplicativity along Single Paths If the variables C and E are connected via a single path with intermediate node X, then $\eta(\mathrm{C}, \mathrm{E}) = \eta(\mathrm{C}, \mathrm{X}) \cdot \eta(\mathrm{X}, \mathrm{E})$.

As a corollary, we obtain that for a causal chain with multiple mediators, e.g., $C \to X_1 \to \ldots \to X_n \to E$,

$$\eta(\mathrm{C},\mathrm{E}) = \eta(\mathrm{C},\mathrm{X}_1) \cdot \eta(\mathrm{X}_1,\mathrm{X}_2) \cdot \ldots \cdot \eta(\mathrm{X}_{n-1},\mathrm{X}_n) \cdot \eta(\mathrm{X}_n,\mathrm{E})$$

It is now possible to characterize all measures that satsify Multiplicativity along Single Paths, in addition to Formality and Difference-Making:

Theorem 6.2 (Representation Theorem for η_d) *All measures of causal effect that satisfy Formality, Difference-Making and Multiplicativity along Single Paths are ordinally equivalent to*

$$\eta_d(\mathrm{C},\mathrm{E}) = p(\mathrm{E}|do(\mathrm{C})) - p(\mathrm{E}|do(\neg\mathrm{C}))$$

This is a simple and intuitive quantity that measures the causal effect of C for E by comparing the effect that different interventions on C have on E. It possesses the *sine qua non* property that two effects in a conjunctive fork (e.g., $E_1 \leftarrow C \to E_2$) do not cause each other. It is also straightforwardly applicable in statistical inference where it is used to quantify effect size for categorical variables under an intervention on C. In clinical trials and epidemiological studies, $\eta_d(\mathrm{C},\mathrm{E})$ is identical to Absolute Risk Reduction, or ARR.

We also state two notable properties of η_d. First, it can be rewritten as

$$\begin{aligned} \eta_d(\mathrm{C},\mathrm{E}) &= p(\mathrm{E}|do(\mathrm{C})) - p(\mathrm{E}|do(\neg\mathrm{C})) \\ &= p(\neg\mathrm{E}|do(\neg\mathrm{C})) + p(\mathrm{E}|do(\mathrm{C})) - 1 \end{aligned}$$

Modulo subtraction of a constant, $\eta_d(\mathrm{C},\mathrm{E})$ is a sum of two quantities that have been called causal/explanatory necessity and causal/explanatory sufficiency by Hempel (1965) and Pearl (2000). The names are natural: $p(\neg\mathrm{E}|do(\neg\mathrm{C}),\mathrm{C},\mathrm{E})$ indicates to what extent C was *necessary* for producing E (in a world where C and E are present, what would have happened if C had not occurred?), and $p(\mathrm{E}|do(\mathrm{C}),\neg\mathrm{C},\neg\mathrm{E})$ indicates to what extent the presence of C was *sufficient* for producing E (in a world where C and E are absent, what would have happened if C had occurred?). $\eta_d(\mathrm{C},\mathrm{E})$ combines these two plausible ways of thinking about causal effect in an intuitive manner.

While this property may be regarded as superficial, the following one is more profound. Consider the proposition E_1 that a certain real-valued quantity E falls into the interval $[e_1^-, e_1^+]$ and the proposition E_2 that E

has values in $[e_2^-, e_2^+]$. Obviously, these two propositions are mutually exclusive if the intervals are. But what is the degree to which C causes E_1 *or* E_2 (that is, $E \in [e_1^-, e_1^+] \cup [e_2^-, e_2^+]$)? This question can be answered in general:

Corollary 6.1 *For* C, E_1 *and* $\mathrm{E}_2 \in \mathcal{L}$,

$$\eta_d(\mathrm{C}, \mathrm{E}_1 \vee \mathrm{E}_2) = \eta_d(\mathrm{C}, \mathrm{E}_1) + \eta_d(\mathrm{C}, \mathrm{E}_2) - \eta_d(\mathrm{C}, \mathrm{E}_1 \wedge \mathrm{E}_2). \tag{6.2}$$

In particular, if E_1 *and* E_2 *are mutually exclusive (that is, if* $\neg(\mathrm{E}_1 \wedge \mathrm{E}_2)$ *is a theorem), then the above equation reduces to*

$$\eta_d(\mathrm{C}, \mathrm{E}_1 \vee \mathrm{E}_2) = \eta_d(\mathrm{C}, \mathrm{E}_1) + \eta_d(\mathrm{C}, \mathrm{E}_2)$$

and we can also formulate the following necessary and sufficient condition on rankings of causal effect according to $\eta_d(\mathrm{C}, \mathrm{E})$*:*

$$\eta_d(\mathrm{C}, \mathrm{E}_1 \vee \mathrm{E}_2) > \eta_d(\mathrm{C}, \mathrm{E}_1) \qquad \textit{if and only if} \qquad \eta_d(\mathrm{C}, \mathrm{E}_2) > 0,$$

and vice versa with E_1 *and* E_2 *reversed.*

The proof is straightforward and left as an exercise. This means that the degree to which a mutually exclusive disjunction of effects is caused is the sum of the individual degrees of causation. In particular, causal effect is enlarged by disjunctively tacking further effects if and only if each of these effects is itself caused to a positive degree.

This corollary has an interesting implication for causal inference with multicategorial variables, such as "place of residence" or "preferred travel destination". Because such variables cannot be measured on a metric scale, they are not easy to use in statistical inference. Often, a multicategorial variable $E \in \{e_1, \ldots, e_N\}$ is encoded by a series of dummy variables, such as $E_1 = \pm e_1$, $E_2 = \pm e_1$, etc. By describing the causal effect of C on a disjunction of several dummy variables $E_i \vee E_j \vee E_k \vee \ldots \vee E_n$, Corollary 6.1 specifies the effect of C on a *range* of values of a multicategorial variable in terms of the effect that it has on the dummy variables $E_1, \ldots, E_N$.

The No Dilution for Irrelevant Effects Principle and Probability Ratio Measures

What is the causal effect for C on the conjunction of two effects—$\mathrm{E}_1 \wedge \mathrm{E}_2$—when C affects only one of them, and the other effect (say, E_2) is independent of C and of E_1? In such circumstances, we may call E_2 an "irrelevant effect". This situation is represented visually in the DAG of Figure 6.3.

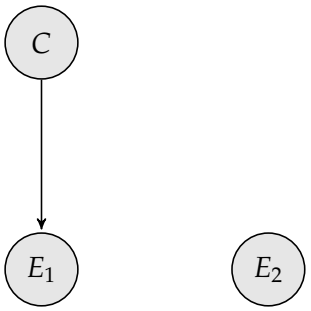

Figure 6.3: An effect E_2 which is irrelevant regarding the causal relation between C and E_1.

There are two basic intuitions about what such effects mean for overall causal effect: either the causal effect of C is diluted when passing from E_1 to $\mathrm{E}_1 \wedge \mathrm{E}_2$, or it is not. Dilution means that adding E_2 to E_1 diminishes the causal effect of C, that is, $\eta(\mathrm{C}, \mathrm{E}_1 \wedge \mathrm{E}_2) < \eta(\mathrm{C}, \mathrm{E}_1)$. A measure is non-diluting if in these circumstances, $\eta(\mathrm{C}, \mathrm{E}_1 \wedge \mathrm{E}_2) = \eta(\mathrm{C}, \mathrm{E}_1)$. This amounts to the following principle:

No Dilution for Irrelevant Effects If $E_2 \perp\!\!\!\perp C$, and $E_2 \perp\!\!\!\perp E_1$ conditional on C and $\neg$C, then $\eta(\mathrm{C}, \mathrm{E}_1 \wedge \mathrm{E}_2) = \eta(\mathrm{C}, \mathrm{E}_1)$.

Non-diluting measures of causal effect that satisfy Difference-Making can be neatly characterized. In fact, they are all ordinally equivalent to Lewis' probability ratio measure (Lewis, 1986), as the following theorem demonstrates.

Theorem 6.3 (Representation Theorem for η_r) *All measures of causal effect that satisfy Formality, Difference-Making, and No Dilution for Irrelevant Effects are ordinally equivalent to*

$$\eta_r(\mathrm{C}, \mathrm{E}) = \frac{p(\mathrm{E}|do(\mathrm{C}))}{p(\mathrm{E}|do(\neg \mathrm{C}))}$$

and its rescaling to the $[-1; 1]$ *range*

$$\eta_{r'}(\mathrm{C}, \mathrm{E}) = \frac{p(\mathrm{E}|do(\mathrm{C})) - p(\mathrm{E}|do(\neg \mathrm{C}))}{p(\mathrm{E}|do(\mathrm{C})) + p(\mathrm{E}|do(\neg \mathrm{C}))}.$$

To some extent, this result can be interpreted as a *reductio* of the probability ratio measure, and the class of measures that are ordinally equivalent to it. After all, given the lack of a causal connection between C and E_2, it is

plausible that C causes $E_1 \wedge E_2$ to a smaller degree than E_1. The probability ratio measure $\eta_r(C, E)$, however, satisfies the Principle of No Dilution for Irrelevant Effects. In particular, since the probability ratio measure is just the Relative Risk measure commonly used in epidemiology, the above arguments undermine the use of that measure in clinical practice, too.

A way around this problem consists in restricting No Dilution for Irrelevant Effects to prevention rather than (positive) causation. According to this reading, if C is a strong preventive cause of E', the degree of prevention is not diluted by adding irrelevant effects. This may be a bit more intuitive than the above principle. After all, lowering degree of prevention can be read as increasing causal effect, and why should be able to achieve this "for free" just by adding irrelevant effects? Formally, this condition reads:

No Dilution for Irrelevant Effects (Prevention) Let C be a preventive cause of E_1. If $E_2 \perp\!\!\!\perp C$, and $E_2 \perp\!\!\!\perp E_1$ conditional on C and $\neg$C, then $\eta(C, E_1 \wedge E_2) = \eta(C, E_1)$.

The adequacy of No Dilution for Irrelevant Effects is a question that will return in Variation 7, when we discuss the principle of Explanatory Justice (Crupi and Tentori, 2012; Cohen, 2016a).

If we combine this restricted version of No Dilution for Irrelevant Effects with Weak Causation-Prevention Symmetry, we get an interesting result:

Theorem 6.4 (Representation Theorem for η_c) *All measures of causal strength that satisfy Formality, Difference-Making, No Dilution for Irrelevant Effects (Prevention) and Causation-Prevention Symmetry are ordinally equivalent to*

$$\eta_c(C, E) = \begin{cases} \frac{p(E|do(C)) - p(E|do(\neg C))}{1 - p(E|do(\neg C))} & \text{if C is a positive cause of E} \\ \frac{p(E|do(C)) - p(E|do(\neg C))}{p(E|do(\neg C))} & \text{if C is a preventive cause of E} \end{cases}.$$

This measure agrees, for the case of positive causation, with two proposals from the literature. The psychologist Patricia Cheng (1997) derived η_c from theoretical considerations about how people perform causal induction and called it the causal power of C on E. I.J. Good (1961a,b) derived a measure that is ordinally equivalent to η_c from a complex set of theoretical

adequacy conditions. Here is Good's rescaling of η_c:

$$\eta_g(\mathrm{C},\mathrm{E}) = \begin{cases} \frac{1-p(\mathrm{E}|do(\neg \mathrm{C}))}{1-p(\mathrm{E}|do(\mathrm{C}))} & \text{if C is a positive cause of E} \\ \frac{p(\mathrm{E}|do(\mathrm{C}))}{p(\mathrm{E}|do(\neg \mathrm{C}))} & \text{if C is a preventive cause of E} \end{cases}$$

The two previous theorems elucidate that η_r and η_c are based on the same principle: the No Dilution for Irrelevant Effects property. The crucial question which separates the two measures is whether this property should hold across the board or just for preventive causes.

Conjunctive Closure and the Logarithmic Ratio Measure

Consider a variable C that affects a set of other variables E_1, E_2, ..., which would be unrelated to each other, were it not for their common cause C. In this scenario, represented visually in Figure 6.4, one could ask how the causal effect of C on each individual effect (E_1, E_2) affects the causal effect that C exerts on the conjunction of these variables. In other words, we ask how $\eta(\mathrm{C}, \mathrm{E}_1 \wedge \mathrm{E}_2)$ depends on $\eta(\mathrm{C}, \mathrm{E}_1)$ and $\eta(\mathrm{C}, \mathrm{E}_2)$, and under which circumstances the former is a function of the latter.

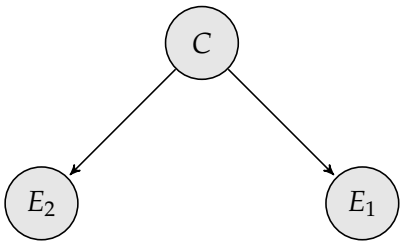

Figure 6.4: A typical common cause structure where C screens off the two effects E_1 and E_2.

A plausible principle for characterizing this dependency is stated below. It is analogous to the conjunction principle in epistemology, which states that justification and/or knowledge is closed under logical conjunction. Shogenji (2012) transfers this principle to quantitative measures of justification: when (i) the degree of justification of H_1 and H_2 given E is both equal to t and (ii) H_1 and H_2 are probabilistically independent (unconditionally and conditionally on E), then the degree of justification of $\mathrm{H}_1 \wedge \mathrm{H}_2$ should also be equal to t. Put differently, justification is not di-

luted under the conjunction of independent propositions. Shogenji (2012) calls this the Special Conjunction Principle.

Transferred to causal effect, this would mean that the causal effect of C on $E_1 \wedge E_2$ equals the causal effect of C on either E_1 or E_2 if $E_1 \perp\!\!\!\perp E_2$, conditional on C and $\neg$C. In other words, causal effect is closed under the conjunction of independent effects. Formally:

Conjunctive Closure If $E_1 \perp\!\!\!\perp E_2$ given C and $\neg$C and $\eta(C, E_1) = \eta(C, E_2) = t$, then also $\eta(C, E_1 \wedge E_2) = t$.

This principle is plausible for doxastic justification and it is appealing for causal effect as well. Imagine, for example, that a medical drug has two side effects—diarrhea and sore throat—which are independent of each other, and that both side effects are caused with the same strength t. It is then natural to say that the overall side effect of the medical drug is also equal to t since there is no interaction between both effects. Apart from the intuitive plausibility, this principle facilitates practical calculations because we can often infer the strength of a complex causal effect from the strength of the individual effects.

It is possible to characterize measures which satisfy Conjunctive Closure up to ordinal equivalence, similar to how Atkinson (2012) described justification measures that satisfy the Special Conjunction Principle. In fact, our theorem and proof follows Atkinson's example quite closely.

Theorem 6.5 (Representation Theorem for η_{lr}) *All measures of causal effect that satisfy Formality, Difference-Making and Conjunctive Closure are ordinally equivalent to*

$$\eta_{lr}(C, E) = \frac{\log p(E|do(C))}{\log p(E|do(\neg C))}$$

We call this measure the Logarithmic Ratio measure since it is based on the ratio of logarithms of $p(E|do(C))$ and $p(E|do(\neg C))$, rather than on the ratio of probabilities, such as in the Lewis measure. Although this measure has not yet been proposed in the literature, it is a serious candidate for a measure of causal effect and deserves our attention.

Application: Quantifying Causal Effect in Medicine

A natural scientific application of probabilistic measures of causal effect consists in quantifying the size of an effect that an intervention on one

variable has on another. As we have already said when discussing Table 6.1, there are several ways of measuring effect size that are employed in the medical literature. They can be related straightforwardly to probabilistic measures of causal effect when we write the relative frequencies of an event as probabilities (e.g., A/(A+B) is just the frequency of E among all C's). In particular, they read as follows:

$$\text{RR} = \frac{p(\text{E}|do(\text{C}))}{p(\text{E}|do(\neg\text{C}))} \qquad \text{(Relative Risk)}$$

$$\text{OR} = \frac{p(\text{E}|do(\text{C}))/p(\neg E|C)}{p(\text{E}|do(\neg\text{C}))/p(\neg E|do(\neg C))} \qquad \text{(Odds Ratio)}$$

$$\text{ARR} = p(\text{E}|do(\text{C})) - p(\text{E}|do(\neg\text{C})) \qquad \text{(Absolute Risk Reduction)}$$

It is not difficult to relate these measures to the measures we discussed. For example, RR is just the familiar probability ratio measure η_r, whereas ARR turns out to be the difference measure η_d. OR is the product of the probability ratio measure η_r and Good's measure η_g. Also normative arguments in favor or against causal effect measures carry over to effect size measures. For example, Multiplicativity along Single Paths—the defining property of η_d—sounds very reasonable in the context of medical inference, whereas the No Dilution for Irrelevant Effect property—the defining property of η_r—is apparently problematic. Our results may thus be used for construing an argument for preferring the Absolute Risk Reduction measure ARR over its more popular competitor RR, the Relative Risk measure. Our theoretical arguments nicely square with decision-theoretic and epistemic arguments for preferring absolute over relative measures of risk reduction in medicine, e.g., the neglect of prior probabilities in relative measures (Stegenga, 2015; Sprenger and Stegenga, 2016). Without pursuing this topic in detail—this would deserve a separate paper—it should be evident that our analysis of causal effect has important applications in scientific inference and medical science in particular.

Discussion

This variation has provided axiomatic foundations for a probabilistic theory of causal effect, proceeding toward a more systematic investigation of that topic. It synthesizes ideas from the manipulability/interventionist view of causation and the probabilistic relevance view of causation. While

causal Bayes nets provide the framework for our analysis, the methods for characterizing the various measures are transferred from various parts of formal epistemology, in particular Bayesian confirmation theory and Bayesian analyses of explanatory power.

After introducing the conceptual and mathematical framework, we have noted that intuitions about measures of causal effect pull into different directions. This makes it difficult to come up with "the one true measure of causal effect", in analogy to what has been tried in confirmation theory (Milne, 1996). However, this does not render the project futile. By contrast, characterizing each measure by a combination of adequacy conditions makes it possible to assess the (possibly context-sensitive) value of the different measures by means of assessing the plausibility of the adequacy conditions. Even if more than a single measure survives the theoretical scrutiny, one can still form informed preferences. Below we make a case for the difference measure η_d.

Notably, the measures which we investigated fall into two major categories: those who do and those who do not satisfy the Difference-Making property (i.e., causal effect is a function of $p(\mathrm{E}|do(\mathrm{C}))$ and $p(\mathrm{E}|do(\neg\mathrm{C}))$, increasing in the first and decreasing in the second argument). Only the first measure in our list—the Suppes-Pearl measure $\eta_{SP}(\mathrm{C},\mathrm{E}) = p(\mathrm{E}|do(\mathrm{C}))$—fails to satisfy this condition because it does not depend on $p(\mathrm{E}|do(\neg\mathrm{C}))$, that is, on the contrastive value that a cause has for an effect. However, it may be suitable for quantifying degree of causal production in cases of actual causation, when we are more interested in questions of attribution and liability than in the predictive value of a cause for an effect (e.g., Kaiserman, 2016a,b). The other measures are more straightforward expressions of counterfactual dependence: how much does a change in the value of C affect the outcome E? See also Beckers and Vennekens (2016) for the role of production and dependence in judgments of causation.

The properties of the investigated measures are summarized in Table 6.4. It is notable that only two measures (η_d and η_g) satisfy the Weak Causation-Prevention Symmetry, although this is an eminently sensible property. The same can be said about Multiplicativity along Single Paths, which is only satisfied by η_d. One should add, however, that this property is significantly stronger since it also varies among measures in one and the same ordinal equivalence class. On the other side, by characterizing the η_r-, η_g- and η_c-measure mathematically, Theorems 6.3 and 6.4 also point

	Property								
Measure	FORM	EP	DM	WCPS	CE	MUL	NDIE	NDIEP	CC
Suppes/Pearl (η_{SP})	yes	yes	no	no	yes	no	no	no	no
Eells (η_d)	yes	no	yes	yes	no	yes	no	no	no
Lewis (η_r, $\eta_{r'}$)	yes	no	yes	no	no	no	yes	yes	no
Cheng/Good (η_g, η_c)	yes	no	yes	yes	no	no	no	yes	no
Log-Ratio (η_{lr})	yes	no	yes	no	no	no	no	no	yes

Table 6.4: A classification of different measures of causal effect according to the adequacy conditions that they satisfy. FORM = Formality, EP = Effect Production, DM = Difference-Making, WCPS = Weak Causation-Prevention Symmetry, CE = Coonditional Equivalence, MUL = Multiplicativity along Single Paths, NDIE = No Dilution for Irrelevant Effects, NDIEP = No Dilution for Irrelevant Effects (Prevention), CC = Conjunctive Closure.

out problems with rivaling measures of causal effect (=that they satisfy the questionable No Dilution principle).

All in all, the above analysis provides good grounds for using η_d as a default measure of causal effect. Indeed, Pearl (2001) bases his quantification of path-specific effects on η_d as underlying the basic measure of causal effect, without justifying this choice further. We are closing this gap. The formal analysis also mirrors and supports practice- and decision-oriented arguments for η_d vis-'a-vis its competitors, e.g., in medical science (Stegenga, 2015; Sprenger and Stegenga, 2016).

What remains to do? First of all, we may aim at generalizing the framework from binary variables to categorical and real-valued variables. Indeed, many measures of effect size for real-valued variables, such as Cohen's d or Glass's Δ, are based on the difference of group means, and η_d might be extended naturally into this direction. As long as the cause is a binary variable, that is, as long as only two different values of C are compared, our analysis holds water. The same calculations still apply even if E is a real-valued variable. Our approach may thus go a longer way toward modeling scientific inference about causal effect than our earlier restriction to binary variables may suggest.

Second, the properties of the above measures, and in particular η_d, in complicated networks (e.g., more than one path linking C and E) have not been investigated. Is it possible to show, for example, how degrees of causation along different paths can be combined in an overall assessment of causal effect, e.g., similar to Theorem 3 in Pearl (2001)?

Third, this work can be connected to information-theoretic approaches

to *causal specificity* (Weber, 2006; Waters, 2007; Korb et al., 2011; Griffiths et al., 2015). The more narrow the range of effects that an intervention is likely to produce, the more specific the cause is to the effect. How does this concept relate to causal effect and to what extent can both research programs learn from each other?

Fourth, we would like to apply this theory to canonical examples in the causation literature and to explore whether our understanding of causal effect squares well with the significance of normality and norms in causal reasoning (Knobe and Fraser, 2008; Hitchcock and Knobe, 2009).

These are all open and exciting questions, and it is not difficult to come up with others. We hope, however, that the results presented herein are promising enough to motivate a further pursuit of an axiomatic theory of causal effect.

Proofs of the Theorems

Proof of Theorem 6.1: The proof relies on a recent result by Michael Schippers (2016) in the field of confirmation theory. Schippers demonstrates that the following three conditions are necessary and sufficient to characterize the posterior probability $c^*(\mathrm{E},\mathrm{H}) := p(\mathrm{H}|\mathrm{E})$ as a measure of degree of confirmation, up to ordinal equivalence.

Formality (Confirmation) There is a measurable function $f' : [0,1]^3 \to \mathbb{R}$ such that for any $h, e \in \mathfrak{L}$ with probability distribution $p(\cdot)$, $c(\mathrm{E},\mathrm{H}) = f'(p(\mathrm{E}), p(\mathrm{H}), p(\mathrm{H} \wedge \mathrm{E}))$.

Final Probability Incrementality For any sentences H, E_1, and $E_2 \in \mathfrak{L}$ with probability measure $p(\cdot)$,

$$c(\mathrm{E}_1,\mathrm{H}) > c(\mathrm{E}_2,\mathrm{H}) \qquad \text{if and only if} \qquad p(\mathrm{H}|\mathrm{E}_1) \; > \; p(\mathrm{H}|\mathrm{E}_2).$$

Local Equivalence If H_1 and H_2 are logically equivalent given E, then $c(\mathrm{E},\mathrm{H}_1) = c(\mathrm{E},\mathrm{H}_2)$.

It is easy to see that Final Probability Incrementality translates into Effect Production when the pair $(\mathrm{H},\mathrm{E}_{1,2})$ is mapped to $(\mathrm{E},\mathrm{C}_{1,2})$:

$$\eta(\mathrm{C}_1,\mathrm{E}) > \eta(\mathrm{C}_2,\mathrm{E}) \qquad \text{if and only if} \qquad p(\mathrm{E}|\mathrm{C}_1) \; > \; p(\mathrm{E}|\mathrm{C}_2)$$

The same is true for Local Equivalence: with $(\mathrm{H}_{1,2},\mathrm{E})$ mapped to $(\mathrm{E}_{1,2},\mathrm{C})$, it postulates that if E_1 and E_2 are logically equivalent given C, then $\eta(\mathrm{C},\mathrm{E}_1) = \eta(\mathrm{C},\mathrm{E}_2)$. This is just the same as Conditional Equivalence.

Thus it remains to show that Formality (Causal Effect) can be transformed into Formality (Confirmation) by a suitable change of variables. We already know that there exists a $f : [0,1]^3 \to \mathbb{R}$ such that $\eta(\mathrm{C},\mathrm{E}) = f(p(\mathrm{C}), p(\mathrm{E}|do(\mathrm{C})), \mathrm{p}(\mathrm{E}|do(\neg\mathrm{C}))$. Since we only want to characterize f mathematically, we restrict ourselves to the case where E is among the descendants of C and they share no common causes. We also assume that $p(\mathrm{C}) \in (0,1)$. This allows us to write

$$p(\mathrm{E} \wedge \mathrm{C}) = p(\mathrm{C})p(\mathrm{E}|do(\mathrm{C})) \quad p(\mathrm{E}) = p(\mathrm{C})p(\mathrm{E}|do(\mathrm{C})) + (1 - \mathrm{p}(\mathrm{C}))\mathrm{p}(\mathrm{E}|do(\neg\mathrm{C}))$$

which we can transform into the equations

$$p(\mathrm{E}|do(\mathrm{C})) = \frac{p(\mathrm{E} \wedge \mathrm{C})}{p(\mathrm{C})} \quad p(\mathrm{E}|do(\neg\mathrm{C})) = \frac{p(\mathrm{E}) - p(\mathrm{C})p(\mathrm{E}|do(\mathrm{C}))}{1 - p(\mathrm{C})} \tag{6.3}$$

Hence, we can write $p(\mathrm{E}|do(\mathrm{C}))$ and $p(\mathrm{E}|do(\neg\mathrm{C}))$ as functions of $p(\mathrm{C})$, $p(\mathrm{E})$ and $p(\mathrm{E} \wedge \mathrm{C})$. In other words, there is a function $f'(p(\mathrm{C}), p(\mathrm{E}), p(C \wedge E))$ that characterizes $\eta(\mathrm{C}, \mathrm{E})$, namely

$$\begin{aligned} f'(p(\mathrm{C}), p(\mathrm{E}), p(C \wedge E)) &:= f\left(p(\mathrm{C}), \frac{p(\mathrm{E} \wedge \mathrm{C})}{p(\mathrm{C})}, \frac{p(\mathrm{E}) - p(\mathrm{C})p(\mathrm{E}|do(\mathrm{C}))}{1 - p(\mathrm{C})}\right) \\ &= f(p(\mathrm{C}), p(\mathrm{E}|do(\mathrm{C})), \mathrm{p}(\mathrm{E}|do(\neg\mathrm{C})) \\ &= \eta(\mathrm{C}, \mathrm{E}) \end{aligned}$$

f' is continuous because f and the functions in Equation (6.3) are. Thus we can extend f' canonically to the set $\{p(\mathrm{C}) \in \{0, 1\}\}$. Hence we can invoke Schippers' theorem which shows that $\eta(\mathrm{C}, \mathrm{E}) = p(\mathrm{E}|\mathrm{C})$ up to ordinal equivalence. □

Proof of Theorem 6.2: The proof of this representation theorem proceeds in several steps. First, we will show that $\eta(\mathrm{C}, \mathrm{E}) = f(p(\mathrm{C}), p(\mathrm{E}|do(\mathrm{C})), \mathrm{p}(\mathrm{E}|do(\neg\mathrm{C}))$ does not depend on $p(\mathrm{C})$.

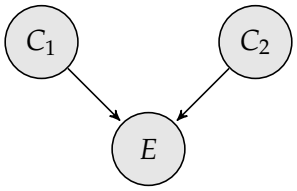

Figure 6.5: A classical collider/joint effect structure in a causal net.

The proof of this first claim proceeds by contradiction. Consider that there are real numbers $x_1, x_2, y, z \in [0, 1]$ such that $f(x_1, y, z) \neq f(x_2, y, z)$. Then choose E, C_1 and C_2 such that E is a joint effect of C_1 and C_2 with $x_1 = p(C_1)$, $x_2 = p(C_2)$, $y = p(\mathrm{E}|do(\mathrm{C}_1)) = p(\mathrm{E}|do(\mathrm{C}_2))$, $z = p(\mathrm{E}|do(\neg\mathrm{C}_1)) = p(\mathrm{E}|do(\neg\mathrm{C}_2))$. In this case, Difference-Making tells us that $\eta(\mathrm{C}_1, \mathrm{E}) = \eta(\mathrm{C}_2, \mathrm{E})$. However, on the other hand, we also know

$$\eta(\mathrm{C}_1, \mathrm{E}) = f(x_1, y, z) \neq f(x_2, y, z) = \eta(\mathrm{C}_2, \mathrm{E})$$

This leads to a straightforward contradiction. Hence, from now on we focus on the function $g : [0, 1]^2 \to \mathbb{R}$ such that $\eta(\mathrm{C}, \mathrm{E}) = g(p(\mathrm{E}|do(\mathrm{C})), \mathrm{p}(\mathrm{E}|do(\neg\mathrm{C}))$.

The second step of the proof consists in deriving the equality

$$g(\alpha, \bar{\alpha}) \cdot g(\beta, \bar{\beta}) = g(\alpha\beta + (1 - \alpha)\bar{\beta}, \bar{\alpha}\beta + (1 - \bar{\alpha})\bar{\beta}) \tag{6.4}$$

To this end, recall the Bayesian network from the main paper. It is reproduced in Figure 6.6. Again, for the purpose of investigating the formal properties of g, we can focus on those cases where $p(\mathrm{E}|\pm\mathrm{C})$ and $p(\mathrm{E}|\pm\mathrm{C})$ agree.

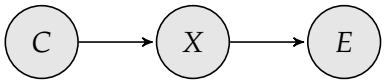

Figure 6.6: The Bayesian Network for causation along a single path.

We know by Multiciplity along Single Paths that

$$\begin{aligned}\eta(\mathrm{C},\mathrm{E}) &= \eta(\mathrm{C},\mathrm{X}) \cdot \eta(\mathrm{X},\mathrm{E}) \\ &= g(p(\mathrm{X}|do(\mathrm{C})), p(\mathrm{X}|do(\neg\mathrm{C}))) \cdot g(p(\mathrm{E}|do(\mathrm{X})), p(\mathrm{E}|do(\neg\mathrm{X}))) \\ &= g(p(\mathrm{X}|\mathrm{C}), p(\mathrm{X}|\neg\mathrm{C})) \cdot g(p(\mathrm{E}|\mathrm{X}), p(\mathrm{E}|\neg\mathrm{X}))\end{aligned}$$

and at the same time,

$$\begin{aligned}\eta(\mathrm{C},\mathrm{E}) &= g(p(\mathrm{E}|do(\mathrm{C})), \mathrm{p}(\mathrm{E}|do(\neg\mathrm{C}))) \\ &= g\left(\sum_{\pm\mathrm{X}} p(\mathrm{X}|\mathrm{C})p(\mathrm{E}|\mathrm{C},\mathrm{X}), \sum_{\pm\mathrm{X}} p(\mathrm{X}|\neg\mathrm{C})p(\mathrm{E}|\neg\mathrm{C},\mathrm{X})\right)\end{aligned}$$

Combining both equations yields

$$\begin{aligned}& g(p(\mathrm{X}|\mathrm{C}), p(\mathrm{X}|\neg\mathrm{C})) \cdot g(p(\mathrm{E}|\mathrm{X}), p(\mathrm{E}|\neg\mathrm{X})) \\ = \; & g\left(\sum_{\pm X} p(\mathrm{X}|\mathrm{C})p(\mathrm{E}|\mathrm{C},\mathrm{X}), \sum_{\pm X} p(\mathrm{X}|\neg\mathrm{C})p(\mathrm{E}|\neg\mathrm{C},\mathrm{X})\right)\end{aligned}$$

With the variable settings

$$\begin{aligned}\alpha &= p(\mathrm{X}|\mathrm{C}) & \beta &= p(\mathrm{E}|\mathrm{X}) \\ \bar{\alpha} &= p(\mathrm{X}|\neg\mathrm{C}) & \bar{\beta} &= p(\mathrm{E}|\neg\mathrm{X})\end{aligned}$$

equation (6.4) follows immediately.

Third, we are going to show that

$$g(x,y) = g(x-y,0) \tag{6.5}$$

To this end, we first note a couple of facts about g:[3]

[3]In the proof, negative arguments of g figure. This may look problematic, but it is not. We just show that any $g(\cdot,\cdot)$ that satisfies Equation (6.4) on$[0,1]^2$ has an extension to a function on $\mathbb{R}^2$ that satisfies certain properties, which can in turn be used for saying something about the behavior of g on $[0,1]^2$.

Fact 1 $g(\alpha,0)g(\beta,0) = g(\alpha\beta,0)$. This follows immediately from equation (6.4) with $\bar{\alpha} = \bar{\beta} = 0$.

Fact 2 $g(1,0) = 1$. With $\beta = 1$, the previous fact entails that $g(\alpha,0)g(1,0) = g(\alpha,0)$. Hence, either $g(\alpha,0) \equiv 0$ for all values of α (which would trivialize g) or $g(1,0) = 1$.

Fact 3 $g(0,1) = -1$. Fact 1 entails (with $\alpha = \beta = 0$, $\bar{\alpha} = \bar{\beta} = 1$) that $g(0,1) \cdot g(0,1) = g(1,0) = 1$. Hence, either $g(0,1) = -1$ or $g(0,1) = 1$. If the latter were the case, then g would take positive values although $p(\mathrm{E}|do(\mathrm{C})) = 0$ and $p(\mathrm{E}|do(\neg\mathrm{C})) > 0$, in violation of Difference-Making. Thus, $g(0,1) = -1$.

Fact 4 $g(-1,0) = -1$. By Fact 1, $g(-1,0) \cdot g(-1,0) = g(1,0) = 1$. Then we apply the same reasoning as in the proof of Fact 3.

Fact 5 $g(0,1) \cdot g(\beta,\bar{\beta}) = g(\bar{\beta},\beta)$. Follows immediately from equation (6.4) with $\alpha = 0$, $\bar{\alpha} = 1$.

These facts will allow us to derive Equation (6.5). Note that (6.5) is trivial if $y = 0$. So we can restrict ourselves to the case that $y > 0$. We choose the variable settings

$$\begin{aligned} \alpha &= \frac{y-x}{y} & \beta &= 0 \\ \bar{\alpha} &= 0 & \bar{\beta} &= y \end{aligned}$$

Then we obtain by means of Equation (6.4) and the previously proven facts

$$\begin{aligned} g(x,y) &= g((y-x)/y,0) \cdot g(0,y) \\ &= g(y-x,0) \cdot g(1/y,0) \cdot g(0,y) \qquad \text{(Fact 1)} \\ &= g(y-x,0) \cdot g(1/y,0) \cdot g(y,0) \cdot g(0,1) \qquad \text{(Fact 5)} \\ &= g(y-x,0) \cdot g(1,0) \cdot g(-1,0) \qquad \text{(Fact 1+3+4)} \\ &= g(x-y,0) \qquad \text{(Fact 1+2)} \end{aligned}$$

This implies

$$\eta(\mathrm{C},\mathrm{E}) = g(p(\mathrm{E}|do(\mathrm{C})),\mathrm{p}(\mathrm{E}|do(\neg\mathrm{C}))) = \mathrm{g}(\mathrm{p}(\mathrm{E}|do(\mathrm{C})) - \mathrm{p}(\mathrm{E}|do(\neg\mathrm{C})),0)$$

Hence, $\eta(\mathrm{C},\mathrm{E})$ is a function of $p(\mathrm{E}|do(\mathrm{C})) - \mathrm{p}(\mathrm{E}|do(\neg\mathrm{C}))$ only. It is easy to see that this function must be monotonic, that is, g is monotonically

increasing in its first argument. Otherwise there would be $x, y \in [0,1]$ with $x > y$ and $g(x,0) < g(y,0)$. In that case, application of Equation (6.5) and Inference to the Only Explanation yields

$$0 > g(x,0) - g(y,0) = g(x-y,0) \geq 0$$

and a contradiction results. This concludes the proof of the Theorem. □

Proof of Theorem 6.3: The proof relies on a move from the proof of Theorem 1 in Schupbach and Sprenger (2011). Consider three variables C, E_1 and E_2 which satisfy the premises of the No Dilution for Irrelevant Effects Principle. This means that

$$\begin{aligned} p(\mathrm{E}_1 \wedge \mathrm{E}_2 | do(\mathrm{C})) &= p(\mathrm{E}_1 | do(\mathrm{C})) \cdot p(\mathrm{E}_2 | do(\mathrm{C})) \\ p(\mathrm{E}_1 \wedge \mathrm{E}_2 | do(\neg \mathrm{C})) &= p(\mathrm{E}_1 | do(\neg \mathrm{C})) \cdot p(\mathrm{E}_2 | do(\neg \mathrm{C})) \\ p(\mathrm{E}_2) &= p(\mathrm{E}_2 | do(\neg \mathrm{C})) = p(\mathrm{E}_2 | do(\mathrm{C})) \end{aligned}$$

In particular it follows that

$$\begin{aligned} p(\mathrm{E}_1 \wedge \mathrm{E}_2 | do(\mathrm{C})) &= p(\mathrm{E}_2)\, p(\mathrm{E}_1 | do(\mathrm{C})) \\ p(\mathrm{E}_1 \wedge \mathrm{E}_2 | do(\neg \mathrm{C})) &= p(\mathrm{E}_2)\, p(\mathrm{E}_1 | do(\neg \mathrm{C})) \end{aligned}$$

According to Formality and Difference-Making, the causal effect measure η can be written as $\eta(\mathrm{C}, \mathrm{E}) = g(p(\mathrm{E}_1 | do(\mathrm{C})), p(\mathrm{E}_1 | do(\neg \mathrm{C})))$ for a continuous function g. From No Dilution for Irrelevant Effects and the above calculations we can infer that

$$\begin{aligned} g(p(\mathrm{E}_1 | do(\mathrm{C})), p(\mathrm{E}_1 | do(\neg \mathrm{C}))) &= \eta(\mathrm{C}, \mathrm{E}_1) \\ &= \eta(\mathrm{C}, \mathrm{E}_1 \wedge \mathrm{E}_2) \\ &= g(p(\mathrm{E}_1 \wedge \mathrm{E}_2 | do(\mathrm{C})), p(\mathrm{E}_1 \wedge \mathrm{E}_2 | do(\neg \mathrm{C}))) \\ &= g(p(\mathrm{E}_2)\, p(\mathrm{E}_1 | do(\mathrm{C})), p(\mathrm{E}_2)\, p(\mathrm{E}_1 | do(\neg \mathrm{C}))) \end{aligned}$$

Since we have made no assumptions on the values of these probabilities, we can infer the general relationship

$$g(x,y) = g(cx, cy). \tag{6.6}$$

for all $0 < c \leq \min(1/x, 1/y)$. Without loss of generality, let $x > y$. Then, choose c := 1/x. In this case, equation 6.9 becomes

$$g(x,y) = g(cx, cy) = g(1, y/x).$$

This implies that g must be a function of y/x only, that is, of the ratio $p(\mathrm{E}|do(\mathrm{C}))/p(\mathrm{E}|do(\neg\mathrm{C}))$. Difference-Making then implies that all such functions must be monotonically increasing, concluding the proof of the theorem. □

Proof of Theorem 6.4: We begin by showing that η_c and η_g are ordinally equivalent. For positive causation,

$$\eta_c(\mathrm{C},\mathrm{E}) = \frac{-p(\neg\mathrm{E}|do(\mathrm{C})) + p(\neg\mathrm{E}|do(\neg\mathrm{C}))}{p(\neg\mathrm{E}|do(\neg\mathrm{C}))} = 1 - \frac{1}{\eta_g(\mathrm{C},\mathrm{E})}$$

and for causal prevention,

$$\eta_c(\mathrm{C},\mathrm{E}) = \frac{p(\mathrm{E}|do(\mathrm{C})) - p(\mathrm{E}|do(\neg\mathrm{C}))}{p(\mathrm{E}|do(\neg\mathrm{C}))} = 1 - \frac{1}{\eta_g(\mathrm{C},\mathrm{E})}$$

After these preliminaries, we start with the real proof. The causal effect measure η that satisfies Formality, Difference-Making, No Dilution for Irrelevant Effects (Prevention) and WCPS can be represented by a function $g(x,y)$ with $x = p(\mathrm{E}|do(\mathrm{C}))$ and $y = p(\mathrm{E}|do(\neg\mathrm{C}))$. Suppose that there are $x > y$ and $x' > y' \in [0,1]$ such that $(1-x)/(1-y) = (1-x')/(1-y')$, but $g(x,y) \neq g(x',y')$. (Otherwise η would just be ordinally equivalent to η_g and η_c.) In that case we can find a probability space such that $p(\mathrm{E}_1|do(\mathrm{C})) = x$, $p(\mathrm{E}_1|do(\neg\mathrm{C})) = y$, $p(\mathrm{E}_2|do(\mathrm{C})) = x'$, $p(\mathrm{E}_2|do(\neg\mathrm{C})) = y'$ and C screens off E_1 and E_2 (proof omitted, but straightforward). Hence $\eta(\mathrm{C},\mathrm{E}_1) \neq \eta(\mathrm{C},\mathrm{E}_2)$. By Weak Causation-Prevention Symmetry, we can then infer $\eta(\mathrm{C},\neg\mathrm{E}_1) \neq \eta(\mathrm{C},\neg\mathrm{E}_2)$.

But this leads to a straightforward contradiction. After all, for cases of causal prevention, we can apply the previous representation theorem relating to a conjunction of Formality, Difference-Making and No Dilution for Irrelevant Effects. This implies that for cases of causal prevention,

$$\eta(\mathrm{C},\mathrm{E}) = f\left(\frac{p(\mathrm{E}|do(\mathrm{C}))}{p(\mathrm{E}|do(\neg\mathrm{C}))}\right)$$

for some monotonically increasing function f. Hence,

$$\begin{aligned}\eta(\mathrm{C},\mathrm{E}_1) &= f\left(\frac{p(\mathrm{E}_1|do(\mathrm{C}))}{p(\mathrm{E}_1|do(\neg\mathrm{C}))}\right)\\ &= f\left(\frac{1-x}{1-y}\right)\end{aligned}$$

and analogously,

$$\begin{aligned}\eta(\mathrm{C},\mathrm{E}_2) &= f\left(\frac{\log p(\mathrm{E}_2|do(\mathrm{C}))}{\log p(\mathrm{E}_2|do(\neg\mathrm{C}))}\right)\\ &= f\left(\frac{1-x'}{1-y'}\right)\end{aligned}$$

Thus, it follows from $\eta(\mathrm{C},\neg\mathrm{E}_1) \neq \eta(\mathrm{C},\neg\mathrm{E}_2)$ that

$$f\left(\frac{1-x}{1-y}\right) \neq f\left(\frac{1-x'}{1-y'}\right)$$

in contradiction with our assumption that $(1-x)/(1-y) = (1-x')/(1-y')$. Hence, there is a function h such that $g(x,y) = h((1-x)/(1-y))$, and a function $h' := 1/h$ such that $g(x,y) = h'((1-y)/(1-x))$. By Difference-Making, this function must be monotonically increasing. This shows that any causal effect measure that satisfies our four conditions must be ordinally equivalent to η_g, and hence also to η_c. □

Proof of Theorem 6.5: By Formality and Difference-Making, we have that $\eta(\mathrm{C},\mathrm{E}) = g(p(\mathrm{E}|do(\mathrm{C})), p(\mathrm{E}|do(\neg\mathrm{C})))$ for some continuous function $g : [0,1]^2 \to \mathbb{R}$. Assume now that $\eta(\mathrm{C},\mathrm{E}_1) = \eta(\mathrm{C},\mathrm{E}_2) = t$, that C screens off E_1 and E_2 and that $p(\mathrm{E}_1|do(\mathrm{C})) = p(\mathrm{E}_2|do(\mathrm{C})) = x$, $p(\mathrm{E}_1|do(\neg\mathrm{C})) = p(\mathrm{E}_2|do(\neg\mathrm{C})) = y$, for some $x, y \in \mathbb{R}$. By the Conjunctive Closure Principle, we can infer

$$\eta(\mathrm{C},\mathrm{E}_1 \wedge \mathrm{E}_2) = \eta(\mathrm{C},\mathrm{E}_1) = g(x,y)$$

Moreover, we can infer

$$\begin{aligned}\eta(\mathrm{C},\mathrm{E}_1 \wedge \mathrm{E}_2) &= g(p(\mathrm{E}_1 \wedge \mathrm{E}_2|do(\mathrm{C})), p(\mathrm{E}_1 \wedge \mathrm{E}_2|do(\neg\mathrm{C})))\\ &= g(p(\mathrm{E}_1|do(\mathrm{C})) \cdot p(\mathrm{E}_2|do(\mathrm{C})), p(\mathrm{E}_1|do(\neg\mathrm{C})) \cdot p(\mathrm{E}_2|do(\neg\mathrm{C})))\\ &= g(x^2,y^2)\end{aligned}$$

Taking both calculations together, we obtain

$$g(x^2,y^2) = g(x,y) \tag{6.7}$$

as a structural requirement on the function g, since we have not made any assumptions on x and y.

Following Atkinson (2012) and his proof idea, we now define $u = \frac{\log x}{\log y}$ and define a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x,u) := g(x,y)$. Equation 6.7 then implies the requirement

$$f(x^2,u) = g(x^2,y^2) = g(x,y) = f(x,u)$$

and by iterating the same procedure, we also obtain

$$f(x^{2n},u) = f(x,u)$$

for some $n \in \mathbb{N}$. Due to the continuity of f, we can infer that f cannot depend on its first argument. Moreover, taking the limit $n \to \infty$ yields $f(x,u) = f(0,u)$. Hence, also

$$g(x,y) = f(0,u) = f(0, \log x / \log y)$$

and we see that

$$\eta(\mathrm{C},\mathrm{E}) = h\left(\frac{\log p(\mathrm{E}|do(\mathrm{C}))}{\log p(\mathrm{E}|do(\neg\mathrm{C}))}\right)$$

for some function $h : \mathbb{R} \to \mathbb{R}$. It remains to show that h is monotonically increasing. Difference-Making implies that $\eta(\mathrm{C},\mathrm{E})$ is an increasing function of $p(\mathrm{E}|do(\mathrm{C}))$ and a decreasing function of $p(\mathrm{E}|do(\neg\mathrm{C}))$. So it must be an increasing function of $\log p(\mathrm{E}|do(\mathrm{C}))/\log p(\mathrm{E}|do(\neg\mathrm{C}))$, too. This implies that h is a monotonically increasing function and concludes the proof that all measures of causal effect that satisfy Formality, Difference-Making and the Conjunctive Closure Principle are ordinally equivalent to

$$\eta_{cc}(\mathrm{C},\mathrm{E}) = \frac{\log p(\mathrm{E}|do(\mathrm{C}))}{\log p(\mathrm{E}|do(\neg\mathrm{C}))}. \qquad \square$$

Variation 7: Explanatory Power

Explanation is a central element of scientific reasoning. Scientists from cognitive science, artificial intelligence and computer science avidly study **abductive inference**, that is, inference where the explanatory value of a hypothesis for a set of phenomena obtains special status (e.g., Hobbs et al., 1988; Bylander et al., 1991; Eiter and Gottlob, 1995; Magnani, 2001; Douven, 2011). Statisticians often give maximum likelihood estimates of an unknown parameter. In other words, they endorse the parameter value that provides the best explanation of the data (e.g., Edwards, 1972; Royall, 1997). Finally, the concept of explanation is also salient in cognitive psychology: explanation-based reasoning affects the way people learn categories, generalize properties and draw inferences (Rips, 1989; Thagard, 1989; Lombrozo, 2006, 2009, 2012).

Explanation is closely related to other important concepts in scientific reasoning, such as prediction, causation, and unification. Explanations differ by discipline and context: phenomena are deduced from natural laws, unified by new and general theories, produced by causal mechanisms, or predicted by statistical models (e.g., Hempel, 1965; Salmon, 1971, 1984; van Fraassen, 1980; Machamer et al., 2000; Lipton, 2004; Woodward, 2014). There is also a general tension between accounts of explanation that emphasize the predictive value of an explanation for a phenomenon (e.g., Hempel and Oppenheim, 1948) and those that stress that explanations provide genuine understanding (e.g., de Regt and Dieks, 2005). Given this wide variety of explanatory reasoning, it is not easy to give a convincing analysis of scientific explanation that transcends the scope of a particular context and unifies reasoning in different disciplines. A plurality of accounts of scientific explanation may be more realistic than a single account that purports to capture all aspects of scientific explanation (Colombo, 2016; Colombo and Wright, 2016).

Consequently, the place of Bayesian reasoning in a theory of explana-

tion will be different than in the case of confirmation, which is essentially captured by a probabilistic explication. Rather, our method will mimic the previous variation on causal effect. There, we described the quantitative dimension of causal judgments—the size of a causal effect—by means of a Bayesian formalism without committing us to a particular qualitative theory of causation. Similarly, in this variation, we focus on a Bayesian explication of **explanatory power**, that is, the degree to which a hypothesis explains a phenomenon, without trying to give a complete Bayesian account of scientific explanation. In particular, we will not make any attempts to reduce explanation to probabilistic relationships. Rather, we show how explanatory power can be explicated within a broadly Bayesian approach to scientific reasoning, and how such an explication can fruitfully inspire further research on scientific explanation.

There is another reason for why a full Bayesian theory of explanation may be difficult to achieve, namely an intrinsic tension between Bayesian and explanatory inference. Philosophers such as Gilbert Harman (1965) and Peter Lipton (2004) have defended **Inference to the Best Explanation (IBE)** as a rational mode of inference: a hypothesis is inferred on the basis of its explanatory virtues. For example, evolutionary psychologists explain features of human behavior and cognition, such as differences in mating or reasoning patterns between males and females, by environmental adaptations evolved during the Pleistoscene (Buss and Schmitt, 1993, e.g.,). Specific theories, such as Parental Investment Theory or Sexual Selection Theory (Buss, 1998; Miller, 1998, 2000), are inferred on the basis of their ability to explain such differences by means of evolutionary stories.

However, if we have no further cues why the theory in question may be empirically adequate, inferring theories on the basis of their explanatory value may just lead us to just-so-stories or improbable conclusions. Certainly we should not infer any implausible story about human life in the Pleistoscene just on its basis to explain features of current behavior (e.g., Gould and Lewontin, 1979). More generally, what are the circumstances where explanatory and Bayesian inferences agree? This question is still open, as witnessed by a lively debate about whether IBE is compatible with, and can be framed in Bayesian terms (van Fraassen, 1989; Okasha, 2000; Lipton, 2001; Salmon, 2001; Schupbach, 2011b, 2016). For this reason, a reduction of explanatory to Bayesian reasoning is at least difficult to achieve.

This variation is structured as follows. First, we motivate why explanatory power should be captured by a statistical relevance measure. In this context, we also compare two different approaches to explanatory power: those motivated by probabilistic theories of causality and those motivated by a structural analogy between prediction and explanation (Section 7.1). Second, we compare different statistical relevance explications of explanatory power and develop arguments for a particular measure (Section 7.2). Finally, we sketch projects for future research on the integration of probabilistic and explanatory inference (Section 7.3).

Toward a Statistical Relevance Account of Explanatory Power

The first characterization of explanatory, abductive inference in terms of probabilistic reasoning can be found in the writings of the American pragmatist philosopher C.S. Peirce:

> Long before I first classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis—which is just what abduction is—was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts or some of them. The form of inference, therefore, is this:
>
> - The surprising fact, E, is observed;
> - But if H were true, E would be a matter of course;
> - Hence, there is reason to suspect that H is true. (Peirce, 1931)

Peirce's characterization contains two crucial premises: First, the phenomenon E is surprising, or expressed in probabilistic terms: $p(\mathrm{E})$ is small. Second, given H, E is "a matter of course", that is, $p(\mathrm{E}|\mathrm{H})$ is close to unity. If these premises are satisfied, Peirce concludes that "there is reason to suspect that H is true"—not necessarily a conclusive reason, but at least *some* reason to accept H. In other words, it is crucial to explanatory inference that the surprising fact E is rationalized by H. This feature of good explanations is also stressed a couple of decades later by Carl G. Hempel:

> [T]he [explanatory] argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*; and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred. (Hempel, 1965, 337, original emphasis)

And, one page later:

> the explanatory information must provide good grounds for believing that X [the explanandum] did in fact occur; otherwise, that information would give us no adequate reason for saying: "That explains it – that does show why X occurred." (Hempel, 1965, 368)

Explanation thus has a central epistemic function—namely to resolve the epistemic puzzle surrounding the explanandum, to make it a matter of course given the explanans. For a recent defense of the predictive value of explanations, see Douglas (2009a).

There is, however, a subtle difference between Peirce and Hempel. Peirce explicitly stresses that E must have been surprising beforehand, Hempel doesn't—at least not explicitly. This difference corresponds to the choice between two different types of probabilistic explications of explanatory power: one that focuses on the statistical relevance of H for E (e.g., by comparing $p(\mathrm{E}|\mathrm{H})$ and $p(\mathrm{E})$), and another that focuses solely on the degree to which E is expected given H. This is, by the way, the same choice that we have already faced in Variations 2 and 6. There, we distinguished confirmation as firmness from confirmation as increase in firmness, and measures of causal production from measures of counterfactual dependence.

Joseph Halpern and Judea Pearl, two renowned researches on causal inference, have proposed an explication of explanatory power that pursues the second option (Halpern and Pearl, 2005a,b). They observe that causation and explanation are a hard-to-separate couple in scientific reasoning. Indeed, the most natural and intuitive account of explanation is an account where phenomena are explained by their causes, e.g., certain mechanisms: causes give an account of how and why the explanandum was produced. Failure of the brakes explains a car accident. Poison in the food explains the death of the king. Exposure to violent video games explains aggressive behavior. In all of these examples, causal efficacy grounds explanatory

power. So perhaps it is not surprising that until the early 20th century, the concept of explanation was subordinate to the concept of causation, e.g., in the writings of David Hume or Immanuel Kant.

In line with these intuitions, the causal theory of explanation sees the role of explanations in tracing causal processes and interactions leading to the explanandum (e.g., Salmon, 1984; Dowe, 2000; Strevens, 2009). More precisely, "the role of explanation is to provide the information needed to establish causation" (Halpern and Pearl, 2005b, 897). Halpern and Pearl use the interventionist account of causality, presented in the previous variation, to redefine the notion of explanatory power: "we view an explanation as a fact that is not known for certain but, if found to be true, would constitute a genuine cause of the explanandum, regardless of the agent's initial uncertainty" (ibid.). Thereby, Halpern and Pearl relativize an explanation to the epistemic state of an agent and introduce a pragmatic, subject-dependent component. They then define that the value of a certain variable $C = x$ counts as an explanation of some fact E roughly if and only of (i) E holds true in all contexts that the agent regards as possible; (ii) $C = x$ is a cause of E; (iii) there are possible contexts where the explanation is false. The last clause serves to rule out vacuous explanations. The goodness of such an explanation is then quantified by the probability $p(\mathrm{E}|do(C = x))$, that is, the conditional probability of E given that C takes value x. Numerically, the Halpern-Pearl measure is identical to Pearl's measure of causal effect η_{SP} that we reviewed in Variation 6.

On the other hand, in the same sense that Halpern and Pearl's account allows for probability-lowering (actual) causes, it allows for probability-lowering explanations. This feature is very unintuitive—explanations should, as noted by Peirce, make a positive difference to the phenomenon they are trying to explain. In particular, the explained phenomenon should not be less likely under the explanans than under the alternative hypotheses. For (actual) causation, this is not necessarily problematic: if a football player hits the ball badly and the ball ends up in the goal nonetheless (e.g., because the goalie was unprepared), his shot has still caused the goal, even if it lowered the probability of a goal compared to a proper shot. But we would be hesitant to say that the player's bad shot *explains* the goal. On this view of explanations, which we will adopt in the remainder, explanations are (probabilistic) arguments in favor of the explanandum; they are prima facie reasons to accept the explanans.

Moreover, the Halpern-Pearl account regards explanation as being secondary to causation. As a consequence, some pragmatic aspects of explanation get out of sight. Bas van Fraassen (1980) illustrates this point with the famous flagpole story from Bromberger (1965). We can explain the length of the shadow of a flagpole by the height of the flagpole, conditional on the angle of the sun. Prima facie, the reverse explanation does not work: the secondary phenomenon, the length of the shadow, does not explain the primary phenomenon, the height of the flagpole. But such judgments depend on pragmatic factors, as van Fraassen pointed out: in a specific context, the height of the flagpole could be explained by the fact that it was manufactured to cast a shadow of a certain length at a certain time of the day. Sundials work in this way, for example. Depending on the context, both explanations can be acceptable, although only one of them is properly causal—the other one is functional.

There is no apparent reason why one of these two explanations should be preferred across the board. After all, there is a great plurality of explanatory reasoning in science, with modes of explanation as different as mathematical explanation (Colyvan, 2001), functional and mechanistic explanation (Machamer et al., 2000; Craver, 2007) and unification (Friedman, 1974; Kitcher, 1981). It is therefore natural to conceive of explanation as a cognitive phenomenon rather than something that is instantiated in the real world (e.g., by means of a causal relationship). This observation supports the plausibility of the approach to conceptualize scientific explanations as arguments. Then, the power of an explanation may be measured by the degree to which the explanans rationalizes the explanandum.

In the following section, we ask the question how to quantify the explanatory power of H with respect to E. We are only concerned with quantifying the strength of that explanation and presuppose that H qualifies as an acceptable explanation of E. This is, of course, no reductive analysis of explanation, but it allows us to focus on the "grammar" of explanatory power without committing ourselves to a specific (and possibly problematic) view on the nature of explanation. At the same time, the question of explicating the concept of explanatory power is complex enough that even in the simple case that H is an undisputed explanation of E, there is sufficiently much room for disagreement about the degree of explanatory power.

Explicating Explanatory Power

Our basic idea, faithful to the principles of Bayesian philosophy of science, is to explicate explanatory power as a function of the joint probability distribution of E and H. This is actually in line with Peirce's rationale that explanation proceeds by making a "surprising fact" a "matter of course". Potentially, this can be extended to a causal-explanatory calculus where, instead of conditional probabilities such as $p(\mathrm{E}|\mathrm{H})$, we reason with counterfactual probabilities such as $p(\mathrm{E}|do(\mathrm{H}))$ and $p(\mathrm{E}|do(\neg\mathrm{H}))$, that is, the probability of E given a causal intervention on the putative explanation. We leave this question open since it does not directly affect our discussion of the various measures of explanatory power.

Like in previous variations, we assume that E and H are among the closed sentences $\mathfrak{L}$ of a first-order language L. Analogous to a confirmation measure, a measure of explanatory power is described by a function $\mathcal{E} : \mathfrak{L}^2 \times \mathfrak{P} \to \mathbb{R}$, where $\mathfrak{P}$ is the set of probability measures on the σ-algebra generated by $\mathfrak{L}$. This function assigns a real-valued degrees of explanatory power $\mathcal{E}(\mathrm{E}, \mathrm{H})$ to any pair of sentences in $\mathfrak{L}$, together with a probability measure p. For the sake of simplicity, we will omit reference to background assumptions and assume that they are implicit in the probability function p.

Three measures of explanatory power have been advanced and discussed in recent years. We shall now present them together with the corresponding representation theorems before delving into the issue of comparing the three. We omit Popper's (2002) measure $\mathcal{E}(\mathrm{E}, \mathrm{H}) = (p(\mathrm{E}|\mathrm{H}) - p(\mathrm{E}))/(p(\mathrm{E}|\mathrm{H}) + p(\mathrm{E}))$ since he provides no independent motivation, and the phrase "explanatory power" is used in a heuristic sense only, in the context of explicating a measure of degree of corroboration.

Among the remaining candidates, the oldest measure in the debate is the one proposed by I.J. Good (1960) and Timothy McGrew (2003):

$$\mathcal{E}_{GMG}(\mathrm{E}, \mathrm{H}) = \log \frac{p(\mathrm{E}|\mathrm{H})}{p(\mathrm{E})} \tag{7.1}$$

The Good-McGrew measure allows for an axiomatic representation, given by Cohen (2016b). To this end, we need to define a number of conditions:

Formality There is a function g such that, for any $\mathrm{E}, \mathrm{H} \in \mathcal{L}$ and any $p \in \mathfrak{P}$, $\mathcal{E}(\mathrm{E}, \mathrm{H}) = g(p(\mathrm{E} \wedge \mathrm{H}), p(\mathrm{E}), p(\mathrm{H}))$.

Formality captures the idea that $\mathcal{E}$ is a function of the joint probability distribution of E and H. The same idea has been applied successfully to measures of confirmation in Variation 2. Next, we have a statistical relevance condition which rules out probability-lowering explanations:

Statistical Relevance For any $\mathrm{E}, \mathrm{H}_1, \mathrm{H}_2 \in \mathcal{L}$ and any $p \in \mathfrak{P}$, $\mathcal{E}(\mathrm{E}, \mathrm{H}_1) > \mathcal{E}(\mathrm{E}, \mathrm{H}_2)$ if and only if $p(\mathrm{E}|\mathrm{H}_1) > p(\mathrm{E}|\mathrm{H}_2)$.

This condition states, in other words, that among two competing explanations for the same phenomenon, we should prefer the one which rationalizes the explanandum to a higher degree. Note that this is not entirely uncontroversial since according to several philosophers of science working on explanation (e.g., Okasha, 2000; Lipton, 2004; Schupbach, 2016), degree of explanatory power may depend on the goodness of an explanation *and* its plausibility. According to Statistical Relevance, however, even a very unlikely explanation is preferred to a likely one, as long as it is better at explaining the data.

The following condition is familiar from the explication of degree of confirmation in Variation 2. It forges a link between explanatory power and degree of confirmation: H explains E_1 better than E_2 if and only if E_1 raises the probability of H to a higher level than E_2 does.

Final Probability Incrementality For any $\mathrm{E}_1, \mathrm{E}_2, \mathrm{H} \in \mathcal{L}$ and any $p \in \mathfrak{P}$, $\mathcal{E}(\mathrm{E}_1, \mathrm{H}) > \mathcal{E}(\mathrm{E}_2, \mathrm{H})$ if and only if $p(\mathrm{H}|\mathrm{E}_1) > p(\mathrm{H}|\mathrm{E}_2)$.

According to this condition, explanatory power is structurally similar to degree of confirmation in so far as a candidate explanans H performs best on those phenomena that are also statistically relevant for it. From these assumptions, Cohen (2016b) derives the following representation theorem:

Theorem 7.1 *Formality, Statistical Relevance, and Final Probability Incrementality hold for a measure of explanatory power $\mathcal{E}(\mathrm{E}, \mathrm{H})$ if and only if there is a strictly increasing function $f : \mathbb{R} \to \mathbb{R}$ such that for any $\mathrm{E}, \mathrm{H} \in \mathcal{L}$ and any $p \in \mathfrak{P}$, $\mathcal{E}(\mathrm{E}, \mathrm{H}) = f(\mathcal{E}_{GMG}(\mathrm{E}, \mathrm{H}))$.*

In other words, the above conditions characterize $\mathcal{E}_{GMG}$ uniquely, up to ordinal equivalence. Cohen's representation theorem transposes the result by Crupi et al. (2013), cited in Variation 2, from degree of confirmation to explanatory power. In the original paper, the same conditions are imposed in order to derive $r(\mathrm{H}, \mathrm{E}) = p(\mathrm{H}|\mathrm{E})/p(\mathrm{H}) = p(\mathrm{E}|\mathrm{H})/p(\mathrm{E})$ as a measure of degree of confirmation.

A similar representation result can be derived for a measure proposed by Crupi and Tentori (2012). It takes the form

$$\mathcal{E}_{CT}(\mathrm{E},\mathrm{H}) = \begin{cases} \frac{p(\mathrm{E}|\mathrm{H})-p(\mathrm{E})}{1-p(\mathrm{E})} & \text{if } p(\mathrm{E}|\mathrm{H}) \geq p(\mathrm{E}) \\ \frac{p(\mathrm{E}|\mathrm{H})-p(\mathrm{E})}{p(\mathrm{E})} & \text{if } p(\mathrm{E}|\mathrm{H}) < p(\mathrm{E}) \end{cases} \tag{7.2}$$

For the $\mathcal{E}_{CT}$-measure, the following condition is characteristic:

Explanatory Justice If E′ is statistically independent from E, H, and their conjunction E ∧ H, then:

i) if $\mathcal{E}(\mathrm{E},\mathrm{H}) > 0$, then $\mathcal{E}(\mathrm{E} \wedge \mathrm{E}',\mathrm{H}) < \mathcal{E}(\mathrm{E},\mathrm{H})$; and

ii) if $\mathcal{E}(\mathrm{E},\mathrm{H}) \leq 0$, then $\mathcal{E}(\mathrm{E} \wedge \mathrm{E}',\mathrm{H}) = \mathcal{E}(\mathrm{E},\mathrm{H})$.

This condition is substantial and shall be the subject of debate later on. The first clause of Explanatory Justice is taken from Schupbach and Sprenger (2011, 115, notation changed), who motivate it as follows:

> "[A]s the evidence becomes less statistically relevant to some explanatory hypothesis H (with the addition of irrelevant propositions), it ought to be the case that the explanatory power of H relative to that evidence approaches the value at which it is judged to be explanatorily irrelevant to the evidence ($\mathcal{E} = 0$)."

Schupbach and Sprenger transfer this property to the case of negative statistical dependence: addition of statistically independent evidence *dilutes* (negative) explanatory power and brings it closer to the neutral value of zero. Crupi and Tentori, on the other hand, think that this property would allow "to indefinitely relieve a lack of explanatory power, no matter how large, by adding more and more irrelevant explananda, simply at will". (Crupi and Tentori, 2012, 370). Hence their demand for the second clause of Explanatory Justice. See also the discussion of No Dilution for Irrelevant Effects in Variation 6.

The Crupi-Tentori measure $\mathcal{E}_{CT}$ also satisfies a similar constraint regarding the relation between positive and negative explanatory power:

Symmetry For any $\mathrm{E}_1, \mathrm{E}_2, \mathrm{H} \in \mathcal{L}$ and any $p \in \mathfrak{P}$, $\mathcal{E}(\mathrm{E}_1,\mathrm{H}) > \mathcal{E}(\mathrm{E}_2,\mathrm{H})$ if and only if $\mathcal{E}(\neg\mathrm{E}_1,\mathrm{H}) < \mathcal{E}(\neg\mathrm{E}_2,\mathrm{H})$.

That is, if H explains E_1 better than E_2, then it also explains $\neg E_2$ better than $\neg E_1$. When Explanatory Justice and Symmetry replace Final Probability Incrementality in Theorem 7.1, this suffices for demonstrating another representation result (Crupi and Tentori, 2012; Cohen, 2016b):

Theorem 7.2 *Formality, Statistical Relevance, Explanatory Justice and Symmetry hold for a measure of explanatory power* $\mathcal{E}(E,H)$ *if and only if there is a strictly increasing function* $f : \mathbb{R} \to \mathbb{R}$ *such that for any* $E, H \in \mathcal{L}$ *and any* $p \in \mathfrak{P}$, $\mathcal{E}(E,H) = f(\mathcal{E}_{CT}(E,H))$.

The primacy of surprise-lowering over the acceptability of the explanation, as coded in Statistical Relevance, is an important characteristic feature of both the Good-McGrew and the Crupi-Tentori measure. It is not shared by the third measure in the debate, proposed by Schupbach and Sprenger (2011). Their measure has the form

$$\mathcal{E}_{SS}(E,H) = \frac{p(H|E) - p(H|\neg E)}{p(H|E) + p(H|\neg E)}. \tag{7.3}$$

This measure can be derived in different ways. Schupbach and Sprenger's original derivation is based on four conditions. The first one is a variation of the Formality condition, which describes their measure as a function of the conditional probabilities $p(H|E)$, $p(H|\neg E)$, and $p(E)$:

Formality* There is a function g such that, for any $E, H \in \mathcal{L}$ and any $p \in \mathfrak{P}$, $\mathcal{E}(E,H) = g(p(E), p(H|E), p(H|\neg E))$.

The second one is a weakened version of Statistical Relevance:

Statistical Relevance* The function $g(p(E), p(H|E), p(H|\neg E))$ from Formality* is not constant in the two latter arguments. That is, there is no function $h : [0,1] \to \mathbb{R}$ such that $g(x,y,z) = h(x)$.

The intuitive idea behind this condition is that explanatory power should be sensitive to probabilistic relations between E and H and not be a function of the unconditional probability of the explanandum only. Next, we have

Irrelevant Conjunctions If H_2 is statistically independent of E, H_1 and E $\wedge$ H_1, then $\mathcal{E}(E,H_1) = \mathcal{E}(E,H_1 \wedge H_2)$.

This condition is similar to the Modularity constraint for measures of confirmation (p. 71). When a scientific hypothesis is irrelevant to a certain

explanandum and a putative explanans, adding it to the explanans neither increases nor decreases the degree of explanatory power. Unified theories just seem to explain a phenomenon as well as that part of the theory that did the explanatory work. Or in other words, embedding an explanation into a general framework leaves its explanatory power for a phenomenon in its original domain unchanged.

Finally, there is the condition

Deductive Entailment If $\neg$H entails E, then $\mathcal{E}(\mathrm{E},\mathrm{H})$ is not sensitive to the values of $p(\mathrm{H})$, ceteris paribus.

While intuitions may not be strong enough to make this condition a compelling constraint on all measures of explanatory power, it does not seem to be implausible or harmful either. One might argue that if $\neg$H is already a perfect explanation of the explanandum, then the (negative) explanatory power of H has nothing to do with its prior probability, but just with the degree to which H accounts for E.

Based on these conditions, Schupbach and Sprenger (2011, Theorem 1) prove the following representation theorem:

Theorem 7.3 *Formality*, Statistical Relevance*, Irrelevant Conjunctions and Deductive Entailment hold for a measure of explanatory power* $\mathcal{E}(\mathrm{E},\mathrm{H})$ *if and only if there is a strictly increasing function* $f : \mathbb{R} \to \mathbb{R}$ *such that for any* $\mathrm{E},\mathrm{H} \in \mathcal{L}$ *and any* $p \in \mathfrak{P}$, $\mathcal{E}(\mathrm{E},\mathrm{H}) = f(\mathcal{E}_{SS}(\mathrm{E},\mathrm{H}))$.

It should be noted that $\mathcal{E}_{SS}$ also satisfies the Symmetry and Statistical Relevance conditions (proof omitted), but not Final Probability Incrementality and only the first, uncontroversial clause of the Explanatory Justice condition. The second clause of Explanatory Justice is a major point of contention in the debate about different measures of explanatory power, as evidenced by Crupi and Tentori (2012) and Cohen (2015, 2016b,a). These papers also offer different and somewhat simpler representation theorems for $\mathcal{E}_{SS}$. However, the price they pay is that the assumptions have to be strengthened a bit. The most interesting alternative characterization (up to ordinal equivalence) is due to Cohen (2015) and consists of three conditions: (i) all tautological hypotheses receive constant explanatory power; (ii) the strong symmetry condition $\mathcal{E}(\neg\mathrm{E},\mathrm{H}) = -\mathcal{E}(\mathrm{E},\mathrm{H})$; (iii) a somewhat stronger version of Deductive Entailment.

We now proceed to a **normative comparison of the three measures** and begin with a critique of the Good-McGrew measure $\mathcal{E}_{GMG}$. For

starters, we note that it allows for the **conjunction of irrelevant evidence** (Schupbach and Sprenger, 2011, 114–115). Suppose that for some piece of evidence E′, $p(\mathrm{E}'|\mathrm{E},\mathrm{H}) = p(\mathrm{E}'|\mathrm{H})$. In that case, $\mathcal{E}_{GMG}(\mathrm{E}\wedge\mathrm{E}',\mathrm{H}) = \mathcal{E}_{GMG}(\mathrm{E},\mathrm{H})$. Schupbach and Sprenger consider this property—possessed by neither their measure $\mathcal{E}_{SS}$ nor the Crupi-Tentori measure $\mathcal{E}_{CT}$—problematic and illustrate their objection with an example. Let E be an observed Brownian motion, let H be an appropriate physical explanation of that motion, and let E′ be a proposition about the mating season of the American tree frog. Clearly, H explains E much better than it explains E $\wedge$ E′—the Brownian motion *and* the tree frog mating season proposition. A substantial part of E $\wedge$ E′ stays unexplained.

This criticism echoes the paradox of irrelevant conjunctions that we have encountered in Variation 2, applied to the ratio measure $r(\mathrm{H},\mathrm{E})$. In that case, r allowed for tacking additional (irrelevant) hypotheses without lowering the degree of confirmation. This was taken as a reason to rule out r as an appropriate measure. The same argument pattern applies here: tacking irrelevant conjunctions to the explanandum should *lower* the degree of explanatory power and not leave it constant.

In defense of $\mathcal{E}_{GMG}$, Cohen (2016b) notes that the bite of Schupbach and Sprenger's objection depends on whether E′ is meant to be explained by H or not. For example, if E′ is just some extra data obtained in an experiment (e.g., demographic data in a psychological survey), then it seems that H should not be penalized for failing to explain E′. Whether the addition of irrelevant evidence is problematic seems to depend on the focus of the explanation: is H supposed to explain all of the evidence or just the part that we consider crucial?

We are not sure that this observation rescues $\mathcal{E}_{GMG}$. To accommodate this kind of context-sensitivity, we would rather conceive of explanatory power as a ternary relation between explanans, explanandum and additional data. What we argue here is that explanatory power should not be invariant under adding data that are *part of the explanandum*, but fail to be rationalized. Hence, the Good-McGrew $\mathcal{E}_{GMG}$ measure remains problematic.

What about the other two measures? Should $\mathcal{E}_{SS}$ or $\mathcal{E}_{CT}$ be preferred? Of course, we are not completely unbiased in answering this question: one of the authors of this monograph (J.S.) developed the $\mathcal{E}_{SS}$-measure together with Jonah Schupbach. We will now advance two arguments in favor of

$\mathcal{E}_{SS}$, both of them due to Cohen (2016a).

The first argument concerns the **scaling properties** of both measures. It is based on a simple coin-flipping example. There are two identical-looking coins, one of which is fair while the other one is biased (say, with a 70/30 bias in favor of heads). We test one of the two coins, but do not know which one and we consider both cases equally probable. Now consider the hypothesis H that the tested coin is biased and the event E_N that all N tosses of the coin turn out to be heads. Certainly, this hypothesis explains E_N to a certain degree—primarily because that course of events would be a truly extraordinary chance under the hypothesis $\neg$H.

However, the Crupi-Tentori measure disagrees: as N increases, $\mathcal{E}_{CT}(\mathrm{E_N}, \mathrm{H})$ quickly approaches zero (e.g, $\mathcal{E}_{CT}(\mathrm{E}_{10}, \mathrm{H}) = 0.014$). In other words, $\mathcal{E}_{CT}$ treats a statistically highly relevant hypothesis as if it were independent of the explanandum. E is surprising under H, but it is much more surprising under $\neg$H, a fact that is not reflected by $\mathcal{E}_{CT}$. By contrast, Schupbach and Sprenger's measure $\mathcal{E}_{SS}$ converges to a reasonable, but not too high value ($\mathcal{E}_{SS}(\mathrm{E_N}, \mathrm{H}) \overset{N\to\infty}{\longrightarrow} 0.33$), indicating that H outperforms $\neg$H while being a far from perfect explanation. In other words, $\mathcal{E}_{SS}$ captures the contrastive nature of scientific explanations (van Fraassen, 1980) better than $\mathcal{E}_{CT}$.

The second and most stringent criticism is based on how $\mathcal{E}_{CT}$ deals with **irrelevant evidence**. If (negative) explanatory power remains constant under the addition of irrelevant evidence, as the downward clause of Explanatory Justice demands, then Crupi and Tentori should also believe that $\mathcal{E}_{CT}$ remains constant under the addition of irrelevant disjunctions to the hypothesis. Hence, they should require that $\mathcal{E}_{CT}(\mathrm{E}, \mathrm{H}) = \mathcal{E}_{CT}(\mathrm{E}, \mathrm{H} \vee \mathrm{H}')$ whenever H′ is statistically independent of E, H and E $\wedge$ H and $p(\mathrm{E}|\mathrm{H}) < p(\mathrm{E})$. However, Cohen (2016a, Claim 1) shows that in this case, $\mathcal{E}_{CT}(\mathrm{E}, \mathrm{H}) < \mathcal{E}_{CT}(\mathrm{E}, \mathrm{H} \vee \mathrm{H}')$. This leads the entire idea of Explanatory Justice that motivated the Crupi-Tentori measure, ad absurdum: explanatory power is not increased for E $\wedge$ E′, but it is increased for H $\vee$ H′. This internal inconsistency can be construed as an argument in favor of the Schupbach-Sprenger measure $\mathcal{E}_{SS}$. Those who fancy the Crupi-Tentori measure may try to evade this objection by making suitable modifications for the case of negative explanatory power. This is a topic for future research, though.

Discussion

This variation motivated, presented and compared various Bayesian accounts of explanatory power. For starters, the two grand traditions for conceiving of scientific explanation—the view of explanations of arguments and the causal-interventionist view—have been introduced and discussed. In the light of that exposition, it seems that none of the views captures completely what scientific explanations are about; yet, each of the views retains important features of scientific explanation that can be used for an explication of explanatory power.

The core of this variation has been the derivation and comparison of three different Bayesian measures of explanatory power. In our view, the results favor the Schupbach-Sprenger measure $\mathcal{E}_{SS}$. The competitors, the Good-McGrew measure $\mathcal{E}_{GMG}$ and the Crupi-Tentori measure $\mathcal{E}_{CT}$, are haunted by general objections pertaining to their functional form that are, at least as things stand now, not easy to answer. Context-specific considerations may have the last word in each application, however.

Another dimension of investigating measures of explanatory power consists in empirical work. In an experiment that transfers the design of Crupi et al. (2007) to the case of explanatory power, (Schupbach, 2011a) has found out that $\mathcal{E}_{SS}$ best describes participants' judgments of explanatory power. For methodological criticism of this design and the statistical analysis, see Glymour (2015). Recent experiments on explanatory power and related cognitive values (e.g., Colombo et al., 2016a,b) confirm that explanatory judgment is sensitive to statistical relevance, lending empirical support to the Bayesian research program on explanatory reasoning and explanatory power. The above studies also revealed a strong link between judgments of explanation, confirmation and rational acceptability. Furthermore, Lombrozo (2007) has investigated how the simplicity of an explanation affects its perceived value. Future studies could transfer this design from Lombrozo's artificial and idealized scenario, involving inhabitants of an alien planet, to an ecologically more valid setting.

More specifically, measures of explanatory power can help to construct a Bayesian account of Inference to the Best Explanation, and to develop a mathematically precise version of IBE. From a descriptive point of view, recent empirical work has shown that people accept hypotheses rather on the basis of their explanatory value than on the basis of objective chances

Douven and Schupbach (2015b,a). This motivates further empirical research into the circumstances under which people's reasoning conforms to IBE. From a normative point of view, Schupbach (2011b, 2016) has used computer simulations in order to show that IBE—conceptualized as inference to the hypothesis with the highest explanatory power—is a reliable mode of inference. Peirce's inference scheme (E, H explains E $\Rightarrow$ H) is replaced by the scheme (E, $\mathcal{E}(\mathrm{E},\mathrm{H}) > \mathcal{E}(\mathrm{E},\mathrm{H_i})$ for all alternatives $\mathrm{H}_i \Rightarrow$ H). This sophisticated form of IBE approximates Bayesian reasoning very well, and in Schupbach's simulations, the explanatorily most valuable hypothesis matches the true hypothesis in an overwhelming number of cases. More research along these lines may help to determine the conditions under which IBE is a sound form of scientific reasoning, and to shed light on issues where IBE takes a prominent role, such as the ongoing debate between realists and anti-realists.

Finally, there is ample room for combining empirical and theoretical research on explanatory inference in the Bayesian paradigm. A particularly salient issue concerns the role and interplay of causal and probabilistic factors in explanatory reasoning. One could, for example, envision a systematic comparison of measures of causal effect and explanatory power. Is there an isomorphism between some measures of explanatory power and causal effect, based on their joint interest in the predictive power of the explanans (=the cause) for the explanandum (=the effect)? How do explanatory, causal and probabilistic reasoning interact (Lombrozo, 2009, 2011, 2012; Sloman and Lagnado, 2015)? Do these differences have correlates on the level of a formal Bayesian analysis?

We hope that our contribution will stimulate further research on the nature of explanation. In particular, we hope that our results will help to promote "the prospects for a naturalized philosophy of explanation" (Lombrozo, 2011, 549), where philosophical theorizing about the nature of explanation is constrained and informed by empirical evidence about the psychology of explanatory power and where, on the other hand, philosophical research stimulates empirical investigations into explanatory reasoning.

Variation 8: Intertheoretic Reduction

Establishing relations between different theories is an important goal of science. Unified theories with a wide scope and a small number of basic postulates have been found attractive by scientists at all times. Take, for example, Newtonian mechanics which can be used to explain terrestial as well as celestial motion, unifying Galilei's invariance principle with Kepler's Laws of planetary motion. Or consider Maxwell's theory of electrodynamics which provides a unified account of electric and magnetic forces, and the laws governing their interaction. There is also the famous example of statistical mechanics, whose micro-level laws about the motion of molecules provide the foundations for a macro-level theory about the behavior of gases and fluids, namely thermodynamics.

The relation between statistical mechanics and thermodynamics is special because it is a paradigm example of **intertheoretic reduction**: accounting for the behaviour of a system at a certain level of organization by describing the behavior of its constituents. What exactly is involved in a reduction is a matter of philosophical controversy (see van Riel and Van Gulick, 2014). The basic idea is that the concepts and **laws of a phenomenological theory $\mathbf{T_P}$**, such as thermodynamics, are "reduced" to **laws of a more fundamental theory $\mathbf{T_F}$**, such as statistical mechanics. Often this reduction is executed by means of deriving the laws of $\mathbf{T_P}$ from those of $\mathbf{T_F}$ (Nagel, 1961; Schaffner, 1967)—more on this below. Following standard terminology, we say that $\mathbf{T_P}$ is the reduced theory and that $\mathbf{T_F}$ is the reducing theory. Other examples of (putative) intertheoretical reductions are chemistry to atomic physics, rigid body mechanics to particle mechanics, psychology to neuroscience, and agent-based modeling in the social sciences.

Reductions are, if successful, celebrated by scientists because they allow for a unified theoretical framework in which one can investigate the phenomenological as well as the fundamental theory. They also allow for

precise predictions on the phenomenological level motivated by assumptions on the fundamental level. They may provide some deep understanding into and explanation of the nature of central concepts of the involved disciplines. For instance, the thermodynamic concept of heat is identified with the energy transfer by a disordered, microscopic action on a system of molecules, described by statistical mechanics. For these reasons, intertheoretic reductions are taken to make large contributions to the cognitive advancement of science.

In this variation, we show how the establishment of intertheoretic reductions boosts the cognitive value of the involved theories by confirming them in the Bayesian sense. More specifically, we show that if there is a reductive relation between two theories, then confirmation flows both from the phenomenological to the fundamental theory and from the fundamental to the phenomenological theory. For instance, evidence that exclusively confirms statistical mechanics before the reduction also confirms (though perhaps to a lower degree) thermodynamics after the reduction, and vice versa.

Section 8.1 sets the scene by outlining the Generalized Nagel-Schaffner (GNS) model of reduction, which serves as the foil for our Bayesian analysis. Section 8.2 contains the main argument: we consider the confirmation of $\mathbf{T_F}$ and $\mathbf{T_P}$ in two scenarios: one with and one without intertheoretic reduction. We conclude that reduction boosts confirmation, and Section 8.3 discusses various implications of this result. In Section 8.4, we sum up our results and outline a number of open problems. Further detail is contained in the articles "Who is Afraid of Nagelian Reduction?" and "Confirmation and Reduction" by Dizadji-Bahmani et al. (2010, 2011), on which this variation is based.

The Generalized Nagel-Schaffner Model

For describing the idea behind the GNS model of reduction which motivates our later Bayesian analysis, it may be useful to begin with the familiar case of thermodynamics and statistical mechanics. Thermodynamics describes systems like gases and solids in terms of macroscopic properties such as volume, pressure, temperature and entropy, and gives a correct description of the behaviour of such systems. The aim of statistical mechanics is to account for the laws of thermodynamics in terms

of dynamical laws governing the microscopic constituents of macroscopic systems (Frigg, 2008). In particular, statistical mechanics aims to show that the Second Law of Thermodynamics is a consequence of the mechanical motion of the molecules of the gas. For example, consider a container divided in two by a partition wall. The left half is filled with a gas, while the right half is empty. If we now remove the partition, the gas will spread and soon be evenly distributed throughout the entire container; the gas's entropy increases as it spreads. This is an instance of a process obeying the Second Law of Thermodynamics. Roughly speaking, the Second Law says that the entropy of a closed system cannot decrease, and usually increases when the system is left on its own in a non-equilibrium state. The aim of statistical mechanics is to account for the Second Law in general in terms of the equations governing the motion of the molecules of the gas and some probabilistic assumptions; that is, it aims to show that the Second Law is a consequence of its basic postulates. Or almost so.

That analogues of the laws of the phenomenological theory (here: thermodynamics) should follow from the laws of the fundamental theory (here: statistical mechanics) is the basic idea of GNS. Consider a phenomenological theory $\mathbf{T_P}$ and a fundamental theory $\mathbf{T_F}$, which are identified with a set of empirical propositions. So let $\mathbf{T_P} := \{T_P^{(1)}, \ldots, T_P^{(n_P)}\}$ and $\mathbf{T_F} := \{T_F^{(1)}, \ldots, T_F^{(n_F)}\}$. The reduction of $\mathbf{T_P}$ to $\mathbf{T_F}$ consists of the following three steps (Schaffner, 1967):

1. Adopt auxiliary assumptions describing the particular setup under investigation. Here, these are assumptions about the mechanical properties of the gas molecules. Then derive from these and $\mathbf{T_F}$ a restricted version of each proposition $T_F^{(i)}$. Denote these by $T_F^{*(i)}$ and the corresponding set by $\mathbf{T_F^*} := \{T_F^{*(1)}, \ldots, T_F^{*(n_F)}\}$.

2. $\mathbf{T_P}$ and $\mathbf{T_F}$ are formulated in different vocabularies. In our example, statistical mechanics talks about trajectories in phase space and probability measures while thermodynamics talks about macroscopic properties such as pressure and temperature. In order to connect the two theories, we adopt bridge laws. These connect terms of one theory with terms of the other, for instance mean kinetic energy in statistical mechanics with temperature in thermodynamics. Substituting the terms in $\mathbf{T}_F^*$ with terms from $\mathbf{T_P}$ as per the bridge laws yields $\mathbf{T}_P^*$, i.e., the set $\{T_P^{*(1)}, \ldots, T_P^{*(n_P)}\}$.

3. Show that each element of $\mathbf{T}_P^*$ is strongly analogous to the corresponding element in $\mathbf{T_P}$.

Figure 8.1: The Generalized Nagel-Schaffner (GNS) model of reduction

If these conditions obtain, we say that $\mathbf{T_P}$ **is reduced to** $\mathbf{T_F}$. See Figure 8.1 for a graphical illustration.

We now explain two central notions that occur in the GNS model of intertheoretic reduction.

First, the notion of **strong analogy**, which may appear inappropriately vague. After all, Nagel himself has stressed the importance of logical and mathematical relations that hold between the reducing and the reduced theory. These strong links between $\mathbf{T_P}$ and $\mathbf{T_F}$ seem to be watered down by introducing a concept which introduces a great deal of subjective judgment on behalf of the scientist. However, it is often impossible to derive the *exact* laws of $\mathbf{T_P}$. For instance, it is not possible to derive the exact Second Law of Thermodynamics from statistical mechanics, which is a probabilistic theory, whereas the Second Law is supposed to hold without exception. Thus *exact* derivability is too stringent a requirement. It suffices to deduce laws that are *approximately the same* as the laws of $\mathbf{T_P}$. For the case of statistical mechanics and thermodynamics, we derive a probabilistic law that is strongly analogous to the Second Law of Thermodynamics: namely the proposition that entropy is highly likely to increase over time, which is known as Boltzmann's Law. This revision of the original model has been developed in a string of publications by Schaffner (1967, 1969, 1976, 1977, 1993), and, indeed, by Nagel (1979) himself. In sum, reduction is the deductive subsumption of a corrected version of $\mathbf{T_P}$ under $\mathbf{T_F}$, where the deduction involves first deriving a restricted version, $\mathbf{T_F^*}$, of the reducing theory by introducing boundary conditions and auxiliary assumptions and then using bridge laws to obtain $\mathbf{T_P^*}$ from $\mathbf{T_F^*}$.

This brings us to the second point: the notion of a **bridge law**. While Nagel himself remains relatively non-committal about the exact form and nature of bridge laws, Schaffner (1976, 614–615) offers a concise characterization of bridge laws, which he calls *reduction functions*. For Schaffner,

a reduction function is a statement to the effect that a term y_P of $\mathbf{T}^*_\mathbf{P}$ and a term y_F of $\mathbf{T}^*_\mathbf{F}$ are coextensional. For example, the terms "temperature" and "mean kinetic energy" are coextensional when applied to a gas (we come back to this qualification below). At least in physics, properties usually have magnitudes: A gas does not have a temperature *simpliciter*, it has a temperature of so and so many degrees Kelvin. Thus, a bridge law does not only establish coextensionality; it also specifies the functional relationship between the magnitudes of the terms and the units of measurements. That is, the bridge law contains a function f such that $\tau_P = f(\tau_F)$, where, respectively, τ_P and τ_F are the values of y_P and y_F. So we can give the following tentative definition of bridge laws (we will qualify this statement below): A bridge law is a statement to the effect that (i) y_P applies if, and only if, y_F applies, and (ii) $\tau_P = f(\tau_F)$.

Both the concept and the epistemology of strong analogy and bridge law have served as the basis for criticism of the GNS model of reduction. For instance, the so-called New Wave Reductionists (e.g., Churchland, 1979, 1985; Bickle, 1998) deny that bridge laws play an important role in the discovery of reductive relations. However, we do not want to engage (again) in a debate about the merits and drawbacks of the GNS model, but to show how it can be used for demonstrating the confirmatory value of reductive relations. Therefore we refer the interested reader to Dizadji-Bahmani et al. (2010), where these and similar criticisms are addressed and, to our mind, convincingly rebutted.

Reduction and Confirmation

Consider how theories are supported by evidence. With regard to our two theories $\mathbf{T_P}$ and $\mathbf{T_F}$, there are three kinds of evidence: evidence that only confirms the phenomenological theory, evidence that only confirms the fundamental theory and evidence that confirms, to some degree, both. We make this clear with examples from thermodynamics and statistical mechanics. For the first case, consider what is known as the Joule-Thomson process: there are two chambers of different dimensions connected to each other by a permeable membrane, filled with a gas. At the end of each chamber, there is a piston which allows the pressure and volume for the gas in each chamber to be varied by applying a force. The pressure in the first chamber is higher than the pressure in the second. Now push the

gas from the first chamber into the second, but so slowly that the pressure remains constant in both chambers and no heat is exchanged with the environment. Then, the gas in the second chamber cools down. The amount of cooling can be calculated using the principles of thermodynamics, and is found to coincide with experimental values. So we have a confirmation of thermodynamics, but not of statistical mechanics since no statistical mechanics assumptions have been used in the argument. For the second, consider the dependence of a metal's electrical conductivity on temperature. From statistical mechanics, one can derive an equation relating the change in the electrical conductivity of certain metals given a change in temperature which is what one finds in experimental thermodynamics, in contrast, is entirely silent about this phenomenon. Third, consider again the gas confined to the left half of the box which spreads evenly when the dividing wall is removed. It follows from thermodynamics that the thermodynamic entropy of the gas increases; at the same time, it is a consequence of statistical mechanics that the Boltzmann entropy increases in that process. So the spreading of the gas confirms both statistical mechanics and thermodynamics. We shall now explicate this intuition in a Bayesian model.

Before the Reduction

We examine the situation before a reduction is attempted. To simplify things, we assume that $\mathbf{T_P}$ and $\mathbf{T_F}$ have only one element, viz. $\mathrm{T_F}$ and $\mathrm{T_P}$ respectively. The generalization to more than one element is conceptually straightforward. Furthermore, E confirms $\mathrm{T_F}$ and $\mathrm{T_P}$, $\mathrm{E_F}$ only confirms $\mathrm{T_F}$ and $\mathrm{E_P}$ only confirms $\mathrm{T_P}$. Introducing corresponding propositional variables T_F, T_P, E , E_F and E_P, we can represent the situation before the attempted reduction in the Bayesian network depicted in Figure 8.2.

Following our methodology, we have to specify the prior probabilities of $\mathrm{T_F}$ and $\mathrm{T_P}$ (i.e., of all root nodes) and the conditional probabilities of E, $\mathrm{E_F}$ and $\mathrm{E_P}$ (i.e., of all child nodes), given their parents. We denote:

$$\begin{aligned}
p(\mathrm{T_F}) = t_F \quad &, \quad p(\mathrm{T_P}) = t_P \\
p(\mathrm{E_F}|\mathrm{T_F}) = p_F \quad &, \quad p(\mathrm{E_F}|\neg\mathrm{T_F}) = q_F \\
p(\mathrm{E_P}|\mathrm{T_P}) = p_P \quad &, \quad p(\mathrm{E_P}|\neg\mathrm{T_P}) = q_P \\
p(\mathrm{E}|\mathrm{T_F},\mathrm{T_P}) = \alpha \quad &, \quad p(\mathrm{E}|\mathrm{T_F},\neg\mathrm{T_P}) = \beta \\
p(\mathrm{E}|\neg\mathrm{T_F},\mathrm{T_P}) = \gamma \quad &, \quad p(\mathrm{E}|\neg\mathrm{T_F},\neg\mathrm{T_P}) = \delta
\end{aligned} \tag{8.1}$$

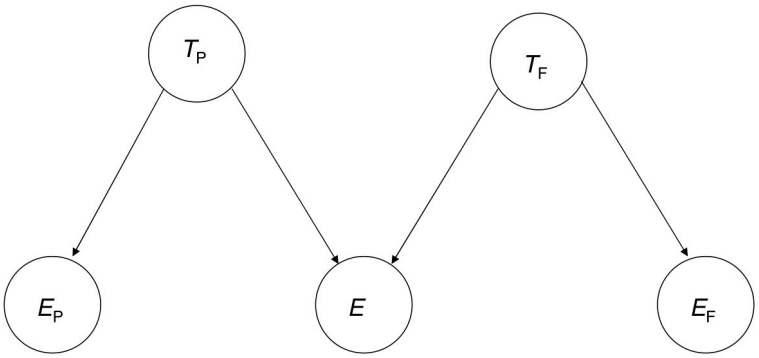

Figure 8.2: The Bayesian network representing the situation *before* the reduction.

These parameters cannot be freely chosen as we assume that the following conditions hold: First, $\mathrm{E_F}$ confirms $\mathrm{T_F}$, hence $p_F > q_F$. Second, $\mathrm{E_P}$ confirms $\mathrm{T_P}$, hence, $p_P > q_P$. Third, E confirms $\mathrm{T_F}$ and fourth E confirms $\mathrm{T_P}$. The last two conditions entail the following constraints on α, β, γ and δ (all proofs are in the final section):

$$(\alpha - \beta)\, t_F + (\gamma - \delta)\, \bar{t}_F \;\; > \;\; 0 \tag{8.2}$$

$$(\alpha - \gamma)\, t_P + (\beta - \delta)\, \bar{t}_P \;\; > \;\; 0 \tag{8.3}$$

These inequalities hold, for example, if $\alpha > \beta, \gamma > \delta$, which seems to be a natural condition. One may also want to require that $p(\mathrm{T_F}|\mathrm{E}, \mathrm{E_F}) > p(\mathrm{T_F})$ and $p(\mathrm{T_P}|\mathrm{E}, \mathrm{E_P}) > p(\mathrm{T_P})$. Note, however, that both inequalities follow from the above four conditions (proof omitted).

Given this network structure and the conditional independences encoded in it, it is easy to see, for example, that the variable E_F is independent of T_P given T_F and that E_P is independent of T_F given T_P. In symbols:

$$E_F \perp\!\!\!\perp T_P | T_F \quad , \quad E_P \perp\!\!\!\perp T_F | T_P \tag{8.4}$$

Hence, $\mathrm{E_F}$ does not confirm (or disconfirm) $\mathrm{T_P}$ and $\mathrm{E_P}$ does not confirm (or disconfirm) $\mathrm{T_F}$:

$$p(\mathrm{T_P}|\mathrm{E_F}) = p(\mathrm{T_P}) \quad , \quad p(\mathrm{T_F}|\mathrm{E_P}) = p(\mathrm{T_F}) \tag{8.5}$$

We conclude that there is no flow of confirmation from one theory to the other. The intuitive reason for this is that there is no chain of arrows from E_F to T_P. Note also that the variables T_F and T_P are probabilistically independent before the reduction:

$$p(\mathrm{T_F}, \mathrm{T_P}) = p(\mathrm{T_F})\, p(\mathrm{T_P}) = t_F\, t_P \tag{8.6}$$

All this may, however, not be right in practice. Scientists may feel, for example, that the two theories are much more intimately connected. An indication for this may be that there is, as we assume, evidence E that supports both theories. Another reason may be that there are formal (or other) relations between the two theories. In this case, scientists will attempt to reduce one theory to the other. Let us now model this situation.

After the Reduction

Recall the three steps involved in reducing one theory to another set out in Section 8.1: First, derive T_F^* from the auxiliary assumptions and T_F. Second, introduce bridge laws and obtain T_P^* from T_F^*. Third, show that T_P^* is strongly analogous to T_P.

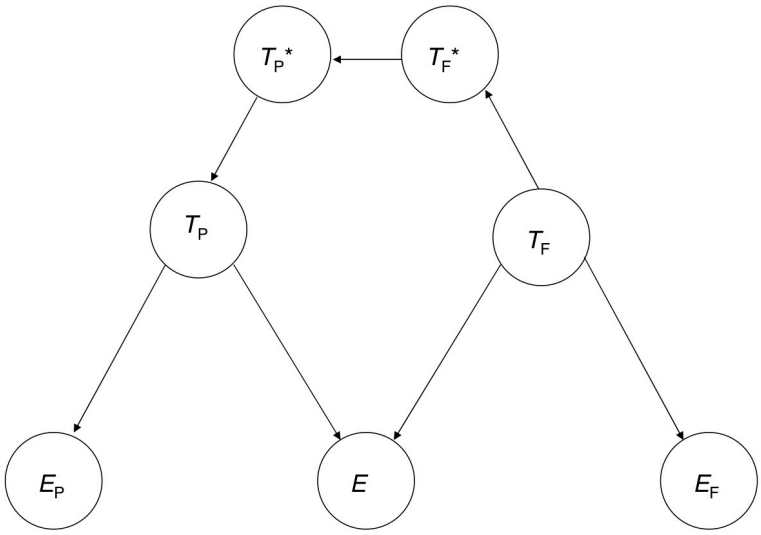

Figure 8.3: The Bayesian Network representing the situation *after* the reduction.

The situation after the reduction can then be represented in the Bayesian network depicted in Figure 8.3. To complete the network, we specify the following conditional probabilities:

$$p'(\mathrm{T_P}|\mathrm{T_P^*}) = p_P^* \quad , \quad p'(\mathrm{T_P}|\neg\mathrm{T_P^*}) = q_P^* \tag{8.7}$$

$$p'(\mathrm{T_F^*}|\mathrm{T_F}) = p_F^* \quad , \quad p'(\mathrm{T_F}|\neg\mathrm{T_F}) = q_F^* \tag{8.8}$$

Note that Equation (8.7) replaces the second equation in first line of Equation (8.1). We also have to represent the bridge law in probabilistic terms.

Naturally, we require:

$$p'(\mathrm{T}_\mathrm{P}^*|\mathrm{T}_\mathrm{F}^*) = 1 \quad , \quad p'(\mathrm{T}_\mathrm{P}^*|\neg\mathrm{T}_\mathrm{F}^*) = 0 \tag{8.9}$$

All other probability assignments hold as in the case of P_1. Requiring this condition makes sure that we can compare the two scenarios later, i.e., the situations before and after the reduction.

Three remarks about the three steps in the reduction are in order. First, T_F^* may be more or less good. How good it is depends on the context (i.e., the application in question and the auxiliary assumptions made) and on the judgment of the scientists involved. In line with our Bayesian approach, we assume that the judgment of the scientists can be expressed in probabilistic terms. Second, the move from T_F^* to T_P^* in virtue of the bridge laws may be controversial amongst scientists. Whilst bridge-laws are non-conventional factual claims, different scientists may assign different credences to them. Third, what counts as strongly analogous will also depend on the specific context and on the judgment of the scientists. For example, whether entropy fluctuations can be neglected or not cannot be decided independently of the specific problem at hand, see Callender (2001). All this fits our Bayesian account well.

Note that, in the Bayesian network in Figure 8.3, there is now a direct sequence of arrows from T_F to T_P: the path through T_F^* to T_P^*. And hence, we expect that E_F is now probabilistically relevant for T_P and that E_P is now probabilistically relevant for T_F. And this is indeed what we find: the independencies formulated in Equation (8.4) do not hold any more. We state our results in the following two theorems:

Theorem 8.1 E_F *confirms* T_P *iff* $(p_F - q_F)\,(p_F^* - q_F^*)\,(p_P^* - q_P^*) > 0$.

This theorem entails that E_F confirms T_P if the following three conditions hold: (i) E_F confirms T_F (i.e., $p_F > q_F$), (ii) T_F confirms T_F^* (i.e.,$p_F^* > q_F^*$), and (iii) T_P^* confirms T_P (i.e., $p_P^* > q_P^*$). These conditions are immediately plausible. Condition (i) was assumed from the beginning, and conditions (ii) and (iii) make sure that there is a positive flow of confirmation from T_F to $\mathrm{T}_\mathrm{F}^* \equiv \mathrm{T}_\mathrm{P}^*$ (*qua* bridge law) and from T_P^* to T_P.

Theorem 8.2 E_P *confirms* T_F *iff* $(p_P - q_P)\,(p_F^* - q_F^*)\,(p_P^* - q_P^*) > 0$.

This theorem is analogous to the previous theorem. It entails that E_P confirms T_F if the following three conditions hold: (i) E_P confirms T_P

(i.e.,$p_P > q_P$), (ii) $\mathrm{T_F}$ confirms $\mathrm{T_F^*}$ (i.e.,$p_F^* > q_F^*$), and (iii) $\mathrm{T_P^*}$ confirms $\mathrm{T_P}$ (i.e.,$p_P^* > q_P^*$).

Note that, in our representation, the bridge law states a *perfect correlation* between $\mathrm{T_F^*}$ and $\mathrm{T_P^*}$. A bridge law is posited by scientists working in a particular field, and it may happen that not everybody in that community is convinced of it. Thus, different scientists may assign different credences to a particular bridge law. In a case where a lower probability is assigned to a bridge law, the reduction may still be epistemically valuable – the flow of confirmation will just be less. How much confirmation will flow depends, of course, on the values of the relevant probabilities.

For future reference, let us calculate the prior probability of the conjunction of both theories. We obtain:

$$p'(\mathrm{T_F}, \mathrm{T_P}) = t_F \ (p_F^* \, p_P^* + \overline{p}_F^* \, q_P^*) \tag{8.10}$$

In a similar way, we may calculate the posterior probability of both theories given the total evidence, i.e., the expression $p'(\mathrm{T_F}, \mathrm{T_P}|\mathrm{E}, \mathrm{E_F}, \mathrm{E_P})$—see Section 8.5.

Finally, let us remark on the specific representation we have chosen in the Bayesian Network in Figure 8.3. Clearly, having a sequence of arrows from T_F to T_P ensures that confirmation can flow from one theory to the other. However, this sequence of arrows is not just driven by our wish to establish a flow of confirmation from the reducing to the reduced theory: it makes scientific sense. First, $\mathrm{T_F^*}$ is an approximation of $\mathrm{T_F}$. It follows from it and depends on it, hence the direction of the arrow. Second, we have drawn an arrow from T_F^* to T_P^* although the propositional variables in question are, *qua* the bridge law, intersubstitutable with each other. This is modeled by assigning appropriate conditional probabilities. The arrow could have also been drawn from T_P^* to T_F^*. In this case we had to require $P(\mathrm{T_F^*}|\mathrm{T_P^*}) = 1$ and $P(\mathrm{T_F^*}|\neg\mathrm{T_P^*}) = 0$. These conditions are, however, equivalent to Equations (8.9) for non-extreme priors. Third, it may look strange that we have drawn an arrow from T_P^* to T_P to model the relation of strong analogy as a symmetrical relation. We would like to reply to this objection that, then, "analogy" is perhaps not the right word as $\mathrm{T_P^*}$ is indeed stronger than $\mathrm{T_P}$, and so it makes sense to draw an arrow from T_P^* to T_P. We conclude that the chain of arrows from T_F to T_P is indeed plausible.

Why Accept a Purported Reduction?

Under what conditions should we accept a proposed reduction? More specifically, given everything we know about the domains of the two theories, when should we accept a proposed reduction and when should we reject it? In the Bayesian framework theories are accepted on the basis of their probabilities and confirmation track record. But which probabilities are relevant? The previous section focused on the probabilities of $\mathrm{T_F}$ and $\mathrm{T_P}$ individually. But perhaps one is interested in the "package" as a whole, that is, the *conjunction* of $\mathrm{T_F}$ and $\mathrm{T_P}$. If so, should we look at the *prior* probability of the conjunction of $\mathrm{T_F}$ and $\mathrm{T_P}$ after the reduction (that is, without accounting for the total evidence)? Or at the *posterior* probability of the conjunction of $\mathrm{T_F}$ and $\mathrm{T_P}$, i.e.,the probability of $\mathrm{T_F}$ and $\mathrm{T_P}$ given the total evidence (i.e., E, $\mathrm{E_F}$ and $\mathrm{E_P}$)? We examine these proposals in turn.

Let us first compare the prior probabilities of the conjunction of $\mathrm{T_F}$ and $\mathrm{T_P}$ before and after the reduction. Before the reduction, the two theories are independent, as expressed in Equation (8.6). For convenience, let us restate the condition:

$$p(\mathrm{T_F}, \mathrm{T_P}) = t_F \, t_P \tag{8.11}$$

We now calculate the prior probability of the conjunction of $\mathrm{T_F}$ and $\mathrm{T_P}$ after the reduction and obtain

$$p'(\mathrm{T_F}, \mathrm{T_P}) = t_F \, (p^*_F \, p^*_P + \overline{p}^*_F \, q^*_P) \, . \tag{8.12}$$

While the expression in Equation (8.11) is an explicit function of t_P, the expression in Equation (8.12) is not. This is because, after the reduction, T_P is no longer a root node, and so it is not assigned a prior probability. In order to meaningfully compare the situation before and after the reduction, we not only have to assume that $p'(\mathrm{E_P}|\mathrm{T_P}) = \mathrm{p}(\mathrm{E_P}|\mathrm{T_P})$ etc., but also that $p'(\mathrm{T_P}) = p(\mathrm{T_P})$. Let us therefore calculate:

$$\tilde{t}_P := p'(\mathrm{T_P}) = t^*_F \, p^*_P + \overline{t}^*_F \, q^*_P \tag{8.13}$$

with

$$t^*_F = p'(\mathrm{T^*_F}) = p'(\mathrm{T^*_P}) = p^*_F \, t_F + q^*_F \, \overline{t}_F \, . \tag{8.14}$$

Alternatively, we have:

$$\tilde{t}_P := (p^*_F p^*_P + \overline{p}^*_F q^*_P) \, t_F + (q^*_F p^*_P + \overline{q}^*_F q^*_P) \, \overline{t}_F \tag{8.15}$$

This equation follows if we insert Equation (8.14) into equation (8.13) or by direct calculation from the Bayesian network depicted in Figure 8.3. We now require $p'(\mathrm{T_P}) = p(\mathrm{T_P})$, i.e.,

$$t_P = \tilde{t}_P \tag{8.16}$$

and replace t_P in Equation (8.11) by the expression for $\tilde{t}_P$ given in Equation (8.15).
With this we calculate the difference,

$$\Delta_0 := p'(\mathrm{T_F}, \mathrm{T_P}) - p(\mathrm{T_F}, \mathrm{T_P}) \tag{8.17}$$

and obtain:

$$\Delta_0 = (p_F^* - q_F^*)\,(p_P^* - q_P^*)\,t_F\,\bar{t}_F \tag{8.18}$$

Hence,

Theorem 8.3 *$\Delta_0 = 0$ iff $(p_F^* = q_F^*)$ or $(p_P^* = q_P^*)$. And $\Delta_0 > 0$ if $(p_F^* > q_F^*)$ and if $(p_P^* > q_P^*)$.*

The first part of the theorem says that if either $\mathrm{T_F}$ and $\mathrm{T_F^*}$ are independent or if $\mathrm{T_P^*}$ and $\mathrm{T_P}$ are independent, then $\mathrm{T_F}$ and $\mathrm{T_P}$ remain independent after the reduction. The second part of the theorem says that the conjunction of $\mathrm{T_F}$ and $\mathrm{T_P}$ is more likely after the reduction if $\mathrm{T_F}$ confirms $\mathrm{T_F^*}$ and if $\mathrm{T_P^*}$ confirms $\mathrm{T_P}$.

Next, let us compare the posterior probabilities of the conjunction of $\mathrm{T_F}$ and $\mathrm{T_P}$ before and after the reduction. To do so, we calculate the difference,

$$\Delta_1 := p'(\mathrm{T_F}, \mathrm{T_P}|\mathrm{E}, \mathrm{E_F}, \mathrm{E_P}) - p(\mathrm{T_F}, \mathrm{T_P}|\mathrm{E}, \mathrm{E_F}, \mathrm{E_P}) \tag{8.19}$$

and obtain:

$$\Delta_1 = (p_F^* - q_F^*)\,(p_P^* - q_P^*)\,t_F\,\bar{t}_F \cdot \alpha\,\tilde{\Delta}_1\,, \tag{8.20}$$

The explicit expression for $\tilde{\Delta}_1$ is given in the appendix. Equation (8.20) then entails the following theorem:

Theorem 8.4 *$\Delta_1 = 0$ if $(p_F^* = q_F^*)$ or $(p_P^* = q_P^*)$. That is, the posterior probability of $\mathrm{T_F} \wedge \mathrm{T_P}$ equals the prior probability if one of the two equalities above are satisfied.*

This result has an intuitive interpretation: If either $\mathrm{T_F}$ and $\mathrm{T_F^*}$ or $\mathrm{T_P^*}$ and $\mathrm{T_P}$ are independent, then the flow of confirmation from $\mathrm{T_F}$ to $\mathrm{T_P}$ (and vice versa) is stopped and the epistemic situation before and after the reduction are the same.

Using the expression for $\tilde{\Delta}_1$, we obtain:

Theorem 8.5 $\Delta_1 > 0$ *if the following three conditions hold: (i)* $\beta, \gamma > \delta$*, (ii)* $0 < x_F, x_P < 1$*, and (iii)* $(p_F^* - q_F^*)(p_P^* - q_P^*) > 0$*. That is, the posterior probability of* $\mathrm{T_F} \wedge \mathrm{T_P}$ *exceeds the prior probability if the above inequalities are satisfied.*

Condition (i) seems natural in the light of inequalities (8.2) and (8.3). In fact, it is a rather weak condition which also holds, for example, for Set 2, below. Condition (ii) makes sure that $\mathrm{E_F}$ confirms $\mathrm{T_F}$ and $\mathrm{E_P}$ confirms $\mathrm{T_P}$; we have assumed this throughout. Condition (iii) is our usual condition on the dependency between $\mathrm{T_P}$ and $\mathrm{T_P^*}$, as well as between $\mathrm{T_F}$ and $\mathrm{T_F^*}$. Hence, none of these conditions is in any way problematic. Given this, we conclude that the posterior probability of the conjunction of $\mathrm{T_F}$ and $\mathrm{T_P}$ indeed increases after a reductive relationship is established between the two theories.

Discussion

We have discussed how the Generalized Nagel-Schaffner model of reduction impacts on the confirmation of theories by evidence. We formulated criteria that help us assess proposed reductions epistemically, and we have shown how a reduction facilitates the flow of confirmation from the reducing theory to the reduced theory and back.

A GNS reduction between two theories, such as thermodynamics and statistical mechanics, is epistemically advantageous in virtue of our main results: Theorem 8.1 and Theorem 8.2. Specifically, we have shown that a reduction ensures that evidence which, prior to reduction, only supported one of the theories, comes to support the other theory as well, due to the reduction. Moreover, a successful reduction increases both the prior and the posterior probability of the conjunction of both theories (Theorem 8.5).

Our Bayesian account also shows to what extent the various judgments depend on the probabilistic judgments of the scientists, connecting—or so we argue—our account to scientific practice. Disagreement about the epistemic value of a reduction can be traced back to disagreement about the assignment of the relevant prior probabilities and probabilities. This need not be a disagreement about exact numbers and may also take the form of qualitative (e.g., ordinal) plausibility judgments.

As usual, we finish the variation with a series of proposals for follow-up projects.

First, one might propose to accept a proposed reduction if the conjunction of T_F and T_P is better confirmed by the evidence after the reduction, compared to the situation before the reduction. Determining whether this is the case requires an analysis in terms of *degree of confirmation*. That is, one has to choose one of the various confirmation measures ($\rightarrow$ Variation 2). Dijadzi-Bahmani, Frigg and Hartmann conduct such an analysis for the difference measure $d(H, E)$ and come to the conclusion that the degree of confirmation is usually greater if a reduction has taken place than if not. Several other confirmation measures have to be checked and the stability of these results has to be explored.

The previous observation suggests that strong coherence between the fundamental and the phenomenological theory may be confirmation-conducive (Dietrich and Moretti, 2005; Moretti, 2007). So second, it would be interesting to compare degrees of coherence before and after an intertheoretic reduction has taken place (Bovens and Hartmann, 2003). Here, one might want to focus on the two theories in question, or on the conjunction of the theories and all available evidence. It might be reasonable to focus on the latter, as the evidence is also uncertain and one might, in the end, be interested in the coherence of the entire package, comprising all available theories and all available evidence. Should coherence considerations play a role when it comes to decide whether a theory should be accepted?

Third, one may want to examine the situation where evidence for, say, the fundamental theory *disconfirms* the phenomenological theory. How shall one assess the value of a reduction in these situations?

Fourth and finally, other types of intertheoretic relation should be studied from a Bayesian point of view. Here, we are thinking of "stories" (Hartmann, 1999) and singular limits (Batterman, 2002). But there will surely be other examples. This project requires the collaboration between philosophers of science, who conduct case studies, and formal philosophers, who provide the corresponding Bayesian analysis. It may also be asked which picture about the structure of science as a whole emerges from all this. It seems plausible to find something like a network structure, with more or less connected theories and models, and it might be interesting to discuss the implications of this for the debate about the (dis-)unity of science.

Proofs of the Theorems

Let us start with the situation before the reduction and the Bayesian network represented in Figure 2. The joint distribution $p(T_F, T_P, E, E_F, E_P)$ is given by the expression

$$p(T_F)\, p(T_P)\, p(E|T_F, T_P)\, p(E_F|T_F)\, p(E_P|E_P)\,.$$

Using the methodology described in Bovens and Hartmann (2003, Ch. 3), we obtain:

$$\begin{aligned} p(\mathrm{T_F}, \mathrm{E}) &= \sum_{T_P, E_F, E_P} p(T_F, T_P, E, E_F, E_P) \\ &= t_F\,(t_P\,\alpha + \bar{t}_P\,\beta) \end{aligned} \tag{8.21}$$

Similarly, we calculate

$$p(\mathrm{T_P}, \mathrm{E}) = t_P\,(t_F\,\alpha + \bar{t}_F\,\gamma) \tag{8.22}$$

$$p(\mathrm{E}) = t_F\,(t_P\,\alpha + \bar{t}_P\,\beta) + \bar{t}_F\,(t_P\,\gamma + \bar{t}_P\,\delta) \tag{8.23}$$

$$= t_P\,(t_F\,\alpha + \bar{t}_F\,\gamma) + \bar{t}_P\,(t_F\,\beta + \bar{t}_F\,\delta) \tag{8.24}$$

To prove Equation (8.2) we note, using the definition of conditional probability, that $p(\mathrm{T_P}|\mathrm{E}) > p(\mathrm{T_P})$ iff $p(\mathrm{T_P}, \mathrm{E}) - p(\mathrm{T_P})\, p(\mathrm{E}) > 0$ and obtain using Equations (8.22) and (8.24)

$$p(\mathrm{T_P}, \mathrm{E}) - p(\mathrm{T_P})\, p(\mathrm{E}) = t_P\, \bar{t}_P \left[(\alpha - \beta)\, t_F + (\gamma - \delta)\, \bar{t}_F\right]\,, \tag{8.25}$$

from which Equation (8.2) immediately follows. The proof of Equation (8.3) proceeds accordingly using Equations (8.22) and (8.23).

Next, we calculate the prior probability of the two theories.

$$\begin{aligned} p(\mathrm{T_F}, \mathrm{T_P}) &= \sum_{E, E_F, E_P} p(T_F, T_P, E, E_F, E_P) \\ &= p(\mathrm{T_F})\, p(\mathrm{T_P}) = t_F\, t_P \end{aligned}$$

Similarly, we obtain for the posterior probability $P_1^* := p(\mathrm{T_F}, \mathrm{T_P}|\mathrm{E}, \mathrm{E_F}, \mathrm{E_P})$:

$$\begin{aligned} P_1^* &= \frac{p(\mathrm{T_F}, \mathrm{T_P}, \mathrm{E}, \mathrm{E_F}, \mathrm{E_P})}{p(\mathrm{E}, \mathrm{E_F}, \mathrm{E_P})} \\ &= \frac{t_F\, t_P\, p_F\, p_P\, \alpha}{t_F\, t_P\, p_F\, p_P\, \alpha + t_F\, \bar{t}_P\, p_F\, q_P\, \beta + \bar{t}_F\, t_P\, q_F\, p_P\, \gamma + \bar{t}_F\, \bar{t}_P\, q_F\, q_P\, \delta} \\ &= \frac{t_F\, t_P\, \alpha}{t_F\,(t_P\,\alpha + \bar{t}_P\, x_P\, \beta) + \bar{t}_F\, x_F\,(t_P\,\gamma + \bar{t}_P\, x_P\, \delta)}\,, \end{aligned} \tag{8.26}$$

with the probability ratios $x_F := q_F / p_F$ and $x_P := q_P / p_P$.

Let us now turn to the situation after the reduction and the Bayesian network represented in Figure 3. The joint distribution $p'(T_F, T_P, T_F^*, T_P^*, E, E_F, E_P)$ is given by

$$p'(T_F)\, p'(E|T_F, T_P)\, p'(E_F|T_F)\, p'(E_P|E_P)\, p'(T_P|T_P^*)\, p'(T_P^*|T_F^*)\, p'(T_F^*|T_F)\,.$$

To simplify our notation, we introduce the following abbreviations:

$$\begin{aligned} \varphi_\alpha &:= p_F^*\, p_P^* + \overline{p}_F^*\, q_P^* \quad , \quad \varphi_\beta := p_F^*\, \overline{p}_P^* + \overline{p}_F^*\, \overline{q}_P^* \\ \varphi_\gamma &:= q_F^*\, p_P^* + \overline{q}_F^*\, q_P^* \quad , \quad \varphi_\delta := q_F^*\, \overline{p}_P^* + \overline{q}_F^*\, \overline{q}_P^* \end{aligned}$$

For later use, we note that $0 < \varphi_\alpha, \varphi_\beta, \varphi_\gamma, \varphi_\delta < 1$ and

$$\varphi_\alpha - \varphi_\gamma = \varphi_\delta - \varphi_\beta = (p_F^* - q_F^*)\,(p_P^* - q_P^*) \tag{8.27}$$

$$\varphi_\alpha + \varphi_\beta = \varphi_\gamma + \varphi_\delta = 1\,. \tag{8.28}$$

We then obtain for the prior probability of the conjunction of both theories after the reduction

$$p'(\mathrm{T_F}, \mathrm{T_P}) = t_F\, \varphi_\alpha\,. \tag{8.29}$$

For the posterior $P_2^* := p'(\mathrm{T_F}, \mathrm{T_P}|\mathrm{E}, \mathrm{E_F}, \mathrm{E_P})$, we obtain:

$$P_2^* = \frac{t_F\, \alpha\, \varphi_\alpha}{t_F\, (\alpha\, \varphi_\alpha + x_P\, \beta\, \varphi_\beta) + \bar{t}_F\, x_F\, (\gamma\, \varphi_\gamma + x_P\, \delta\, \varphi_\delta)} \tag{8.30}$$

Similarly, we calculate

$$p'(\mathrm{T_P}) = t_F\, \varphi_\alpha + \bar{t}_F\, \varphi_\gamma \tag{8.31}$$

$$p'(\mathrm{T_P}|\mathrm{E_F}) = \frac{t_F\, \varphi_\alpha + \bar{t}_F\, x_F\, \varphi_\gamma}{t_F + \bar{t}_F\, x_F} \tag{8.32}$$

$$p'(\mathrm{T_F}|\mathrm{E_P}) = \frac{t_F\, (\varphi_\alpha + x_P\, \varphi_\beta)}{t_F\, (\varphi_\alpha + x_P\, \varphi_\beta) + \bar{t}_F\, (\varphi_\gamma + x_P\, \varphi_\delta)}\,. \tag{8.33}$$

We now calculate

$$\begin{aligned} p'(\mathrm{T_P}|\mathrm{E_F}) - p'(\mathrm{T_P}) &= \frac{t_F\, \bar{t}_F\, (\varphi_\alpha - \varphi_\gamma)\, (1 - x_F)}{t_F + \bar{t}_F\, x_F} \\ &= \frac{t_F\, \bar{t}_F\, (p_F - q_F)\, (p_F^* - q_F^*)\, (p_P^* - q_P^*)}{p_F\, (t_F + \bar{t}_F\, x_F)}\,. \end{aligned}$$

This proves Theorem 8.1. Similarly, we calculate

$$p'(\mathrm{T_F}|\mathrm{E_P}) - p'(\mathrm{T_F}) = \frac{t_F\, \bar{t}_F\, (\varphi_\alpha - \varphi_\gamma)\, (1 - x_P)}{t_F\, (\varphi_\alpha + x_P\, \varphi_\beta) + \bar{t}_F\, (\varphi_\gamma + x_P\, \varphi_\delta)}$$

$$= \frac{t_F\,\bar{t}_F\,(p_P - q_P)\,(p_F^* - q_F^*)\,(p_P^* - q_P^*)}{p_P\,\left[t_F\,\left(\varphi_\alpha + x_P\,\varphi_\beta\right) + \bar{t}_F\,\left(\varphi_\gamma + x_P\,\varphi_\delta\right)\right]}\,,$$

which proves Theorem 8.2.

To prove Equation (8.18), we note that, using Equation (8.29)

$$\Delta_0 = (\varphi_\alpha - t_P)\,t_F\,.$$

We now use Equations (8.16) and (8.31) and obtain

$$\begin{aligned}\Delta_0 &= (\varphi_\alpha - t_F\,\varphi_\alpha - \bar{t}_F\,\varphi_\gamma)\,t_F \\ &= (\varphi_\alpha - \varphi_\gamma)\,t_F\,\bar{t}_F\,.\end{aligned}$$

Equation (8.18) then follows using Equation (8.27).

Let us finally calculate Δ_1 using Equations (8.26) and (8.30). We obtain

$$\Delta_1 = (\varphi_\alpha - \varphi_\gamma)\,t_F\,\bar{t}_F \cdot \alpha\,\tilde{\Delta}_1\,, \tag{8.34}$$

with

$$\tilde{\Delta}_1 = N_1^{-1}\,N_2^{-1} \cdot \tilde{\Delta}_1' \tag{8.35}$$

and

$$\begin{aligned}N_1 &= t_F\,(t_P\,\alpha + \bar{t}_P\,x_P\,\beta) + \bar{t}_F\,x_F\,(t_P\,\gamma + \bar{t}_P\,x_P\,\delta) \\ N_2 &= t_F\,\left(\alpha\,\varphi_\alpha + x_P\,\beta\,\varphi_\beta\right) + \bar{t}_F\,x_F\,(\gamma\,\varphi_\gamma + x_P\,\delta\,\varphi_\delta)\,.\end{aligned}$$

Note that $N_1, N_2 > 0$. We are therefore most interested in $\tilde{\Delta}_1'$, which is given by

$$\begin{aligned}\tilde{\Delta}_1' &= t_F\,x_F\,(\varphi_\alpha - \varphi_\gamma)\,(\gamma - \delta\,x_P) + t_F\,x_P\,(\beta - \delta\,x_F) \\ &+ \gamma\,\varphi_\gamma\,x_F + \delta\,\overline{\varphi}_\gamma\,x_F\,x_P\,.\end{aligned}$$

From conditions (i) and (ii) of Theorem 8.5, we conclude that $\gamma > \delta\,x_P$ and $\beta > \delta\,x_F$. Hence $\tilde{\Delta}_1' > 0$, which proves the theorem.

Variation 9: Hypothesis Testing and Corroboration

Scientific reasoning often proceeds by testing hypotheses and appraising how well they have stood up to the test. For critical rationalists such as Karl R. Popper (2002), the critical attitude that we express by repeatedly testing our best scientific theories even constitutes the basis of rational inquiry about the world. Arguably, such tests have already been conducted in antiquity—think of Erastothenes' test of the hypothesis that the Earth is round, conducted by comparing the height of the sun in two different places at the same time. However, only in the middle of the 20th century, the design and interpretation of hypothesis tests has been formalized and standardized. The emergence of the discipline of statistics, the science of analyzing and interpreting data, played a crucial role in this process. It provided science with probabilistic tests, above all **null hypothesis significance tests (NHST)**, which have acquired a predominant role in scientific reasoning.

NHST test a precise hypothesis H_0—the "null" or default hypothesis—against an unspecific alternative H_1. In the most common form of NHST, the null hypothesis posits a precise value for a real-valued parameter θ ($H_0 : \theta = \theta_0$), while the alternative ($H_1 : \theta \neq \theta_0$) is a disjunction of infinitely many precise hypotheses (e.g., Neyman and Pearson, 1933, 1967; Fisher, 1956). The null hypothesis standardly denotes an absent or negligible effect (e.g., a new medical drug is not better than a placebo treatment) whereas the alternative stands for a sizeable effect. NHST are applied across all domains of science, but they are especially prominent in psychology and medicine.

Despite their popularity in scientific inference, the philosophical foundations of NHST are shaky at best. NHST are used for quantifying evidence that the data accumulate against the null hypothesis. When this

level of evidence is high enough, i.e., greater than a prespecified significance threshold, the null hypothesis is rejected. See Figure 9.1 for an illustration. The more the observed value lies in the tail of the distribution, the more it counts as evidence against the null hypothesis (Fisher, 1956; Mayo, 1996, e.g.,). Mathematically, the significance level is captured by the notorious ***p*-value**: for a random variable X with realization x and a function $z(X)$ that measures the distance to the null hypothesis (e.g., the difference between the null mean and the actual sample mean), the p-value describes the probability of obtaining a result that speaks as least as much against the null hypothesis as the actual result.

$$p := p_{\mathrm{H}_0}(|z(X)| \geq |z(x)|), \tag{9.1}$$

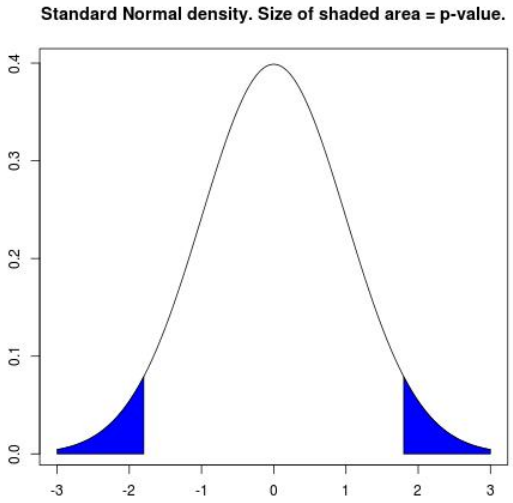

Figure 9.1: The shaded area indicates the set of observations where the null hypothesis H_0 is "rejected". Here, H_0 denotes the hypothesis that the observations follow a standard Normal distribution with mean value $\theta = 0$ as opposed to $\theta \neq 0$.

However, there is barely any methodological guidance on **how we should interpret a non-significant result**, that is, a result where we fail to reject the null hypothesis. Statistics textbooks (e.g., Chase and Brown, 2000; Wasserman, 2004) restrict themselves to a purely negative interpretation: failure to reject the null means failure to demonstrate a statistically significant phenomenon.

To illustrate this point, consider a Binomial model where we are testing the hypothesis that the coin is fair. To what extent does a result of 52 heads in 100 independent tosses corroborate the null hypothesis? Neither $x = 52$ nor $x = 58$ qualifies as significant evidence against the null hypothesis at the $p = 0.05$ level, but there is certainly a difference in the performance of the null hypothesis on that particular dataset. The classical statistical methodology, which refuses to interpret p-values greater than 0.05, fails to quantify this difference. For example, p-values of .15 and .35 have, for all practical purposes, the same meaning: they are above the range where results are statistically significant, and therefore no evidence against the null.

All in all, the standard NHST method does not address the question whether the results **corroborate the null hypothesis**. Should we prefer the null hypothesis to the alternative hypotheses and preliminarily accept it? Whenever the null hypothesis is of substantial scientific interest, e.g., independence of two variables in a causal model, the safety of a medical drug or the adequacy of a phylogenetic tree, such judgments are urgently required. This fact is also acknowledged by numerous scientists. For two recent examples from psychology, see Gallistel (2009) and Morey et al. (2014).

Explicating (degree of) corroboration is thus central for a sound interpretation of NHST. Karl R. Popper, one of the few philosophers engaging in this debate, proposed the following characterization:

> By the degree of corroboration of a theory I mean a concise report evaluating the state (at a certain time t) of the critical discussion of a theory, with respect to the way it solves its problems; its degree of testability; the severity of tests it has undergone; and the way it has stood up to these tests. Corroboration (or degree of corroboration) is thus an evaluating report of past performance. Like preference, it is essentially comparative. (Popper, 1979, 18)

In Popper's view, corroboration judgments positively appraise the performance of the null hypothesis in a severe test, rather than just stating the failure to find significant evidence against it. Notably, high degrees of corroboration need not guide us to the truth (Popper, 1979, 21). Instead, the function of corroboration is comparative and pragmatic: it guides our

practical preferences over competing hypotheses, for example the choice of the hypothesis on which we base the next experiment (Popper, 2002, 416). This is exactly what most scientists are after when testing a complex set of hypotheses.

The corroboration-based approach to scientific reasoning should not be confused with confirmation-based reasoning. While (Bayesian) confirmation is based on increase in degree of belief, corroboration does not imply any *confidence* in the tested hypothesis: it is just the statement that a hypothesis has survived severe tests. Popper (2002, ch. 8 and 10, appendix vii) even argued for the impossibility of inductive (Bayesian) probability while defending the epistemic role of corrorboration. According to Popper's corroboration-centered perspective, scientific progress occurs through successive elimination of hypotheses, and degrees of corroboration guide practical preferences over the competing hypotheses. This is something very different from a probabilistic inductive logic in the style of Carnap (1950).

This variation, which partly builds on results from Sprenger (2016d), explores the prospects for a corroboration-based epistemology of NHST. We begin with a conceptual demarcation of degree of corroboration versus degree of confirmation (Section 9.1). Then, we discuss Popper's own explication of corroboration (Section 9.2) and address the more general question of whether testability and past performance may be synthesized into a single measure of corroboration (Section 9.3). The answer is negative: no such measure can simultaneously satisfy a set of desirable constraints. This seems to create insurmountable problems for the project of explicating corroboration, but they can be solved by moving to a different statistical framework. We construct a measure of corroboration that fruitfully applies Popperian thinking to hypothesis tests and that can be understood as a generalization of Bayesian inference (Section 9.4). Finally, we compare this measure to p-values in NHST and standard Bayesian inference (Section 9.5) and we provide the proofs of our results (Section 9.6). While the practical merits of the new corroboration measure are still to be evaluated, it demonstrates two important theoretical insights: First, we can provide a valid interpretation of non-significant results in NHST. Second, Popperian and Bayesian approaches to hypothesis testing may, in the end, be less fierce opponents in hypothesis testing than the popular picture has it.

Confirmation versus Corroboration

The point of measuring corroboration is to quantify the extent to which a hypothesis has stood up to an attempt to refute it. Thus, degree of corroboration gives an evaluating report of past performance. For the case of a hypothesis that makes deterministic predictions, corroborating evidence is intuitively defined as evidence that conforms to the predictions of the tested hypothesis. The more specific the evidence, the more it corroborates the hypothesis.

This rationale essentially corresponds to the hypothetico-deductive model of theory confirmation (e.g., Gemes, 1998): observed logical consequences of a theory confirm it. While this model may be adequate as a qualitative theory of corroboration, it is not applicable to NHST. Here, a different, quantitative model has to be developed that applies to statistical inference (see also Popper, 2002, 265–266).

However, do we really need the concept of corroboration to explicate this aspect of NHST? Can't we just describe the results of NHST in terms of degree of confirmation? According to Bayesian Confirmation Theory, evidence E confirms hypothesis H if and only if $p(\mathrm{H}|\mathrm{E}) > p(\mathrm{H})$, where p represents an agent's subjective degrees of belief. That is, E confirm H if and only if E increases the agent's subjective degree of belief in H (e.g., Fitelson, 2001b, see also Variation 2 of this book). Before introducing a new and complex concept—corroboration—we first need to argue why it is not coextensive with confirmation as increase in firmness.

In other words, we have to address the **Monism Thesis**: the epistemic function of the concept of corroboration can be taken over by the Bayesian concept of confirmation as increase in firmness. The monist replaces a judgment of corroboration by a judgment of confirmation. This line of argument gains support from authors such as Howson and Urbach (2006) and Wagenmakers et al. (2011), who argue that NHST should be abandoned and be replaced by Bayesian hypothesis testing.

We shall now present three objections to the Monism thesis. This does not rule out that explications of corroboration and confirmation agree numerically: rather, the point is to show that the two *concepts* are not redundant and need different explication strategies.

> **Objection 1:** Corroboration does not aim at inferring *probable* hypotheses, or raising our degree of belief in the tested hypoth-

esis.

This objection contends that scientific hypotheses and models are idealizations of the external world, which are judged by their ability to capture relevant causal relations and to predict future events (see the survey of Frigg and Hartmann, 2012). The epistemic function of corroboration consists in determining whether the data are consistent with the tested hypothesis, or whether the results agree "well enough" with the null hypothesis H_0 that we may use it as a proxy for a more general statistical model.

Consequently, corroborated hypotheses should not be regarded as true or empirically adequate, but as useful and tractable idealization of a general statistical model (Bernardo, 2012; Gelman and Shalizi, 2012, 2013). Corroboration is a guide to practical preference over competing hypothesis, but it does not ground confidence in the truth of the tested hypothesis (Popper, 2002, 281–282).

Degree of confirmation, on the other hand, is defined by the change of confidence in a hypothesis. Characteristically, all confirmation measures $c(\mathrm{H}, \mathrm{E})$ possess the Final Incrementality Property familiar from Variation 2:

$$c(\mathrm{H}, \mathrm{E}) > c(\mathrm{H}, \mathrm{E}') \qquad \text{if and only if} \qquad p(\mathrm{H}|\mathrm{E}) > p(\mathrm{H}|\mathrm{E}'). \tag{9.2}$$

This condition demands that E confirms H more than E′ if and only if E raises the probability of H to a higher level than E′ does (Festa, 2012; Crupi, 2013). However, corroboration is about past performance, not about epistemic or psychological attitude. In a nutshell, rather than a (subjective) measure of belief change, corroboration ought to be an (objective) measure of past performance. Indeed, even if we did not have subjective degrees of belief in the tested hypothesis or were unable to elicit them, we should still be able to assess the past performance of the null hypothesis by a judgment of corroboration.

> **Objection 2:** On a Bayesian account, hypotheses with prior probability $p(\mathrm{H}) = 0$ cannot be confirmed. Yet, they are perfectly acceptable candidates for being corroborated.

This point was first raised by Karl R. Popper (2002, appendix vii). As a consequence of Bayes' Theorem, any hypothesis with prior probability

$p(\mathrm{H}) = 0$ also has posterior probability $p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H})\, p(\mathrm{E}|\mathrm{H})/p(\mathrm{E}) = 0$. No such hypothesis can be confirmed in the sense of increase in firmness. But certainly, they can be *corroborated*: after all, scientists often deal with an uncountable set of candidate hypotheses where all singleton hypotheses receive zero weight (e.g., different values of a physical parameter). Testing whether such hypotheses are good idealizations of reality certainly makes sense.

This argument is in line with the practice of Bayesian statistics. Bayesian hypothesis tests often assign zero weight to the null hypothesis $\mathrm{H}_0 : \theta = \theta_0$, e.g., by assigning a continuous prior over the entire parameter space. Whatever the measure of evidence that the Bayesian uses for appraising the null in such tests (e.g., a density-based measure such as the Bayes factor), it cannot be a classical Bayesian confirmation measure that compares $p(\mathrm{H}|\mathrm{E})$ to $p(\mathrm{H})$. The Bayesian apparatus may be a convenient mathematical tool for performing such hypothesis tests, but it stands in need of a philosophical rationale regarding the outcomes of the analysis.

> **Objection 3:** Corroboration is a way more asymmetric notion than confirmation.

The logic of NHST is asymmetric: unlike the null hypothesis, the alternative is usually not a precise hypothesis, like in our introductory example of testing $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_1 : \theta \neq \theta_0$. As explained above, NHST often aims at finding out whether the null hypothesis is a good proxy for the more general model represented by the alternative. Finding the null hypothesis highly corroborated is a precise conclusion in favor of the null whereas a "rejection" of the null leaves open which of the alternatives is corroborated. Confirmation judgments, however, are symmetric: disconfirmation of H is also confirmation of $\neg\mathrm{H}$, and sometimes, it is also demanded that $c(\neg\mathrm{H}, \mathrm{E}) = -c(\mathrm{H}, \mathrm{E})$ (Crupi et al., 2007). That is, while confirmation measures pitch a hypothesis H against its negation $\neg\mathrm{H}$, measures of corroboration pitch H_0 against a set of distinct alternatives. Section 9.4 will elaborate this idea.

These objections undermine the Monism Thesis sufficiently to motivate that the concept of corroboration stands in need of an independent explication and cannot be reduced to degree of confirmation. We begin by discussing Popper's classical proposal for a measure of corroboration.

Popper's Measure of Degree of Corroboration

Popper's first writings on degree of corroboration, in Chapter 10 of "Logic of Scientific Discovery" (1934/2002), do not engage in a quantitative explication. Apparently, this task is deferred to a scientist's common sense (see, e.g., Popper, 2002, 265–267). However, this move makes the entire concept of corroboration vulnerable to the charge of subjectivism: without a quantitative criterion, it is not clear which corroboration judgments are sound and which aren't (Good, 1968b, 136). Especially if we aim at gaining objective knowledge from hypothesis tests, we need a precise explication of degree of corroboration.

Popper faces this challenge in a couple of *BJPS* articles (Popper, 1954, 1957, 1958) that form, together with a short introduction, appendix ix of "Logic of Scientific Discovery". In these articles, Popper develops and defends a measure of degree of corroboration. Popper argues that this measure cannot be a probability in the sense of Carnap (1950), that is, the plausibility of the tested theory (or hypothesis) conditional on the observed evidence:

> [...] the probability of a statement [...] simply does not express an appraisal of the severity of the tests a theory has passed, of the manner in which it has passed these tests. (Popper, 2002, 411)

In particular, logical content and informativity contribute to the testability of a theory and to its degree of corroboration:

> The main reason for this is that the *content* of a theory—which is the same as its *improbability*—determines its *testability* and *corroborability*. (ibid., original emphasis)

Recall that testability, identified with the empirical content or informativity of a hypothesis, is an essential cognitive value for Popper: being testable is a hallmark of science as opposed to pseudo-scientific theories that can be reconciled with all types of empirical evidence. Popper's classical examples are psychoanalysis and Marxist economics. While pseudo-scientific theories are a lens to watch the world rather than statements about the world (e.g., Marxists interpret all economic developments as following the logic of class struggle), genuinely scientific theories make testable predictions and may be refuted empirically.

Also in Popper's characterization of corroboration, testability is assigned a crucial role. Corroboration should be sensitive to the informativity and logical content of a theory, which is again related to the improbability of a theory. If one considers that degree of corroboration should guide our judgments of acceptance in NHST, this makes a lot of sense: good theories should agree with observed evidence and be informative (see the discussions in Hempel, 1960; Levi, 1963; Huber, 2005). Popper confirms that scientific theory assessment pursues both goals at once:

> Science does not aim, primarily, at high probabilities. It aims at a *high informative content*, well backed by experience. But a hypothesis may be very probable simply because it tells us nothing, or little. (Popper, 2002, 416, original emphasis)

Such a characterization of corroboration is attractive because it amalgamates two crucial cognitive values in theory assessment: high informative content and empirical confirmation. Also in NHST, both values play a role since a precise hypothesis (the null) is tested against a continuum of alternatives. However, this variation shows that such a tradeoff is unattainable if further reasonable assumptions are made.

Let us now look at how Popper characterizes degree of corroboration. Transcribed to modern notation, Popper assumes that evidence E and hypothesis H are among the closed sentences $\mathfrak{L}$ of a first-order language L. A corroboration measure is described by a function $C : \mathfrak{L}^2 \times \mathfrak{P} \to \mathbb{R}$, where $\mathfrak{P}$ is the set of probability measures on the σ-algebra generated by $\mathfrak{L}$. This function assigns a real-valued degree of corroboration $C(\mathrm{H}, \mathrm{E})$ to any pair of sentences in $\mathfrak{L}$, together with a probability measure p. This measure may be interpreted as a function of the logical structure of L, but also as objective chance or degree of belief—our discussion is independent of this point. For the sake of simplicity, we will omit reference to background assumptions and assume that they are implicit in the probability function p.

Note that such a probabilistic measure of corroboration does not capture all aspects of corroboration. Popper (2002, 265–266, 402, 437) and also his modern followers (e.g., Rowbottom, 2008, 2011) emphasize that corroborating evidence has to report the results of sincere and severe effort to overturn the tested hypothesis. Obviously, such requirements cannot be formalized completely (see also Popper, 1983, 154). We cannot infer

reversely from a high (probabilistic) degree of corroboration to a sound experimental design. The point of a probabilistic measure is rather to describe the degree of corroboration of a hypothesis if all important methodological requirements are met.

Popper then specifies a set of adequacy criteria I–IX for degree of corroboration as a function of empirical performance.

I $C(\mathrm{H},\mathrm{E}) >/=/< 0$ if and only if $p(\mathrm{E}|\mathrm{H}) >/=/< p(\mathrm{E})$.

This is a classical statistical relevance condition: E corroborates H just in case supposing H makes E more expected. This condition is also in line with Popper's remark that corroboration is, like preference, essentially contrastive (Popper, 1979, 18).

II $-1 = C(\neg\mathrm{H},\mathrm{H}) \leq C(\mathrm{H},\mathrm{E}) \leq C(\mathrm{H},\mathrm{H}) \leq 1$.

III $C(\mathrm{H},\mathrm{H}) = 1 - p(\mathrm{H})$.

IV If $\mathrm{E} \models \mathrm{H}$ then $C(\mathrm{H},\mathrm{E}) = 1 - p(\mathrm{H})$.

V If $\mathrm{E} \models \neg\mathrm{H}$ then $C(\mathrm{H},\mathrm{E}) = -1$.

These conditions determine under which conditions the measure of corroboration takes its extremal values. Minimal degree of corroboration is obtained if the evidence refutes the hypothesis (V). Conversely, the most corroborating piece of evidence E is a verification of H (II). In that case, degree of corroboration is equal to $1 - p(\mathrm{H})$ (III, IV), which expresses the informativity, testability and logical content of H. This is especially plausible in Carnap's logical interpretation of probability, which Popper adopts for $p(\mathrm{H})$. But it also makes sense for a subjective Bayesian interpretation. See Popper (2002, 268–269), Popper (1963, 385–387), Rowbottom (2012, 741–744).

VI $C(\mathrm{H},\mathrm{E}) \geq 0$ increases with the power of H to explain E.

VII If $p(\mathrm{H}) = p(\mathrm{H}')$, then $C(\mathrm{H},\mathrm{E}) > C(\mathrm{H}',\mathrm{E}')$ if and only if $p(\mathrm{H}|\mathrm{E}) > p(\mathrm{H}'|\mathrm{E}')$.

These conditions reiterate the statistical relevance rationale from condition I, and make it more precise. Regarding condition VI, Popper (2002, 416) defines explanatory power according to the formula $\mathcal{E}(e,h) =$

$(p(\mathrm{E}|\mathrm{H}) - p(\mathrm{E}))/(p(\mathrm{E}|\mathrm{H}) + p(\mathrm{E}))$, another measure of the statistical relevance between E and H. But the details need not bother us here. Condition VII states that corroboration essentially co-varies with posterior probability whenever two hypotheses are equiprobable at first. In that case, posterior probability is a good indicator of past performance. In comparison to Popper's original formulation, we have dropped the requirement $p(\mathrm{H}) > 0$ because by Bayes' Theorem, the case $p(\mathrm{H}) = p(\mathrm{H}') = 0$ would imply $p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H}'|\mathrm{E}') = 0$ and trivialize the condition.

VIII If $\mathrm{H} \models \mathrm{E}$, then

a) $C(\mathrm{H}, \mathrm{E}) \geq 0$;

b) $C(\mathrm{H}, \mathrm{E})$ is an increasing function of $1 - p(\mathrm{E})$;

c) $C(\mathrm{H}, \mathrm{E})$ is an increasing function of $p(\mathrm{H})$.

IX If $\neg\mathrm{H}$ is consistent and $\neg H \models \mathrm{E}$, then

a) $C(\mathrm{H}, \mathrm{E}) \leq 0$;

b) $C(\mathrm{H}, \mathrm{E})$ is an increasing function of $p(\mathrm{E})$;

c) $C(\mathrm{H}, \mathrm{E})$ is an increasing function of $p(\mathrm{H})$.

Condition VIII demands that corroboration gained from a successful deductive prediction co-vary with the informativity of the evidence and the prior probability of the hypothesis. Condition IX mirrors this requirement for the case $\neg H \models \mathrm{E}$. These conditions can be motivated from the idea that if $\mathrm{H} \models \mathrm{E}$, then corroboration should not automatically transfer to hypotheses $\mathrm{H} \wedge \mathrm{H}'$ that contain an "irrelevant conjunct" H' which has not yet been tested. See the next section for more detailed discussion of this point.

Popper (1954, 359) then proposes the corroboration measure $C_P(\mathrm{H}, \mathrm{E})$ which satisfies all of his constraints:

$$C_P(\mathrm{H}, \mathrm{E}) = \frac{p(\mathrm{E}|\mathrm{H}) - p(\mathrm{E})}{p(\mathrm{E}|\mathrm{H}) + p(\mathrm{E}) - p(\mathrm{E}|\mathrm{H})\, p(\mathrm{H})}. \tag{9.3}$$

But we can easily see that an essential motivation behind a measure of degree of corroboration is not satisfied. $C_P(\mathrm{H}, \mathrm{E})$ is an increasing function of $p(\mathrm{H})$ for all values of $p(\mathrm{E}|\mathrm{H})$ and $p(\mathrm{E})$. Hence, the informativity and testability of the hypothesis, as measured by $1 - p(\mathrm{H})$, never contributes to its degree of corroboration. This violates Popper's informal characterization of the concept and does not square well with the practice of

NHST. Diez (2011) provides even more reasons why Popper's explication is at odds with the tenets of critical rationalism. We shall now phrase this problem more generally and show that it does not only arise for Popper's measure $C_P(\mathrm{H}, \mathrm{E})$, but for all corroboration measures that are motivated from the same intuitions; that is, measures that aim at capturing statistical relevance and testability at the same time.

The Impossibility Results

Popper's nine adequacy conditions are quite specific requirements and too strong for the purpose of a general analysis of degree of corroboration. We will therefore weaken them and retain only those adequacy conditions that are indispensable for a conceptual analysis of corroboration. We then proceed to showing two impossibility results for corroboration measures that (i) build on statistical relevance between H and E and the predictive success of H for E; and (ii) preserve the intuition that corroboration should be responsive to the informativity and testability of the tested hypothesis.

First, we impose a condition which is mainly representational in nature and is frequently used in Bayesian Confirmation Theory and formal epistemology more generally (see Variation 2, 6 and 7 of this book for details). Popper's own measure $C_P(h, e)$ also conforms to it.

Formality There exists a function $f : [0,1]^3 \times \{(x,y,z)|1 + xz - z \geq y \geq xz\} \to \mathbb{R}$ such that for all $\mathrm{E}, \mathrm{H} \in \mathfrak{L}$ and $p \in \mathfrak{P}$,

$$C(\mathrm{H}, \mathrm{E}) = f(p(\mathrm{E}|\mathrm{H}), p(\mathrm{E}), p(\mathrm{H})).$$

This condition relates degree of corroboration to the joint probability distribution of E and H. The three arguments of f determine that distribution in all non-degenerate cases, and they are the same quantities that figure in Popper's measure of corroboration $C_P(\mathrm{H}, \mathrm{E})$. This makes comparisons easier. Formality means that two scientists who agree about all relevant probabilities will make the same corroboration judgments.[4]

In a Popperian spirit, we now demand that corroboration track predictive success (e.g., Popper, 1983, 241–243):

[4]Note that the corroboration measure is not defined on the entire unit cube $[0,1]^3$ since not all assignments of $p(\mathrm{E}|\mathrm{H})$, $p(\mathrm{E})$ and $p(\mathrm{H})$ are compatible with each other. This is evident from the equality

$$p(\mathrm{E}) = p(\mathrm{E}|\mathrm{H})p(\mathrm{H}) + p(\mathrm{E}|\neg\mathrm{H})(1 - p(\mathrm{H}))$$

Weak Law of Likelihood (WLL) For mutually exclusive hypotheses $H_1, H_2 \in \mathfrak{L}$, $E \in \mathfrak{L}$ and $p \in \mathfrak{P}$, if

$$p(E|H_1) \geq p(E|H_2) \qquad \text{and} \qquad p(E|\neg H_1) \leq p(E|\neg H_2) \tag{9.4}$$

with one inequality being strict, then $C(H_1, E) > C(H_2, E)$.

The WLL has been defended as capturing a "core message of Bayes' Theorem" (Joyce, 2008): if H_1 predicts E better than H_2, and $\neg H_2$ predicts E better than $\neg H_1$, then E favors H_1 over H_2. Since WLL is phrased in terms of predictive performance, it is even more compelling for corroboration than for degree of confirmation. After all, $p(E| \pm H_1)$ and $p(E| \pm H_2)$ measure how well H_1 and H_2 have stood up to a test with outcome E. The version given here is in one sense stronger and in one sense weaker than Joyce's original formulation: it is stronger because only one inequality has to be strict (see also Brössel, 2013, 395–396); it is weaker because the WLL has been restricted to mutually exclusive hypotheses, where our intuitions tend to be more reliable.

The next condition deals with the role of irrelevant evidence in corroboration judgments:

Screened-Off Evidence Let $E_1, E_2, H \in \mathfrak{L}$ and $p \in \mathfrak{P}$. If E_2 is probabilistically independent of E_1, H, and $E_1 \wedge H$ and $p(E_2) > 0$, then $C(H, E_1) = C(H, E_1 \wedge E_2)$.

Structurally identical versions of this condition prominently figure in explications of confirmation and explanatory power (e.g., Kemeny and Oppenheim, 1952; Schupbach and Sprenger, 2011). It is a weaker version of condition (9.2) which demands, translated to corroboration, that $C(H, E) = C(H, E')$ if and only if $p(H|E) = p(H|E')$. To see this, just choose $E := E_1$, $E' := E_1 \wedge E_2$ and note that under the independence conditions of Screened-Off Evidence,

$$p(H|E_1 \wedge E_2) = \frac{p(H \wedge E_1|E_2)}{p(E_1|E_2)} = p(H|E_1)$$

which implies, by setting $p(E|\neg H)$ to its extremal values, the inequalities

$$p(E) \geq p(E|H)p(H) \qquad\qquad p(E) < p(E|H)p(H) + 1 - p(H).$$

Hence, anybody who accepts condition (9.2) for measures of corroboration also needs to endorse Screened-Off Evidence. However, Screened-Off Evidence is also very sensible on independent grounds: in an experiment where H has been tested and (relevant) evidence E_1 has been observed, completely irrelevant extra evidence ($E_2 \perp\!\!\!\perp E_1, H, E_1 \wedge H$) should not change the evaluation of the results. Imagine, for example, that a scientist tests the hypothesis that voices with high pitch are recognized more easily. As her university is interested in improving the planning of lab experiments, the scientist also collects data on when participants drop in, which days of the week are busy, which ones are quiet, etc. Plausibly, these data satisfy the independence conditions of Screened-Off Evidence. But equally plausibly, they do not influence the degree of corroboration of the hypothesis under investigation.

The next adequacy condition is motivated by the problem of irrelevant conjunctions for confirmation measures (e.g., Hawthorne and Fitelson, 2004). Assume that hypothesis H asserts the wave nature of light. Taken together with a body of auxiliary assumptions, H implies the phenomenon E: the interference pattern in Young's double slit experiment. Such an observation apparently corroborates the wave nature of light.

However, once we tack an utterly irrelevant proposition such as H′ = "the chicken came before the egg" to the hypothesis, it seems that E corroborates $H \wedge H'$—the conjunction of the wave theory of light and the chicken-egg hypothesis—not more than H, if at all. After all, H′ was in no way tested by the observations we made. It has no record of past performance to which we could appeal. This problem, familiar from Bayesian Confirmation Theory (see Variation 2), motivates the following constraint:

Irrelevant Conjunctions Assume the following conditions on $H, H', E \in \mathfrak{L}$ and $p \in \mathfrak{P}$ are satisfied:

(1) H and H′ are consistent and $p(H \wedge H') < p(H)$;

(2) $p(E) \in (0, 1)$;

(3) $H \models E$;

(4) $p(E|H') = p(E)$.

Then it is always the case that $C(H \wedge H', E) \leq C(H, E)$.

This requirement states that for any non-trivial hypothesis H′ that is consistent with H (condition 1) and irrelevant for E (condition 4), $H \wedge H'$ is

corroborated no more than H whenever H non-trivially entails E (conditions 2 and 3). A similar requirement has been defended for measures of empirical justification (Atkinson, 2012, 50–51). Indeed, it would be strange if corroboration (or justification) could be increased for free by attaching irrelevant conjunctions. That would also make it nearly impossible to reply persuasively to Duhem's problem, and to separate innocuous from blameworthy hypotheses. Degree of corroboration is supposed to guide our evaluation of hypotheses in the light of experimental results. But a measure which is invariant under logical conjunction of hypotheses (for deductively implied evidence) cannot fulfil this function.

Interestingly, the preceding adequacy conditions can be derived from Popper's original adequacy conditions (all proofs are given in the appendix):

Theorem 9.1 *The following statements are true:*

- *Popper's condition VII implies Weak Law of Likelihood for the case of equiprobable hypotheses.*
- *Popper's condition VII implies Screened-Off Evidence.*
- *Popper's condition VIIIc implies Irrelevant Conjunctions.*

This shows that our adequacy conditions are motivated in the right way: they are either weaker versions of Popper's criteria, or closely related to them. We can thus be confident that our formal analysis of corroboration is on target and that our adequacy conditions do not track a different, incompatible concept.

However, unlike confirmation, corroboration contains an element of severe testing: the hypothesis should run a risk of being falsified. High informativity and testability contribute to this goal. As Popper states, "in many cases, the more improbable [...] hypothesis is preferable" (Popper, 1979, 18–19), and the purpose of a measure of degree of corroboration is "to show clearly in which cases this holds and in which it does not hold" (ibid.). This motivates the following desideratum:

Weak Informativity Degree of corroboration $C(\mathrm{H},\mathrm{E})$ does not generally increase with the probability of H. That is, there are $\mathrm{H},\mathrm{H}',\mathrm{E} \in \mathfrak{L}$ and $p \in \mathfrak{P}$ such that

(1) $p(\mathrm{E}|\mathrm{H}) = p(\mathrm{E}|\mathrm{H}') > p(\mathrm{E})$;

(2) $1/2 \geq p(\mathrm{H}) > p(\mathrm{H}')$;

(3) $C(\mathrm{H}, \mathrm{E}) \leq C(\mathrm{H}', \mathrm{E})$.

The intuition behind Weak Informativity can also be expressed as follows: corroboration does not, in the first place, assess the probability of a hypothesis; therefore $C(\mathrm{H}, \mathrm{E})$ should not always increase with the probability of H. To this, the following condition—Strong Informativity—adds that low probability/high logical content can in principle be corroboration-conducive. Note that the requirement $1/2 \geq p(\mathrm{H}), p(\mathrm{H}')$ is purely technical and philosophically innocuous.

Strong Informativity The informativity/logical content of a proposition can increase degree of corroboration, ceteris paribus. That is, there are $\mathrm{H}, \mathrm{H}', \mathrm{E} \in \mathfrak{L}$ and $p \in \mathfrak{P}$ such that

(1) $p(\mathrm{E}|\mathrm{H}) = p(\mathrm{E}|\mathrm{H}') > p(\mathrm{E})$;

(2) $1/2 \geq p(\mathrm{H}) > p(\mathrm{H}')$;

(3) $C(\mathrm{H}, \mathrm{E}) < C(\mathrm{H}', \mathrm{E})$.

To our mind, any account of corroboration that denies these properties has stripped itself of its distinctive features with respect to degree of confirmation. At the very least, the Popperian characaterization of corroboration as capturing both predictive success and testability would have to be abandoned, and links with NHST would have to be loosened. The idea behind Strong/Weak Informativity has also recently been defended by Roberto Festa in his discussion of the "Reverse Matthew Effect": successful predictions reflect more favorably on powerful general theories than on restricted or weakened versions of them (Festa, 2012, 95–100). Note that neither Strong nor Weak Informativity postulates that corroboration decreases with prior probability; they just deny the "Matthew Effect" that corroboration co-varies with prior probability (see also Roche, 2014).

We will now demonstrate that the listed adequacy conditions are incompatible with each other. First, as a consequence of Weak Law of Likelihood, corroboration increases with the prior probability of a hypothesis. This clashes directly with Strong/Weak Informativity:

Theorem 9.2 *No measure of corroboration $C(\mathrm{H}, \mathrm{E})$ constructed according to Formality can satisfy Weak Law of Likelihood and Weak/Strong Informativity at the same time.*

Since Formality is a purely representational condition, this result means that Weak Law of Likelihood and Weak/Strong Informativity pull in different directions: the first condition emphasizes the predictive performance of the tested hypothesis, the second its logical strength. It is perhaps surprising that these two conditions are already incompatible, since it is a popular tenet of critical rationalism that informative hypotheses are also more valuable predictively.

Second, Strong Informativity clashes with Irrelevant Conjunctions and Screened-Off Evidence:

Theorem 9.3 *No measure of corroboration* $C(\mathrm{H},\mathrm{E})$ *constructed according to Formality can satisfy Screened-Off Evidence, Irrelevant Conjunctions and Strong Informativity at the same time.*

Thus, the intuition behind Strong Informativity cannot be satisfied if other plausible adequacy constraints on degree of corroboration are accepted. In particular, if a measure of corroboration is insensitive to irrelevant evidence and does not reward adding irrelevant conjunctions, then it cannot give any bonus to informative hypotheses. The less informative and testable a hypothesis is, the higher its degree of corroboration, ceteris paribus.

Finally, the result of Theorem 9.3 can be extended to Weak Informativity if we make the assumption that irrelevant conjunctions dilute the degree of corroboration, rather than not increasing it (proof omitted). See also the corresponding remark in the motivation of Irrelevant Conjunctions (page 218).

Note that these results are meaningful even for those who are not interested in the project of explicating Popperian corroboration (e.g., because they are radical subjective Bayesians). Some of the above adequacy conditions have been proposed for measures of confirmation or explanatory power as well; others could be potentially interesting in this context. For instance, Brössel (2013) has recently discussed the condition Continuity, which is similar to Strong/Weak Informativity: if the posterior probabilities of two hypotheses are almost indistinguishable from each other, we should prefer the hypothesis which was initially less probable. Hence, the above results are also meaningful in the framework of Bayesian Confirmation Theory: they indicate the impossibility of statistical relevance measures that capture informativity and predictive success at the same time.

All this does not yet imply that explicating degree of corroboration is a futile project. Rather, it reveals a fundamental and insoluble tension between the two main contributing factors of corroboration that Popper identifies: predictive success and testability/informativity. Weak Law of Likelihood, Screened-Off Evidence and Irrelevant Conjunctions all speak to the predictive success intuition, whereas Strong/Weak Informativity rewards informative and testable hypotheses. In other words, the pre-theoretic concept of corroboration is overloaded with desiderata that point in different directions and create insoluble tensions. The point of Theorem 9.2 and 9.3 is to lay bare these tensions and to suggest ways out of the dilemma. Basically, we have four options: (i) to reject one of the (substantial) adequacy conditions; (ii) to split up degree of corroboration into different sub-concepts that preserve subsets of these intuitions; (iii) to conclude that the explication of degree of corroboration is hopeless and not worthy of further pursuit, and (iv) to reconcile the various desiderata in a different mathematical and conceptual framework.

Option (i) would come down to either giving up Weak Law of Likelihood, Screened-Off Evidence, Irrelevant Conjunctions or Strong/Weak Informativity. But each of these adequacy conditions for degree of corroboration has been carefully motivated in the preceding section. Such a step would therefore appear arbitrary and unsatisfactory.

For example, one could propose to endorse a statistical relevance measure of degree of confirmation as a measure of corroboration, giving up the informativity intuition. This has the advantage of relating corroboration to a bunch of statistical and philosophical literature on degree of confirmation, but it comes at the price of stripping corroboration of its defining characteristics, and it runs into the objections presented in section 9.1.

Also, statistical relevance measures generally depend on $p(\mathrm{E}|\neg\mathrm{H})$, either explicitly or via the calculation of $p(\mathrm{E})$ and $p(\mathrm{H}|\mathrm{E})$. This creates a variety of problems. Consider, for example, a Binomial model where we test the null hypothesis $\mathrm{H}_0 : \theta = 0.5$ against the alternative $\mathrm{H}_1 : \theta \neq 0.5$. If the observed relative frequency of successes is close to 0.5, for example $\bar{x} = 0.53$, the degree of corroboration of the null hypothesis should not depend on the likelihoods $p(\bar{x}|\theta)$ for very large and very small values of θ. Such alternatives are logically possible, but apparently irrelevant for testing the *adequacy of the point null hypothesis* $\theta = 0.5$. However, for statistical relevance measures in the spirit of Bayesian Confirmation Theory,

this conclusion is inevitable since $p(\bar{x}|\theta \neq \theta_0) = \int_0^1 p(\bar{x}|\theta)\, p(\theta)\, d\theta$. The probability of the data under the alternative is just the weighted average of all the likelihoods.

Option (ii) amounts to endorsing pluralism for degree of corroboration. The model case for this option are probabilistic analyses of degre of confirmation: some measures, like $d(\mathrm{H}, \mathrm{E}) = p(\mathrm{H}|\mathrm{E}) - p(\mathrm{H})$ capture the boost in degree of belief in H provided by E, while others, like $l(\mathrm{H}, \mathrm{E}) = p(\mathrm{E}|\mathrm{H})/p(\mathrm{E}|\neg\mathrm{H})$, aim at the discriminatory power of E with respect to H and ¬H. However, it is not clear what similarly interesting subconcepts could look like for degree of corroboration. Right now, this option does not appear to be viable.

Neither does the pessimistic option (iii) have much appeal, unless convincing reasons are given why scientists can dispense with the concept of corroboration, and hypothesis testing in general.

This leaves us with option (iv): to change the mathematical framework for explicating degree of corroboration. Perhaps it is neither necessary nor sufficient to base a corroboration judgment on the joint probability distribution of H and E? As noted above, statistical relevance measures of corroboration compare the merits of H with the merits of ¬H, defined as the aggregate of alternatives to H. However, a comparison to such an aggregate does not make much sense in many NHST contexts where we deal with a multitude of distinct alternatives H_i, $i \in \mathbb{N}$. Perhaps corroboration judgments should be made with respect to the best-performing alternative in the hypothesis space, and not with respect to all possible alternatives. This is the option that we explore in the next section.

Toward a New Explication of Corroboration

Statistical relevance measures of corroboration compare the merits of H with the merits of ¬H, defined as the *aggregate* of alternatives to H. For example, in the above example from the Binomial model, with $\mathrm{H} : \theta = 0.5$ and $\mathrm{E} : \bar{x} = 0.53$, $p(\mathrm{E}|\neg\mathrm{H})$ would be equal to the value of $p(x|\theta \neq \theta_0) = \int_0^1 p(\theta)p(x|\theta)d\theta$. This is a property that the above statistical relevance measures of corroboration have in common with Bayesian confirmation measures, such as the Bayes factor.

However, a comparison to such an aggregate does not make much sense in many NHST contexts where we deal with a multitude of distinct

alternatives H_i, $i \in \mathbb{N}$. It seems to be essential that we have many alternatives in such a testing problem from which we can choose, and not just a "one-size-fits-all" probabilistic mixture of them. Perhaps degree of corroboration should be measured by comparing H_0 to the best available alternative, rather than to the collective of alternatives, which inevitably contains some very implausible hypotheses.

In the remainder of this section, we sketch an explication of degree of corroboration in a framework with **many distinct alternatives to the tested hypothesis** H_0. As a consequence, Formality has to be dropped and degree of corroboration becomes **partition-relative**: testing H_0 with alternative $\neg\mathrm{H}$ can lead to different corroboration judgments than testing H_0 with alternatives $\mathcal{H} = \{H_1, H_2, \ldots, H_n\}$ even if $\neg H = \bigvee_{1 \leq i \leq n} H_i$ (cf. Good, 1960, 1968b,a, 1975). Consider, for example, a test whether a medical drug is effective. The null corresponds to a particular parameter value $\mathrm{H}_0 : \theta = \theta_0$, indicating efficacy at placebo level, and the alternative to $\mathrm{H}_1 : \theta \neq \theta_0$. Dependent on the practical implications of certain effect sizes, we may divide the hypotheses in the following coarse-grained intervals: "worse than a placebo", "as good as a placebo", "slightly better than a placebo", "clearly better than a placebo", etc. Whereas in other testing contexts (e.g., determining the value of a natural constant), a very fine grained partition of the alternatives would seem more appropriate.

We now derive such a measure of corroboration on axiomatic grounds. In the explication, we focus exclusively on measuring past performance and neglect the testability intuition. It will resurface later, though. The first and most substantial requirement states that corroboration judgments are made with respect to the best-performing alternative in the hypothesis space, and not with respect to *all* possible alternatives.

CA1 Corroboration is the **minimal weight of evidence in favor of** H_0 when compared to all relevant alternatives, up to rescaling. That is, the degree of corroboration that E provides for H_0 relative to $\mathcal{H}$ can be defined as

$$C_{\mathcal{H}}(\mathrm{H}_0, \mathrm{E}) = \min_{h_i \in \mathcal{H}} W(\mathrm{H}_0, \mathrm{H}_\mathrm{i}, \mathrm{E}) \tag{9.5}$$

where $W(\mathrm{H}_0, \mathrm{H}_\mathrm{i}, \mathrm{E})$ quantifies the weight of evidence that E provides for H_0 and against the specific alternative H_i.

The idea is that positive corroboration requires that no genuine alternative $\mathrm{H}_\mathrm{i} \in \mathcal{H}$ be evidentially favored over H_0. On the weight of evidence

function W, we make the following constraints:

CA2 There exists a real-valued, continuous function $g : [0,1]^2 \to \mathbb{R}$ such that $W(\mathrm{E}, \mathrm{H}_0, \mathrm{H}_1) := g(p(\mathrm{E}|\mathrm{H}_0), p(\mathrm{E}|\mathrm{H}_1))$. In other words, weight of evidence only depends on the probability of E under H_0 and H_1.

The idea is that weight of evidence is a function of the predictive performance of both hypotheses, in line with Popper's characterization of corroboration as indicating past performance. Similar requirements are made in Good (1952); Bernardo (1999) and Williamson (2010). Finally, we make a convenience-based constraint on g: its range is normalized to [-1,1] and it should be represented as a rational function, in the mathematical sense of the word:

CA3 $g(x, y)$ is the simplest function of the form

$$g(x,y) = \frac{\sum_{j=1}^{m} \sum_{k=1}^{m} c_{jk}\, x^j\, y^k}{\sum_{j=1}^{n} \sum_{k=1}^{n} d_{jk}\, x^j\, y^k} \tag{9.6}$$

with the properties

$$g(1,0) = 1 \qquad\qquad g(0,1) = -1$$
$$g(x,x) = 0$$

From these constraints, we can derive

Theorem 9.4 *CA1–CA3 jointly determine the unique weight of evidence function*

$$W(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E}) = \frac{p(\mathrm{E}|\mathrm{H}_0) - p(\mathrm{E}|\mathrm{H}_\mathrm{i})}{p(\mathrm{E}|\mathrm{H}_0) + p(\mathrm{E}|\mathrm{H}_\mathrm{i})} \tag{9.7}$$

and the corroboration measure

$$C_{\mathcal{H}}(\mathrm{H}_0, \mathrm{E}) = \min_{H_i \in \mathcal{H}} \frac{p(\mathrm{E}|\mathrm{H}_0) - p(\mathrm{E}|\mathrm{H}_\mathrm{i})}{p(\mathrm{E}|\mathrm{H}_0) + p(\mathrm{E}|\mathrm{H}_\mathrm{i})} \tag{9.8}$$

In other words, we obtain the Kemeny-Oppenheim measure of (contrastive) confirmation, familiar from Variation 2, as a measure of weight of evidence. It is ordinally equivalent to the likelihood ratio, that is, the Bayes factor. This may be seen as the Bayesian foundation in the explication of corroboration. The corroboration measure itself is then equal to the degree of confirmation that H_0 obtains when pitched against the best-performing alternative. Positive degree of corroboration entails that there is no superior alternative.

We shall now apply this measure to our example of statistical inference in a Binomial model. A series of independent and identically distributed (i.i.d.) coin tosses is performed. In the figures below, we have plotted degree of corroboration as a function of the number of successes (e.g., "heads") for the null hypothesis $H_0 : \theta = 0.5$ and the sample size $N = 100$. The degree of corroboration is plotted as a function of the number of observed successes (x-axis), for three different partitions of the alternative hypotheses. The green dots correspond to classical Bayesian hypothesis testing with only a single alternative (=the probabilistic mixture of the H_i). The orange dots report the results for the best-performing alternative in the set of intervals $[0, 0.1)$, $[0.1, 0.2)$, etc. The blue dots, finally, are derived from the maximally fine-grained partition, that is, the best-performing point alternative in the interval $[0, 1]$. The left figure uses a uniform weighting of point values within the intervals that represent alternative hypotheses. The second figure uses a slightly centered weighting ($\beta(2, 2)$). As visible from the plots, there are no significant qualitative differences between the weightings, so we will disregard them from now onwards.

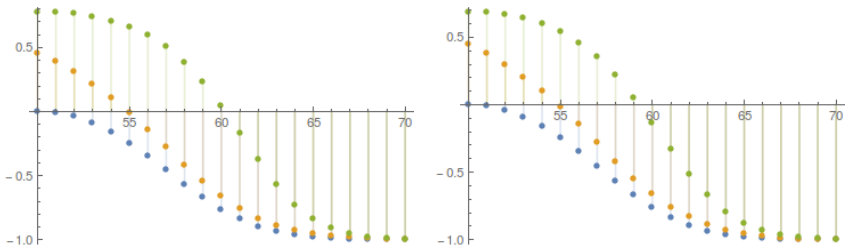

Figure 9.2: Degree of corroboration of the hypothesis $H_0 : \theta = \theta_0$ plotted against number of observed successes, for sample size $N = 100$. The green dots correspond to the alternative $\mathcal{H} = \{[0, 1]\}$. The orange dots correspond to $\mathcal{H} = \{[0; 0.1), [0.1, 0.2), \ldots\}$. The blue dots correspond to $\mathcal{H} = [0, 1]$. Left figure: weighting $\beta(1, 1)$; right figure: weighting $\beta(2, 2)$.

We see that for the coarse-grained partition (green dots), degree of corroboration is positive until $x = 60$: the performance of the alternative is dragged down by the extreme alternatives close to zero and one (their score is mixed with the well-fitting hypotheses). This is also the result yielded by the Bayes factor. Degree of corroboration diminishes if the alternatives are more fine-grained (e.g., for alternatives of the type $[0, 0.1)$, $[0.1, 0.2)$, etc.). The break-even point is $N = 55$. For the maximally fine-

grained alternative (=every point hypothesis is a potential alternative), degree of corroboration is always negative. This is actually very natural: when each parameter value is a serious scientific option, how can a point null hypothesis be ever corroborated unless the sample mean agrees *exactly* with the hypothesized parameter value?

These findings suggest that more fine-grained partitions lead to a smaller degree of corroboration, ceteris paribus. Indeed, we can verify this claim:

Theorem 9.5 *If $\mathcal{H}$ is a subpartition of $\mathcal{H}'$, then $C_{\mathcal{H}}(\mathrm{H}_0, \mathrm{E}) < C_{\mathcal{H}'}(\mathrm{H}_0, \mathrm{E})$, provided that the alternative hypotheses are weighted in the same way.*

This property shows that the testability of the alternatives affects the degree corroboration of the null hypothesis. If the alternatives are very specific and testable, the degree of corroboration of the null hypothesis is lower than if the alternatives are quite unspecific. Hence, Popper's two crucial aspects of corroboration—past performance and testability—have finally been reconciled, although not with respect to the null hypothesis itself, but with respect to the alternative hypothesis.

Discussion

After studying the formal properties of our corroboration measure, we now proceed to a philosophical evaluation. First, we list several essential features of $C_{\mathcal{H}}$ that distinguish it from p-values in NHST and Bayesian confirmation measures.

1. The explication of corroboration is sensitive to the partition of hypotheses against which the null is tested. This is the key conceptual move in this section. The Bayesian framework conceptualizes the alternative hypothesis as the probabilistic mixture of all point values different from $\theta = \theta_0$. However, we argue that it is often fruitful to think about the alternative as a set of hypotheses, e.g., intervals that correspond to a certain scientific conclusion (e.g., small/sizeable/very large effect).

2. The explication is entirely independent of raising one's confidence in the tested hypothesis and therefore distinctive of corroboration as opposed to confirmation ($\rightarrow$ Objection 1). The measure of corroboration is constructed as a comparison of the null hypothesis with the

best possible alternative. This is in agreement with scientific reasoning, where we accept a theory only if it outperforms the alternatives.

3. Hypotheses with prior probability zero can be corroborated as well ($\rightarrow$ Objection 2). That we are not prepared to bet on the truth of a precise point hypothesis, regardless of the betting odds, does not preclude that this hypothesis can perform well and be corroborated with respect to certain competitors.

4. The explication is asymmetric, respecting that the role of the null hypothesis and the alternative are not interchangeable ($\rightarrow$ Objection 3). This preserves an important feature of NHST without buying into their methodological flaws.

Notably, subjective elements are still present in corroboration-based hypothesis testing. First and foremost, in the partitioning of alternatives. To what extent is the null hypothesis a good idealization of reality, and what are the error margins that we are willing to accept? What is a scientifically meaningful effect size, and which differences can be neglect? Second, in the weighting of point hypotheses *within* the alternatives. This may be negligible for very fine-grained hypotheses, but it may substantially affect the outcome for fairly coarse-grained partitions. Third, the Bayes factor emerges as the degree of corroboration for a maximally coarse-grained partition ($\mathcal{H} = \neg \mathrm{H}_0$). In other words, Bayesian hypothesis testing can be represented as a special case of evaluating hypothesis tests in terms of degrees of corroboration.

This brings us to questions for future research. An obvious project is the aforementioned reconciliation of Bayesian inference and NHST within a corraboration-centered perspective, and in particular, examining the hypothesis that our explication of corroboration unifies Bayesian and non-Bayesian hypothesis testing. Second, the proposed corroboration measure needs to be applied to more complicated cases of statistical inference, including nuisance parameters, hierarchical models and model selection. In this context, it would also be challenging to see what kind of meaning the notorious p-value obtains within a corroboration-based framework. Third, one may conduct case studies that reconstruct specific episodes of scientific reasoning as guided by corroboration judgments, and to see whether these episodes fit into a probabilistic explication of degree of corroboration.

The next variation stays with the topic of statistical inference and focuses on the problem of model selection: comparing classes of statistical hypotheses that are parametrized by one or several variables. Similar to the question of whether testability is conducive to degree of corroboration, investigated in this variation, we will ask the question of whether simplicity should be a cognitive value in model selection.

Proofs of the Theorems

Proof of Theorem 9.1: We begin with showing that condition VII implies the Weak Law of Likelihood (WLL). Assume $p(\mathrm{H}_1) = p(\mathrm{H}_2)$. We distinguish two jointly exhaustive cases in which WLL may apply:

Case 1: $p(\mathrm{E}|\mathrm{H}_1) > p(\mathrm{E}|\mathrm{H}_2)$ Case 2: $p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2)$ and $p(\mathrm{E}|\neg\mathrm{H}_1) < p(\mathrm{E}|\neg\mathrm{H}_2)$.

For the first case, the proof is simple in virtue of the inequality

$$p(\mathrm{H}_1|\mathrm{E}) = p(\mathrm{H}_1)\frac{p(\mathrm{E}|\mathrm{H}_1)}{p(\mathrm{E})} > p(\mathrm{H}_2)\frac{p(\mathrm{E}|\mathrm{H}_2)}{p(\mathrm{E})} = p(\mathrm{H}_2|\mathrm{E}).$$

Then, VII guarantees that $c(\mathrm{H}_1, \mathrm{E}) > c(\mathrm{H}_2, \mathrm{E})$.

For the second case, let $x := p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2)$ and $y := p(\mathrm{H}_1) = p(\mathrm{H}_2)$. We know that

$$\begin{aligned} p(\mathrm{E}|\neg\mathrm{H}_1) &= \frac{1}{1-p(\mathrm{H}_1)}\left[p(\mathrm{E}|\mathrm{H}_2)p(\mathrm{H}_2) + p(\mathrm{E}|\neg\mathrm{H}_1,\neg\mathrm{H}_2)p(\neg\mathrm{H}_1,\neg\mathrm{H}_2)\right] \\ &= \frac{1}{1-y}(xy + p(\mathrm{E}|\neg\mathrm{H}_1,\neg\mathrm{H}_2)p(\neg\mathrm{H}_1,\neg\mathrm{H}_2)) \\ p(\mathrm{E}|\neg\mathrm{H}_2) &= \frac{1}{1-p(\mathrm{H}_2)}\left[p(\mathrm{E}|\mathrm{H}_1)p(\mathrm{H}_1) + p(\mathrm{E}|\neg\mathrm{H}_1,\neg\mathrm{H}_2)p(\neg\mathrm{H}_1,\neg\mathrm{H}_2)\right] \\ &= \frac{1}{1-y}(xy + p(\mathrm{E}|\neg\mathrm{H}_1,\neg\mathrm{H}_2)p(\neg\mathrm{H}_1,\neg\mathrm{H}_2)). \end{aligned}$$

Hence, $p(\mathrm{E}|\neg\mathrm{H}_1) = p(\mathrm{E}|\neg\mathrm{H}_2)$. On the other hand, we have assumed that $p(\mathrm{E}|\neg\mathrm{H}_1) < p(\mathrm{E}|\neg\mathrm{H}_2)$. This shows that the second case can never occur and may be dismissed.

We now prove the second implication, that is, VII $\Rightarrow$ Screened-Off Evidence. To this end, remember that condition VII reads

VII If $p(\mathrm{H}) = p(\mathrm{H}')$, then $c(\mathrm{H}, \mathrm{E}) \leq c(\mathrm{H}', \mathrm{E}')$ if and only if $p(\mathrm{H}|\mathrm{E}) \leq p(\mathrm{H}'|\mathrm{E}')$.

Assuming H = H′, it is easy to see that VII implies

VII′ If $p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H}|\mathrm{E}')$, then $c(\mathrm{H}, \mathrm{E}) = c(\mathrm{H}, \mathrm{E}')$.

The reason is simple: If $p(\mathrm{H}|\mathrm{E}) = p(\mathrm{H}|\mathrm{E}')$, then also $p(\mathrm{H}|\mathrm{E}) \leq p(\mathrm{H}|\mathrm{E}')$ and the '$\Leftarrow$' direction of VII implies $c(\mathrm{H}, \mathrm{E}) \leq c(\mathrm{H}, \mathrm{E}')$, where H has been substituted for H'. Now we repeat the same trick with the premise $p(\mathrm{H}|\mathrm{E}') \leq p(\mathrm{H}|\mathrm{E})$ and we obtain $c(\mathrm{H}, \mathrm{E}') \leq c(\mathrm{H}, \mathrm{E})$. Taking both inequalities together yields the conclusion $c(\mathrm{H}, \mathrm{E}) = c(\mathrm{H}, \mathrm{E}')$ and thereby VII'.

Notice that under the conditions of Screened-Off Evidence, $p(h|\mathrm{E}_1 \wedge \mathrm{E}_2) = p(\mathrm{H}|\mathrm{E}_1)$. This is so because

$$\begin{aligned} p(h|\mathrm{E}_1 \wedge \mathrm{E}_2) &= p(\mathrm{H}) \frac{p(\mathrm{E}_1 \wedge \mathrm{E}_2|\mathrm{H})}{p(\mathrm{E}_1 \wedge \mathrm{E}_2)} \\ &= p(\mathrm{H}) \frac{p(\mathrm{E}_1|\mathrm{H})\, p(\mathrm{E}_2)}{p(\mathrm{E}_1)\, p(\mathrm{E}_2)} = p(\mathrm{H}) \frac{p(\mathrm{E}_1|\mathrm{H})}{p(\mathrm{E}_1)} = p(\mathrm{H}|\mathrm{E}_1). \end{aligned}$$

Hence, we can apply VII' to the case of Screened-Off Evidence, with $e := e_1$ and $e' := \mathrm{E}_1 \wedge \mathrm{E}_2$. This implies

$$c(\mathrm{H}, \mathrm{E}_1 \wedge \mathrm{E}_2) = c(\mathrm{H}, \mathrm{E}_1),$$

completing the proof.

Finally, we have the implication VIIIc $\Rightarrow$ Irrelevant Conjunctions. Let for $\mathrm{H}, \mathrm{H}', \mathrm{E} \in \mathfrak{L}$ and $p \in \mathfrak{P}$ the conditions of Irrelevant Conjunctions ([1] to [4]) be satisfied. Since $\mathrm{H} \models \mathrm{E}$, VIIIc implies that $c(\mathrm{H}, \mathrm{E})$ and $c(H \wedge \mathrm{H}', \mathrm{E})$ are increasing functions of the probability of the tested hypothesis—$p(\mathrm{H})$ and $p(\mathrm{H} \wedge \mathrm{H}')$, respectively. But by assumption, we have $p(\mathrm{H} \wedge \mathrm{H}') < p(\mathrm{H})$. Hence, it follows that $c(\mathrm{H} \wedge \mathrm{H}', \mathrm{E}) \leq c(\mathrm{H}, \mathrm{E})$. □

Proof of Theorem 9.2: By Weak Informativity and Formality, there are $x > y$ and $z > z'$ with $z + z' < 1$, $1 + xz - z \geq y \geq xz$ and $1 + xz' - z' \geq y \geq xz'$ such that

$$f(x, y, z) \leq f(x, y, z').$$

Choose a probability function p such that $p(\mathrm{H}_1) = z$, $p(\mathrm{H}_2) = z'$, $p(\mathrm{H}_1 \wedge \mathrm{H}_2) = 0$, $p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2) = x$, $p(\mathrm{E}) = y$. We now verify that this distribution satisfies the axioms of probability. Because of $xz > xz'$ and $1 + xz - z < 1 + xz' - z'$, it suffices to verify the inequalities $y \geq xz$ and $y \leq 1 + xz - z$.

First note that

$$p(\mathrm{E}) = p(\mathrm{E}|\mathrm{H}_1)p(\mathrm{H}_1) + p(\mathrm{E}|\mathrm{H}_2)p(\mathrm{H}_2) + p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)(1 - p(\mathrm{H}_1) - p(\mathrm{H}_2))$$

which translates, setting $\omega := p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)$, as

$$y = xz + xz' + \omega(1 - z - z').$$

This equation allows us to show the desired inequalities:

$$\begin{aligned} y - xz &= xz + xz' + \omega(1 - z - z') - xz \\ &= xz' + \omega(1 - z - z') \\ &\geq 0 \\ 1 + xz - z - y &= 1 + xz - z - xz - xz' - \omega(1 - z - z') \\ &= (1 - z - xz') + \omega(1 - z - z') \\ &\geq 0 \end{aligned}$$

In both cases, all summands are greater or equal than zero because $z + z' < 1$ by assumption. This completes the proof that the above probability distribution is well-defined.

Now it is straightforward to show that

$$\begin{aligned} p(\mathrm{E}|\neg\mathrm{H}_1) &= \frac{1}{1 - p(\mathrm{H}_1)} \left[p(\mathrm{E}|\mathrm{H}_2)p(\mathrm{H}_2) + p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)p(\neg\mathrm{H}_1, \neg\mathrm{H}_2)\right] \\ &= \frac{1}{1 - p(\mathrm{H}_1)} \left[p(\mathrm{E}|\mathrm{H}_1)p(\mathrm{H}_2) + p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)p(\neg\mathrm{H}_1, \neg\mathrm{H}_2)\right] \\ p(\mathrm{E}|\neg\mathrm{H}_2) &= \frac{1}{1 - p(\mathrm{H}_2)} \left[p(\mathrm{E}|\mathrm{H}_1)p(\mathrm{H}_1) + p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)p(\neg\mathrm{H}_1, \neg\mathrm{H}_2)\right] \end{aligned}$$

because by assumption, $p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2)$. From this we can infer

$$\begin{aligned} & p(\mathrm{E}|\neg\mathrm{H}_1) - p(\mathrm{E}|\neg\mathrm{H}_2) \\ = \; & \frac{p(\mathrm{E}|\mathrm{H}_1)\, p(\mathrm{H}_2)}{1 - p(\mathrm{H}_1)} + \frac{p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)\, p(\neg\mathrm{H}_1, \neg\mathrm{H}_2)}{1 - p(\mathrm{H}_1)} - \frac{p(\mathrm{E}|\mathrm{H}_1)\, p(\mathrm{H}_1)}{1 - p(\mathrm{H}_2)} \\ & - \frac{p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)\, p(\neg\mathrm{H}_1, \neg\mathrm{H}_2)}{1 - p(\mathrm{H}_2)} \\ = \; & p(\mathrm{E}|\mathrm{H}_1) \left[\frac{p(\mathrm{H}_2)}{1 - p(\mathrm{H}_1)} - \frac{p(\mathrm{H}_1)}{1 - p(\mathrm{H}_2)}\right] + p(\mathrm{E}|\neg\mathrm{H}_1\neg h_2)(1 - p(\mathrm{H}_1) - p(\mathrm{H}_2)) \end{aligned}$$

$$\cdot \left[\frac{1}{1-p(\mathrm{H}_1)} - \frac{1}{1-p(\mathrm{H}_2)}\right]$$

$$= p(\mathrm{E}|\mathrm{H}_1)\frac{p(\mathrm{H}_2) - p(\mathrm{H}_2)^2 - p(\mathrm{H}_1) + p(\mathrm{H}_1)^2}{(1-p(\mathrm{H}_1))\,(1-p(\mathrm{H}_2))} + p(\mathrm{E}|\neg\mathrm{H}_1\neg h_2)$$

$$\cdot (1 - p(\mathrm{H}_1) - p(\mathrm{H}_2))\,\frac{p(\mathrm{H}_1) - p(\mathrm{H}_2)}{(1-p(\mathrm{H}_1))\,(1-p(\mathrm{H}_2))}$$

$$= p(\mathrm{E}|\mathrm{H}_1)\frac{(p(\mathrm{H}_1) - p(\mathrm{H}_2)) \cdot (p(\mathrm{H}_1) + p(\mathrm{H}_2) - 1)}{(1-p(\mathrm{H}_1))\,(1-p(\mathrm{H}_2))} + p(\mathrm{E}|\neg\mathrm{H}_1\neg h_2)$$

$$\cdot (1 - p(\mathrm{H}_1) - p(\mathrm{H}_2))\,\frac{p(\mathrm{H}_1) - p(\mathrm{H}_2)}{(1-p(\mathrm{H}_1))\,(1-p(\mathrm{H}_2))}$$

$$= \frac{(p(\mathrm{H}_1) - p(\mathrm{H}_2)) \cdot (p(\mathrm{H}_1) + p(\mathrm{H}_2) - 1)}{(1-p(\mathrm{H}_1))\,(1-p(\mathrm{H}_2))}\,(p(\mathrm{E}|\mathrm{H}_1) - p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)).$$

If we look at the signs of the involved factors, we notice first that $p(\mathrm{H}_1) - p(\mathrm{H}_2) = z - z' > 0$ and $p(\mathrm{H}_1) + p(\mathrm{H}_2) - 1 = z + z' - 1 < 0$. Then we observe that H_1 and H_2 were disjoint and that $p(\mathrm{E}|\mathrm{H}_1)$ and $p(\mathrm{E}|\mathrm{H}_2)$ are both greater than $p(\mathrm{E})$, implying $p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2) > p(\mathrm{E}|\neg\mathrm{H}_1, \neg\mathrm{H}_2)$. Taken together, we can then conclude

$$p(\mathrm{E}|\neg\mathrm{H}_1) - p(\mathrm{E}|\neg\mathrm{H}_2) < 0.$$

Hence, the conditions for applying Weak Law of Likelihood are satisfied: H_1 and H_2 are two mutually exclusive hypotheses with $p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2)$ and $p(\mathrm{E}|\neg\mathrm{H}_1) < p(\mathrm{E}|\neg\mathrm{H}_2)$. Thus we can conclude

$$f(x, y, z) = c(\mathrm{H}_1, \mathrm{E}) > c(\mathrm{H}_2, \mathrm{E}) = f(x, y, z'),$$

in contradiction with the inequality $f(x, y, z) \leq f(x, y, z')$ that we got from Weak Informativity. □

Lemma 1 Any measure of corroboration $c : \mathfrak{L}^2 \times \mathfrak{P} \to \mathbb{R}$ that satisfies Screened-Off Evidence and Formality also satisfies the equality

$$f(ax, ay, z) = f(x, y, z) \tag{9.9}$$

for $x > y > 0$, $z > 0$ and $0 < a \leq 1$ with $1 + xz - z \geq y \geq xz$.

Proof of Lemma 1: For any $0 < a \leq 1$, $x > y > 0$ and $z > 0$ with $1 + xz - z \geq y \geq xz$, we can choose sentences $\mathrm{H}, \mathrm{E}_1, \mathrm{E}_2 \in \mathfrak{L}$ and a probability function $p \in \mathfrak{P}$ such that

$$\begin{aligned} a &:= p(\mathrm{E}_2) & p(\mathrm{E}_2, \mathrm{H}) &= p(\mathrm{E}_2)p(\mathrm{H}) \\ x &:= p(\mathrm{E}_1|\mathrm{H}) & p(\mathrm{E}_1 \wedge \mathrm{E}_2) &= p(\mathrm{E}_2)p(\mathrm{E}_1) \\ y &:= p(\mathrm{E}_1) & p(\mathrm{E}_1 \wedge \mathrm{E}_2|\mathrm{H}) &= p(\mathrm{E}_2)p(\mathrm{E}_1|\mathrm{H}) \\ z &:= p(\mathrm{H}). \end{aligned}$$

Since our choice of p is not restricted, this is always possible. Now, the conditions of Screened-Off Evidence are satisfied, and it follows that $c(\mathrm{H}, \mathrm{E}_1 \wedge \mathrm{E}_2)) = c(\mathrm{H}, \mathrm{E}_1)$. By Formality, we can also derive the equalities

$$\begin{aligned} c(\mathrm{H}, \mathrm{E}_1 \wedge \mathrm{E}_2)) &= f(p(\mathrm{E}_1 \wedge \mathrm{E}_2|\mathrm{H}), p(\mathrm{E}_1 \wedge \mathrm{E}_2), p(\mathrm{H})) \\ &= f(p(\mathrm{E}_2)p(\mathrm{E}_1|\mathrm{H}), p(\mathrm{E}_2)p(\mathrm{E}_1), p(\mathrm{H})) \\ &= f(ax, ay, z) \\ c(\mathrm{H}, \mathrm{E}_1) &= f(x, y, z). \end{aligned}$$

Taking all these equalities together delivers the desired result:

$$f(ax, ay, z) = c(\mathrm{H}, \mathrm{E}_1 \wedge \mathrm{E}_2)) = c(\mathrm{H}, \mathrm{E}_1) = f(x, y, z).$$

Finally we note that (ax, ay, z) is always in the domain of f when $a \leq 1$ and $1 + xz - z \geq y \geq xz$:

$$\begin{aligned} (ay) &\geq (ax)/z & ay &\leq a(1 + xz - z) \\ & & &= axz + a(1 - z) \\ & & &\leq 1 + (ax)z - z \end{aligned}$$

□

Proof of Theorem 9.3: Choose sentences $\mathrm{H}_1, \mathrm{H}_2, \mathrm{E} \in \mathfrak{L}$ and a probability function $p \in \mathfrak{P}$ such that the conditions of Strong Informativity are satisfied:

(1) $p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2) > p(\mathrm{E})$;

(2) $1/2 \geq p(\mathrm{H}_1) > p(\mathrm{H}_2)$;

(3) $c(\mathrm{H}_1, \mathrm{E}) < c(\mathrm{H}_2, \mathrm{E})$.

Writing $x := p(\mathrm{E}|\mathrm{H}_1) = p(\mathrm{E}|\mathrm{H}_2)$, $y := p(\mathrm{E})$, $z = p(\mathrm{H})$ and $z' := p(\mathrm{H}')$, we then obtain

$$f(x, y, z) = c(\mathrm{H}_1, \mathrm{E}) < c(\mathrm{H}_2, \mathrm{E}) = f(x, y, z'). \tag{9.10}$$

Since $c(\mathrm{H}, \mathrm{E})$ satisfies Formality and Screened-Off Evidence, by Lemma 1 it also satisfies the equality

$$f(ax, ay, z) = f(x, y, z)$$

for $x > y > 0$, $z > 0$ and $0 < a \leq 1$. It is easy to see that $(1, y/x, z)$ is in the domain of f if (x, y, z) is. Applying the above equality to $f(1, y/x, z)$ and choosing $a := x$, we now obtain

$$f(1, y/x, z) = f(x, y, z) \qquad f(1, y/x, z') = f(x, y, z').$$

Then it follows from inequality (9.10) and the above equalities that

$$f(1, y/x, z) < f(1, y/x, z') \tag{9.11}$$

for these specific values of x, y, z and z'.

We can now find sentences H, H′, E′ and a probability function $p'(\cdot)$ such that the conditions of Irrelevant Conjunctions are satisfied and at the same time, $p'(\mathrm{H}) = z$, $p'(\mathrm{H} \wedge \mathrm{H}') = z'$, $p'(\mathrm{E}') = y/x$. This implies $c(\mathrm{H} \wedge \mathrm{H}', \mathrm{E}') \leq c(\mathrm{H}, \mathrm{E}')$. By Formality, this also implies

$$f(1, y/x, z) \geq f(1, y/x, z').$$

However, this inequality contradicts Equation (9.11) that we have shown before. Hence, the theorem is proven. □

Proof of Theorem 9.4: We have to show that $W(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E})$ is indeed of the form (9.7). From CA2, we know that it must be a function of $p(\mathrm{E}|\mathrm{H}_0)$ and $p(\mathrm{E}|\mathrm{H}_1)$. Now we use CA3 to prove its specific from.

Assume first that $m = 0$ and $n = 1$. In that case, the neutrality condition $f_S(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E}) = 0$ if $p(\mathrm{E}|\mathrm{H}_0) = p(\mathrm{E}|\mathrm{H}_1)$ cannot be satisfied unless $c_{00} = 0$ because the numerator is a constant. Hence, we can neglect this possibility.

Now assume that $m = 1$ and $n = 0$. Here, the neutrality condition $f_S(\mathrm{h}_0, \mathrm{H}_1, \mathrm{E}) = 0$ if $p(\mathrm{E}|\mathrm{H}_0) = p(\mathrm{E}|\mathrm{H}_1)$ leads to the equation

$$c_{00} + (c_{10} + c_{01})p(\mathrm{E}|\mathrm{H}_0) + c_{11}p(\mathrm{E}|\mathrm{H}_0)^2 = 0 \tag{9.12}$$

which is satisfied in general if and only if $c_{00} = c_{11} = 0$ and $c_{10} = -c_{01}$. Clearly, the resulting function $f(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E}) = p(\mathrm{E}|\mathrm{H}_0) - p(\mathrm{E}|\mathrm{H}_1)$ is not ordinally equivalent to $S(\mathrm{H}_0, \mathrm{E}) - S(\mathrm{H}_1, \mathrm{E}) = \log p(\mathrm{E}|\mathrm{H}_0) - \log p(\mathrm{E}|\mathrm{H}_1)$, regardless of the value of c_{10} and the base of the logarithm. Hence, we can neglect this possibility, too.

Now assume that $m = n = 1$. Again, the neutrality condition leads to the conclusion $c_{00} = c_{11} = 0$ and $c_{10} = -c_{01}$. Now, let us set $p(\mathrm{E}|\mathrm{H}_0) = 1$, $p(\mathrm{E}|\mathrm{H}_1) = 0$, and vice versa. Then, the maximality constraint implies $d_{10} = d_{01} = 1$ and the simplest function that maintains ordinal equivalence with $S(\mathrm{H}_0, \mathrm{E}) - S(\mathrm{H}_1, \mathrm{E})$, as demanded by CA1′, is obtained by setting $d_{00} = d_{11} = 0$.

From this result, the theorem follows by a simple application of CA1. □

Proof of Theorem 9.5: Without loss of generality, we can restrict ourselves to the case of two disjoint alternatives $\mathcal{H} = \{\mathrm{H}_1 \vee \mathrm{H}_2\}$, $\mathcal{H}' = \{\mathrm{H}_1, \mathrm{H}_2\}$, with $p(\mathrm{E}|\mathrm{H}_1) > p(\mathrm{E}|\mathrm{H}_2)$, This is because the relative weighting of the elements of H_1 and H_2 stays the same. Now we observe

$$\begin{aligned} p(\mathrm{E}|\mathrm{H}_1 \vee \mathrm{H}_2) &= \frac{1}{p(\mathrm{H}_1 \vee \mathrm{H}_2)} \left(p(\mathrm{E}|\mathrm{H}_1)p(\mathrm{H}_1) + p(\mathrm{E}|\mathrm{H}_2)p(\mathrm{H}_2)\right) \\ &> \frac{1}{p(\mathrm{H}_1) + p(\mathrm{H}_2)} \left(p(\mathrm{E}|\mathrm{H}_2)p(\mathrm{H}_1) + p(\mathrm{E}|\mathrm{H}_2)p(\mathrm{H}_2)\right) \\ &= p(\mathrm{E}|\mathrm{H}_2) \end{aligned}$$

Hence $C_{\mathcal{H}}(\mathrm{H}_0, \mathrm{E}) > C_{\mathcal{H}'}(\mathrm{H}_0, \mathrm{E})$, irrespective of the value of $p(\mathrm{E}|\mathrm{H}_0)$. □

Variation 10: Simplicity and Model Selection

> "Numquam ponenda est pluralitas sine necessitate."
> (William of Occam)

Is simplicity a cognitive value? Is it indicative of a good scientific theory? Few questions in philosophy of science are older, and few have been debated more controversially. The thesis that simple demonstrations, scientific theories or ontological systems are more valuable than complex ones has already been defended by great philosophical minds such as Aristotle, Aquinas and Kant, e.g.:

> If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices. (Aquinas, 1945, 129)

Indeed, the belief that simplicity is a **cognitive value** is often backed by an ontological assumption that among different ways nature could be, the simple one is more likely to be true. In a weaker version of that thesis, simplicity is not necessarily seen as truth-conducive, but as contributing to the success, verisimilitude, predictive accuracy or rational acceptability of a theory. Thomas S. Kuhn (1977a) also includes simplicity in the list of standard criteria for scientific theory choice.

Opponents reply that this belief is unjustified: we have no reason to assume that nature is simple rather than complex. On that view, simplicity is nothing more than a **pragmatic value** related to our cognitive limitations as human beings (e.g., van Fraassen, 1980). Simple theories are easier to handle than complex ones, be it for purposes of prediction, explanation, or further theoretical development. Thus, to what extent is simplicity related to the success of science?

To answer this question from a Bayesian perspective, we have to distinguish between different dimensions of simplicity in scientific inference. The first distinction concerns the **syntactic** and the **semantic dimension of simplicity**. The semantic dimension is concerned with the ontological implications of that theory. How many entities does the theory postulate? Are they all of the same kind? And so on. This dimension of simplicity is called **parsimony** (Nolan, 1997; Baker, 2003, 2010). Again, the thesis that parsimonious theories are to prefer to less parsimonious theories has two aspects, one pertaining to the number of entities postulated by the theory in question, and one pertaining to plurality in the kinds of postulated entities. Both of them are fundamental questions in the metaphysics of science. We feel that there is little that the Bayesian framework—which is primarily a tool for uncertain reasoning—can contribute to deciding these questions, and leave them aside in what follows.

The syntactic dimension of simplicity is more interesting for our purposes. It deals with the way scientific theories are formulated: How many hypotheses are postulated? How complex are they? Can they be related to each other in a straightforward way? Discussions about the role of simplicity in curve-fitting and other everyday elements of scientific practice belong into this systematic dimension which the survey article by Baker (2010) calls **elegance**. In the remainder of the variation, whenever we write "simplicity", we refer to the syntactic dimension of simplicity as elegance. We focus on Bayesian accounts of simplicity, that is, on the question of whether it is rational to prefer simpler theories to more complex ones for purely epistemic reasons. These accounts will also be contrasted with non-Bayesian explications of simplicity in model selection.

Simplicity in Model Selection

The debate about simplicity as elegance has, in recent decades, focused on the role of simplicity in model selection. This involves the comparison of different statistical models on the basis of their fit with a dataset and the intrinsic properties of these models. A nice feature of model selection is that simplicity can be quantified neatly, in terms of the number of free parameters that a statistical model posits. This understanding of simplicity is also manifest in various model selection criteria and allows for a quite rigorous treatment of the role of simplicity. Here and in the sequel, the

term **model selection** is used as shorthand for the more general term of statistical model comparison or model evaluation. That is, model selection need not refer to a choice of the modeler, or a decision to reject/accept a particular model. It can also lead to a judgment about which model is, all things taken together, superior to its competitors.

In the last 20 years, there has been a boom of papers on the value of simplicity in statistical inference, prompted by Forster and Sober's 1994 paper in *BJPS*. In that paper, the authors challenge van Fraassen's view that simpler models are preferred solely on non-empirical, pragmatic grounds, such as mathematical convenience. They substantiate their thesis with the help of the Akaike Information Criterion (AIC), a statistical model selection criterion that involves a tradeoff of simplicity and goodness-of-fit. Along these lines, they argue that the simplicity of a model contributes to its predictive success.

In Forster and Sober's view, the real epistemological question surrounding simplicity is not whether simple models are more likely to be true, but whether they are **predictively accurate** (Forster, 2002; Sober, 2002). After all, in modeling economic growth, climate change, social decision-making, etc., statistical models are just idealizations of an excessively complex reality. Simple models, however, may capture salient aspects of reality, whereas complicated models only muddle the waters by introducing effects that do not really matter. According to this argument, simplicity is a genuine cognitive value because it contributes to attaining another cognitive value, namely predictive accuracy, whose centrality for the scientific enterprise stands undisputed (Kuhn, 1977a; McMullin, 1982, 2008; Douglas, 2013).

We can distinguish two questions regarding the relation between simplicity and predictive accuracy:

1. The *qualitative* question: Do simple models tend to be more predictively accurate, ceteris paribus, thereby vindicating simplicity as a genuine cognitive value?

2. The *quantitative* question: What is the weight of simplicity vis-à-vis other cognitive values (e.g., goodness-of-fit) in model selection)?

In this variation, we shall argue for an affirmative answer to the first question. Forster and Sober, however, go beyond this: they argue that the

epistemic pull of simplicity is established by means of the mathematical properties of a *particular* model comparison criterion, Akaike's information criterion AIC. By contrast, we argue that a tradeoff between simplicity and goodness-of-fit is bound to be highly context-dependent and cannot be captured by a single criterion.

The rest of the variation is structured into three major sections: one dealing with the qualitative rationale for preferring simpler models in model selection (Section 10.2), and two dealing with the quantitative tradeoff rate between simplicity and goodness-of-fit. The first of these sections reviews a non-Bayesian model selection criterion (Section 10.3), the second a Bayesian model selection criterion (Section 10.4). We also investigate whether there is a specific Bayesian way to make simplicity matter, that is, a model selection criterion where a direct link between simplicity and posterior probability can be established. We observe that Bayesian inference often plays an instrumental role in model selection, rather than providing a genuine philosophical underpinning of a particular procedure. Finally, we summarize our findings (Section 10.5) and provide additional mathematical details (Section 10.6).

Curve Fitting and Estimation Error

Statistical model analysis compares a large set of candidate models to given data, in the hope to select the best model on which predictions, explanations and further inferences can be based. Often, it is unrealistic to assume that the "true model" (i.e., the data-generating process) is found among the candidate models: data sets are often huge and messy, the underlying processes are complex and hard to describe theoretically, and contain lots of noise and confounding factors. Furthermore, the candidate models are often generated by automatic means (e.g., as linear combinations of potential predictor variables). This means that they usually do not provide the most striking mechanism, or the best scientific explanation of the data. Rather, they are constructed from the data, supposed to explain them reasonably well and to be a reliable device for future predictions. This "bottom-up" approach (Burnham and Anderson, 2002; Sober, 2002) to curve-fitting and model-building is complementary to a "top-down" approach where complex models are derived from general theoretical principles (Weisberg, 2007).

How does this work in a concrete case? Consider fitting a scatterplot such as Figure 10.1 with polynomial curves. Assume that we describe the relationship between an independent ("input") variable x and a dependent ("output") variable y either by a linear or by a quadratic polynomial, together with i.i.d. noise terms ε_i (e.g., $\varepsilon_i \sim N(0,1)$) and coefficients α, β, and γ. Then, the data points $D = \{(x_i, y_i), 1 \leq i \leq N\}$ are described by the equations

$$\text{(LIN)} \qquad y_i = \alpha + \beta x_i + \varepsilon_i \tag{10.1}$$

$$\text{(PAR)} \qquad y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i \tag{10.2}$$

The linear model (LIN) now corresponds to the null hypothesis $H_0 : \gamma = 0$ and the quadratic model (PAR) to the more general alternative $H_1 : \gamma \neq 0$. The ordinary method for fitting the linear model to the data is the method of ordinary least squares (OLS): the parameters $\hat{\alpha}$ and $\hat{\beta}$ are chosen such that the thus-defined curve $y = \hat{\alpha} + \hat{\beta}x$ makes the data $D = (x_i, y_i)$ most likely among all pairs (α, β). If the error terms are i.i.d. and follow a Gaussian (=Normal) distribution, this is equivalent to minimizing the sum of the square of the residuals, $\sum_i (y_i - \alpha - \beta x_i)^2$, that is, the variation in the data that cannot be explained by assuming the model (LIN).

For simple linear regression, the values (α, β) that minimize the above sum can be calculated analytically: $\hat{\beta} = Cov(X, Y)/Var(X)$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. For more complex regression models, numerical methods are normally used. But the idea stays the same: to use the **maximum likelihood estimate (MLE)**, that is, the parameter values that account best for the residuals.

Figure 10.1 shows how linear and quadratic curves are fitted to the data. Notably, the values of y increase in both directions in the quadratic model, but not in the linear model. For higher order polynomials, the effect would even be greater; the curve would be highly oscillatory and there would be many additional extremal and inflection points for which we would not have a scientific explanation.

In general, complex models can achieve a satisfactory fit for more data sets than simple models can. These superior fitting resources can also be a vice as the problem of **overfitting** illustrates: the more degrees of freedom a model has, the more difficult it is to simultaneously estimate all model parameters. We will often fit noise to the data. This is sometimes phrased in the words that complex models have high **estimation**

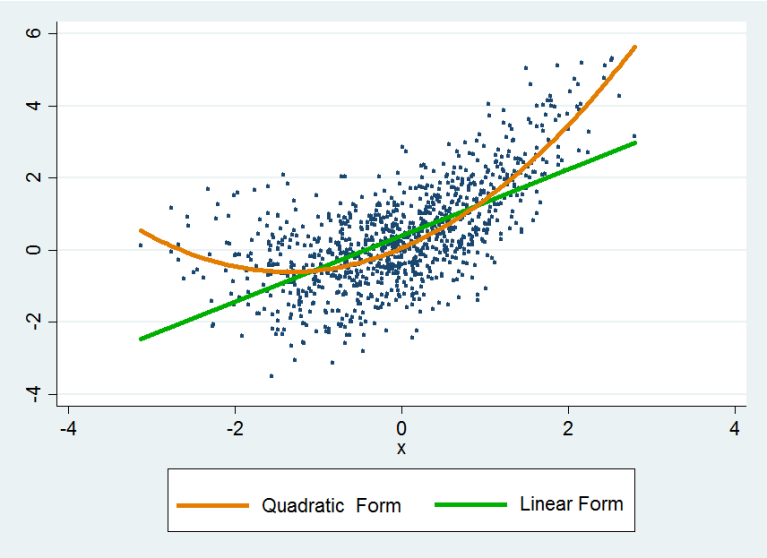

Figure 10.1: A linear model (LIN, green line) and a quadratic model (PAR, orange line) are fitted to a scatterplot of data according to the ordinary least squares method.

variance. Simultaneously estimating numerous parameters is more difficult and error-prone than only estimating one or two of them. Notably, complex models can perform worse than simpler models even if they are closer to the data-generating process.

This problem is aggravated by the frequent use of MLE's: maximum likelihood estimation is usually overoptimistic with respect to the predictive performance of the chosen model, especially when the number of adjustable parameters is high. An MLE always selects the best-fitting model in a model class. Projecting the current goodness-of-fit to future predictive accuracy just neglects the problem of high estimation variance and the danger of overfitting. While there is a strong epistemic intuition that we should try to find the correct model, even it is very complex, a predictive perspective may advocate to prefer a wrong, but stable simple model: the joint estimation of the coefficients in the complex model will almost always lead to misleading parameter values, and thus to bad predictions.

Therefore it appears natural to assign an epistemic value to simplicity in curve fitting and model selection. Null hypothesis significance testing takes this into account. When a null hypothesis $H_0 : \theta = \theta_0$ is tested against the alternative $H_1 : \theta \neq \theta_0$, the alternative is more complex than the null since it has one additional degree of freedom: the parameter θ

takes no definite value under H_1. This allows H_1 to fit the data better than H_0. On the other hand, if the population mean is not exactly equal to θ_0, but quite close to it, the null hypothesis still does a good job. In general, the null hypothesis makes more precise (though not necessarily more accurate) predictions than the alternative, and it is easier to use in theoretical developments. Therefore, the statistical thresholds for rejecting the null in favor of the alternative are typically high (e.g., the famous $p < 0.05$). By imposing high standards before calling observed data significant evidence against H_0, we compensate for the fact that complex hypotheses have it easier to achieve a good fit, even if they are mistaken.

There is a striking resemblance, by the way, between this viewpoint and Popper's idea that good scientific theories should trade off simplicity—which is associated with being informative—and predictive accuracy: "Science does not aim, primarily, at high probabilities. It aims at a *high informative content*, well backed by experience." (Popper, 2002, 416, original emphasis) Such a viewpoint can, in turn, be related to truthlikeness or verisimilitude as a primary goal of science, and it is possible to find a fruitful role for simplicity in that paradigm (e.g., Oddie, 1986; Niiniluoto, 1999).

One note of caution, though. The understanding of simplicity as number of free parameters works well for polynomial models and similar cases, but not across the board. Some models are deceptively simple. For example, with only two free parameters (α and β) we can construct a functional dependency $f(x) = \alpha \sin(\beta x)$ such that *all* data points in a dataset D are fitted up to an arbritrary amount of precision, notwithstanding the size of D. But certainly, this model is has many features that are hard to make sense of scientifically, such as the extreme oscillation of f as a function of x. Comparing different hypotheses in terms of simplicity is thus relative to a particular model family (e.g., polynomials) in which they are embedded.

The problem of estimation variance establishes that simplicity can be a cognitive value in statistical inference: it helps us to make more reliable estimates. But can we also determine an optimal tradeoff rate between simplicity and goodness-of-fit? This thesis has been defended with respect to Akaike's model comparison criterion AIC—a criterion that is widely applied in scientific reasoning and has attracted particular interest in ecological modeling (Burnham and Anderson, 2002, 2004).

The Akaike Information Criterion (AIC)

This section takes a closer look at the Akaike Information Criterion (AIC) proposed by Akaike (1973) and Sakamoto et al. (1986) as a representative of non-Bayesian ways to make simplicity matter for predictive accuracy. Partly, we chose AIC because of its historical role in model selection and partly, because it has been used by philosophers of science as a showcase for making general epistemic claims about the role of simplicity in scientific inference (Forster and Sober, 1994; Forster, 1999, 2000; Forster and Sober, 2010; Sober, 2002, 2008).

The AIC tries to estimate the discrepancy between the candidate model and the unknown true model. A popular metric for this discrepancy is the *Kullback-Leiber divergence*

$$\begin{aligned} D_{KL}(f||g_\theta) &:= \int f(x)\, \log \frac{f(x)}{g_\theta(x)}\, dx \\ &= \int f(x)\, \log f(x)\, dx \,-\, \int f(x)\, \log g_\theta(x)\, dx \end{aligned} \tag{10.3}$$

known from Variation 1. Here, f denotes the probability density of the unknown true model f, g_θ is a class of candidate models indexed by parameter θ, and the integral is taken over the sample space (=the set of observable results). As stated before, Kullback-Leiber divergence is used in information theory to measure the loss of content when estimating the unknown distribution f by an approximating distribution g_θ.

Of course, we cannot compute KL-divergence directly for a given candidate model g_θ. First, we do not know the true probability density f. This implies that we can only *estimate* KL-divergence. Second, g_θ is no single model, but stands for an entire class of models with parameter θ. We have to use a particular element of g_θ for the estimation procedure. The maximum likelihood estimator $g_{\hat{\theta}}$ is a particularly natural candidate: it is the model whose parameter values maximize the likelihood of the data, given the model. However, if one used the maximum likelihood estimator to estimate KL-divergence without any corrective terms, one would overestimate the closeness to the true model. Third, we are not interested in KL-divergence per se, but in predictive success. So we should relate (10.3) in some way to the predictive performance of a model. Akaike's (1973) famous mathematical result addresses these worries:

Akaike's Theorem: *For observed data y and a candidate model*

> *class g_θ with K adjustable parameters (or an adjustable parameter of dimension K), the model comparison criterion*
>
> $$AIC(g_\theta, N) := -\log g_{\hat{\theta}(y)}(y) + K \tag{10.4}$$
>
> *is an asymptotically unbiased estimator of $\mathbb{E}_x\mathbb{E}_y[\log(f(x)/g_{\hat{\theta}(y)}(x))]$—the "expected predictive success of $g_{\hat{\theta}}$".*

In the above equation, $g_{\hat{\theta}(y)}$ denotes the probability density of the maximum likelihood estimate $\hat{\theta}(y)$. To better understand the double expectation in the last term, note that the maximum likelihood estimate $g_{\hat{\theta}}$ is determined with the help of the data set y. Then, $g_{\hat{\theta}}$'s KL-divergence to the true model f is evaluated with respect to another set of data x. This justifies the name predictive success, and taking the expectation two times—over training data y and test data x—justifies the name **expected predictive success**.

In other words, AIC estimates expected predictive success by subtracting the number of parameters K from the log-likelihood of the data under the maximum likelihood estimate $g_{\hat{\theta}}$. It gives an *asymptotically unbiased estimate* of predictive success—an estimate that will, in the long run, center around the true value. The more parameters a model has, the more do we have to correct the MLE estimate in order to obtain an unbiased estimate. We are then to favor the model which minimizes AIC among all candidate models. According to Forster and Sober,

> Akaike's theorem shows the relevance of goodness-of-fit *and* simplicity to our estimate of what is true. But of equal importance, it states a precise rate-of-exchange between these two conflicting considerations: it shows how the one quantity should be traded off against the other. (Forster and Sober, 1994, 11)

Moreover, they use Akaike's theorem to counter the (empiricist) idea that simplicity is a merely pragmatic virtue and that "hypothesis evaluation should be driven by data, not by *a priori* assumptions about what a 'good' hypothesis should look like [such as being simple, the author]" (Forster and Sober, 1994, 27). By means of Akaike's theorem, simplicity is assigned a specific weight in model selection and established as a cognitive value.

However, this argument does not stand on firm grounds. First, unbiasedness is not *sufficient* to ensure the goodness of an estimator. The goodness of an estimator $\hat{\theta}$ relative to the true value θ is usually measured by the mean square error, which can be written as the square of the bias plus its variance.

$$MSE[\hat{\theta}] = (\mathbb{E}[\hat{\theta} - \theta])^2 + \mathbb{E}[(\hat{\theta} - \theta)^2]$$

If $\hat{\theta}$ is unbiased, the first term will disappear, but this does not ensure low overall error—an unbiased estimator may have high variance, dissipate far from the true value and be awfully bad in practice. In particular, it may be outperformed by an estimator with low variance that is only slightly biased. By itself, unbiasedness does not warrant good performance.

This objection may be countered by noting that unbiasedness is an advantage, ceteris paribus. Forster and Sober (2010) note that AIC and one of its rivals, the Bayesian Information Criterion BIC, just differ by a constant. If estimators differ by a constant, they have the same variance. Since mean square error = square of bias + variance, the unbiased estimator will have the lower mean square error. Hence BIC seems to be a worse estimator of predictive accuracy than the unbiased AIC.

However, this argument is based on an oversight which many authors in the debate commit (Forster and Sober, 1994, 2010; Kieseppä, 1997). AIC is *not* an unbiased estimator—it is just *asymptotically* unbiased, in other words, the property of unbiasedness is only realized for very large samples. To the excuse of these authors, it should be added that the standard AIC textbook by Sakamoto et al. (1986, 69) sometimes uses this formulation in passing. Several other passages (Sakamoto et al., 1986, 65,77,81) make clear, however, that the word "unbiased", when applied to AIC, is merely used as a shortcut for "asymptotically unbiased".

To see this with your own eyes, we invite you to have a look at the mathematical details in Section 10.6. There, the dependence of Akaike's Theorem on the asymptotical, and not the actual normality of the maximum likelihood estimator becomes clear. This has substantial consequences, and speaks against a normative interpretation of Akaike's findings. AIC outperforms BIC as an estimator of predictive accuracy only for an infinitely large sample, whereas actual applications usually deal with medium-size finite samples. As long as we don't know the speed of convergence—and this varies from data set to data set—, the asymptotic

properties are unwarranted. This is good news for those who want to defend Bayesian model selection against its non-Bayesian competitors such as AIC.

0Finally, the contribution of simplicity relative to goodness-of-fit in Akaike's Theorem diminishes as sample size N increases, as Forster (2002) notes himself (see Section 10.6 for details). The goodness-of-fit term is of the order of N whereas the simplicity-based contribution remains constant. Thus, with increasing sample size, simplicity drops out of the picture. In the light of these observations, claims that AIC establishes a tradeoff rate between simplicity and goodness-of-fit (thereby establishing simplicity as an epistemic value) shine in a dim light.

The Bayesian Information Criterion (BIC)

Are there model selection criteria that are firmly anchored within, and derived from Bayesian reasoning? The classical, subjective view of Bayesian inference consists in reasoning from a prior to a posterior probability. A model selection procedure is called Bayesian if it is based on the posterior distribution of degrees of belief, or on the difference between prior and posterior probabilities. This was also the rational behind the manifold measures of confirmation presented in Variation 2.

An example for such a procedure is model selection based on Bayes factors. They compare the performance of the rivalling models H_0 and H_1 by means of the ratio

$$B_{01}(\mathrm{E}) := \frac{p(\mathrm{H}_0|\mathrm{E})}{p(\mathrm{H}_1|\mathrm{E})} \cdot \frac{p(\mathrm{H}_1)}{p(\mathrm{H}_0)} = \frac{p(\mathrm{E}|\mathrm{H}_0)}{p(\mathrm{E}|\mathrm{H}_1)}$$

Citing work by Spiegelhalter and Smith (1980), Kass and Raftery (1995, 790) argue that Bayes factors act as a "fully automatic Occam's razor" for nested models: when the Bayes factor favors a simple model, the complex model will be penalized for hosting lots of poor-fitting hypotheses. In that case, the loss in predictive accuracy by accepting the simpler model will be negligible. The search for an *explicit* tradeoff rate between simplicity and goodness-of-fit is replaced by the rate that is implicit in Bayesian inference and Bayesian measures of evidence.

However, this orthodox Bayesian model selection is not that frequently put into practice. First of all, there is a plethora of practical and methodological problems, such as are the computational costs of calculating pos-

terior distributions or handling nested models in a Bayesian framework. Second, when prior probabilities are assigned, reliable expert opinion is usually hard to elicit so that the choice of the prior is often dominated by mathematical convenience. Furthermore, results may be highly sensitive to the prior distribution. This has triggered the search for model selection criteria that can play a useful role for approximating Bayes factors. Schwarz's Bayesian Information Criterion (BIC) has often been claimed to fulfil that role; so we will investigate its foundations in some detail. We claim that the findings in our analysis of BIC are somewhat typical of Bayesian model selection in general. For a philosophical analysis of further Bayesian model selection techniques, such as the Minimum Message Length Criterion (MML), see Dowe et al. (2007) and Sprenger (2013c).

The BIC is an estimation procedure that aims at the posterior probability of a parametric model M_θ, that is, at the weighted sum of the posterior probabilities of the hypotheses in M_θ that correspond to different values of θ. Thus, it has a different target than AIC, which compares the best-performing representatives of a class of models. We will now reconstruct and analyze the motivation of BIC, following Schwarz (1978).

Assume that M_θ is one of our candidate models, whose elements are indexed by a vector θ with dimension K. We would like to approximate the posterior probability of M_θ. Assume further that all probability densities for data x (with respect to the Lebesgue measure θ) belong to the exponential family and that they can be written as

$$p(x|\theta) = e^{N(A(x)-\lambda|\theta-\hat{\theta}(x)|^2)}. \tag{10.5}$$

Here, $\hat{\theta}(x)$ denotes the maximum likelihood estimate of the unknown θ, and N the sample size, assuming independent sampling. This specific form of the likelihood function seems to make a substantial presumption, but in fact, the densities in (10.5) comprise the most familiar distributions, such as the Normal, Uniform, Fisher, Poisson and Student's $t-$distribution. For that reason, the assumption is plausible from a practical point of view.

Then we take a standard Bayesian approach and write the posterior probability of M_θ as proportional to the prior probability $p(M_\theta)$ and the averaged likelihood of the data x under M_θ:

$$p(M_\theta|x) \quad \sim \quad p(M_\theta) \int_{\theta\in\Theta} e^{N(A(x)-\lambda|\theta-\hat{\theta}(x)|^2)} d\theta(\theta)$$

$$= \quad p(M_\theta)\, e^{NA(x)} \int_{\theta\in\Theta} e^{-N\lambda|\theta-\hat{\theta}(x)|^2} d\theta(\theta).$$

Substituting the integration variable θ by $\theta/\sqrt{N\lambda}$, and realizing that for the maximum likelihood estimate $\hat{\theta}(x)$, $p(x|\hat{\theta}(x)) = e^{NA(x)}$, we obtain

$$\begin{aligned}
\log p(M_\theta|x) &\sim \log p(M_\theta) + NA(x) + \log\left(\frac{1}{N\lambda}\right)^{K/2} + \log\int_{\theta\in\Theta} e^{-|\theta-\hat{\theta}(x)|^2} d\theta(\theta) \\
&= \log p(M_\theta) + NA(x) + \frac{1}{2}K\log\left(\frac{1}{N\lambda}\right) + \log\sqrt{\pi}^K \\
&= \log p(M_\theta) + \log p(x|\hat{\theta}(x)) - \frac{1}{2}K\log\left(\frac{N\lambda}{\pi}\right). \qquad (10.6)
\end{aligned}$$

Let us take stock. On the left hand side, we have the log-posterior probability, which can be interpreted as a subjective Bayesian's natural model selection criterion. As we see from (10.6), this term is proportional to the sum of three terms: log-prior probability, the log-likelihood of the data under the maximum likelihood estimate, and a penalty proportional to the number of model parameters. This derivation, whose assumptions are relaxed subsequently in order to yield more general results, forms the mathematical core of BIC. The number of parameters K enters the calculations because the expected likelihood of the data depends on the dimension of the model, via the skewness of the likelihood function.

In practice, it is difficult to elicit sensible subjective prior probabilities of the candidate models, and the computation of posterior probabilities involves high computational efforts. Therefore, Schwarz suggests to estimate log-posterior probability in (10.6) by a large sample approximation. We neglect the terms that make only constant contributions and focus on the terms that increase in N: $\log p(M_\theta)$ drops out of the picture. Therefore, in the long run, the model with the highest posterior probability will be the model that minimizes

$$BIC(M_\theta, x) = -2\log p(x|\hat{\theta}(x)) \; + \; K\,\log N. \qquad (10.7)$$

BIC is intended to select the model that accumulates, in the long run, the most posterior mass. However, it neglects the contribution of the priors when comparing the models to each other. Keeping in mind the identity

$$\log p(\mathrm{H}|\mathrm{E}) = \log p(\mathrm{H}) + \log\left(p(\mathrm{E}|\mathrm{H}) \cdot \frac{1}{p(\mathrm{E})}\right) \qquad (10.8)$$

and comparing it to Equation 10.6, wee see that BIC could as well be described as an approximation to the log-ratio measure of confirmation

$\log p(\mathrm{H}|\mathrm{E}) - \log p(\mathrm{H}) = \log(p(\mathrm{H}|\mathrm{E})/p(\mathrm{H}))$, up to addition of a constant ($\rightarrow$ Variation 2).

Therefore, BIC can only partially be described as having a properly Bayesian justification: while (log-ratio) confirmation may be suitable for comparing models on the basis of past performance, it does not conform to classical subjective Bayesian inference: the priors drop out of the picture, as witnessed by the transition from (10.6) to (10.7). Instead of conforming to the subjective Bayesian rationale, the BIC is a **hybrid procedure**: it does not primarily aim at an accurate representation of subjective uncertainty. Rather, it uses the Bayesian calculus as a convenient mathematical tool for meeting goals that a statistician or modeler may encounter in inference, such as selecting models with strong performance on past data. There is nothing specifically Bayesian about the estimation target of BIC. This finding is, by the way, in agreement with Schwarz' note that BIC extends "beyond the Bayesian context" (Schwarz, 1978, 461). See also Forster and Sober (1994, 23–24). Even more, frequentist properties are sometimes invoked in an attempt to justify the practical use of BIC (e.g., Burnham and Anderson, 2002).

To further strengthen this conclusion, note that BIC is quite different from a numerical large sample approximation for posterior degrees of belief: the posterior approximated by BIC is detached from subjective prior probability. So BIC is not just a practical approximation to Bayesian coherence. Compare BIC to techniques such as Gibbs sampling or Monte Carlo Markov Chains (Han and Carlin, 2001): those techniques aim at numerical approximations of subjective posterior distributions, and offer computational help for tricky multi-dimensional integrals. BIC follows a philosophical rationale that is much less tied to Bayesian reasoning.

Neither does the statistical consistency of BIC provide a genuinely Bayesian justification. Here, consistency does not denote logical consistency with another proposition, but a long-run property of statistical estimators. An estimator is consistent if and only if it converges in probability to the true model as sample size increases. Both Bayesians and frequentists regard consistency as a necessary constraint on good estimators, and BIC is consistent as long as the overall model is not misspecified (i.e., if it contains the true model). Apart from the fact that this condition may often fail in practice, consistency alone has no implications on speed of convergence to the true value. Hence, it is not a sufficient reason for us-

ing a particular method. So neither is consistency in any way peculiar to Bayesian inference, nor is it strong enough to make a case for BIC as opposed to other methods.

Our diagnosis that BIC lacks, in spite of the extensive use of Bayesian formalism, a fully Bayesian rationale, is supported by the variety of purposes to which BIC is put. Sometimes it is regarded as an approximation to the Bayes factor (Kass and Raftery, 1995). Raftery (1995) proposes an interpretation of BIC as an approximation to marginal likelihood, which is easily derived on the basis of the above calculations. Romeijn et al. (2012) see different worries with a Bayesian understanding of BIC and propose to anchor it more securely in Bayesian reasoning by taking into account the size of the parameter space. Hence, what the asymptotic analysis of BIC approximates is not determined by the mathematics only: it depends on the general perspective one adopts.

Neither is the derivation of BIC committed to the true model being in the set of candidate models which is a standard premise of Bayesian convergence-to-truth theorems (Blackwell and Dubins, 1962; Gaifman and Snir, 1982). All this shows that for BIC, Bayesianism constitutes no philosophical underpinning (e.g., as a logic of belief revision), but only a convenient framework which motivates using a specific estimator of log-posterior probability.

Discussion

Simplicity is a complex and ambiguous concept. This ambiguity may explain why people are so divided over whether or not it does have cognitive value in science. From a Bayesian point of view, it is most fruitful to investigate the concept of simplicity as elegance, that is, as referring to the number and the complexity of hypotheses in a scientific theory. The ontological dimension of simplicity ("parsimony") is left out of the picture.

The cognitive value of simplicity as elegance has been investigated in the framework of model selection and curve-fitting, that is, in fitting a scientific hypothesis (=a particular curve) to a set of data points. In this context, simplicity has epistemic value as an antidote to estimation error, that emerges from the multiple estimation of various parameters. However, this qualitative finding does not answer the question of whether there is an optimal tradeoff rate between simplicity and other cognitive values

in model selection, most notably goodness-of-fit.

In answering this quantitative question, we have reviewed two model selection criteria: AIC and BIC. First, we have taken issue with Forster and Sober's claim that an optimal tradeoff rate between simplicity and goodness-of-fit is established by the non-Bayesian Akaike's Information Criterion AIC. Then we moved on to the Bayesian Information Criterion BIC. However, while the derivation BIC is couched into Bayesian language, properly Bayesian elements, such as the prior probability of different models, are swept under the carpet. This sheds doubt on whether BIC is, as its name promises, a fully Bayesian model selection criterion.

The finding from these case studies is that most model selection criteria do not emerge from rigorous derivations, backed by a particular philosophical approach (e.g., frequentism, Bayesianism or likelihoodism). These approaches provide the conceptual and mathematical framework for deriving a model selection criterion (such as AIC or BIC), but no waterproof justifications. Even more, it is also possible to find a non-Bayesian justification for BIC, and a Bayesian justification for AIC (Romeijn et al., 2012). So it is misleading to attach model selection criteria to particular philosophical schools. The judgment on when these criteria do and do not work is highly context-sensitive. The Bayesianism involved in the derivation of BIC might be characterized as an **instrumental Bayesianism**—an approach to statistical inference which is happy to Bayes's Theorem as a scientific modeling tool, but without taking the Bayesian elements too literally, as expressions of subjective uncertainty.

On the positive side, our findings imply that we can reject some of the bolder anti-Bayesian conclusions drawn from the investigation of model selection criteria. For instance, Forster and Sober (1994) write:

> Bayesianism is unable to capture the proper significance of considering *families* of curves [...]. Akaike's reconceptualization of statistics does recommend that the foundations of Bayesian statistics require rethinking. (Forster and Sober, 1994, 26, original emphasis)

In the light of the above results, we can safely conclude that this conclusion is mistaken. First, Bayesian and non-Bayesian model selection criteria stand on equal footing. Second, the link between particular model selection criteria and philosophical schools is not particularly tight. Third,

Bayesianism considers, in calculating Bayes factors and approximations to them (such as the BIC), the significance of families of curves (models) as opposed to single curves (fitted models). It may also be noted that Forster and Sober do not repeat this claim in later publications, probably realizing that their conclusion went too far.

I.J. Good (1971) once quipped that there are at least 46656 varieties of Bayesians. After studying Bayesian model selection, we may add that there are also many different ways of doing Bayesian model selection—as long as one conceives Bayesian reasoning in a more general way than just an inference from prior to posterior probability. In model selection, Bayesian reasoning is often applied in an instrumental manner. While these Bayesian methods are adequate frameworks for investigating our core question about why simplicity matters in model selection, they fail, like their non-Bayesian counterparts, to state an optimal tradeoff rate for simplicity and goodness-of-fit. As our examples and analyses have shown, that latter question may just depend too much on the specific context to allow for an illuminating general answer.

This observation also leads to several avenues for further research. First, if foundational philosophical arguments fail to ground model selection criteria, what are the context-sensitive considerations that lead us to prefer one of the model selection criteria over others? What are the reasons why we decide to work with AIC instead of BIC, or vice versa? Second, AIC and BIC are only two of a large number of model selection criteria, and it would be good to extend the analysis in order to see whether the conclusions of this variation remain valid. One of us (Sprenger, 2013c) has conducted such an analysis with respect to the Minimum Message Length (MML, Dowe et al., 2007) and the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002), but a more detailed study, that also involves further model selection criteria, would be highly welcome. Third, it would be exciting to see whether our thesis of instrumental Bayesianism (and instrumental frequentism, for the case of AIC) also extends to other cases of scientific inference. That is, can we find cases of scientific reasoning where schools of uncertain reasoning are treated as a quarry for mathematical formalisms rather than as a philosophical basis for justifying one's inference? Fourth, one may compare the role of simplicity in model selection to its role in other forms of scientific reasoning, e.g., causal inference, and check whether the results remain the same.

The last two variations have demonstrated the fruitfulness of Bayesian reasoning in statistical inference. The final variation will respond to a foundational objection that is frequently raised against the (subjective) Bayesian: that Bayesian inference is not objective enough to be of any value in science.

Sketch of the Derivation of Akaike's Information Criterion

At the end of this variation, we summarize the main steps of the derivation of AIC below, with a focus on the philosophical rationale that motivates this model selection criterion. Detailed treatments can be found in Chapter 4.3 of Sakamoto et al. (1986) and Chapter 7.2 in Burnham and Anderson (2002).

The AIC aims at estimating the "expected predictive success" of a model, identified with its maximum likelihood estimate (MLE) $g_{\hat{\theta}(y)}$:

$$\mathbb{E}_x\mathbb{E}_y\left[\log\frac{f(x)}{g_{\hat{\theta}(y)}(x)}\right] = \mathbb{E}_x\mathbb{E}_y\left[\log f(x)\right] - \mathbb{E}_x\mathbb{E}_y\left[\log g_{\hat{\theta}(y)}(x)\right] \quad (10.9)$$

The first term on the right hand side of (10.9) is equal for all candidate models. When comparing them, it drops out as a constant. Hence we can neglect it in the remainder and focus on the second term in (10.9).

The AIC is usually derived by a double Taylor expansion of the log-likelihood function. The general formula of Taylor expansion for an analytic, real-valued functions f is

$$f(x) = \sum_{k=0}^{\infty} f^{(k)}(x_0)(x - x_0)^k.$$

In our case, we expand the term $\log g_{\hat{\theta}(x)}(y)$—our MLE—around θ_0, the value of θ that minimizes Kullback-Leibler divergence to the true model. The expansion is trunctated at $k = 2$ yielding

$$\begin{aligned}\log g_{\hat{\theta}(y)}(x) \approx\ & \log g_{\theta_0}(x) + N\left(\left(\frac{\partial}{\partial\theta}\log g_\theta(x)\right)(\theta_0)\right)(\hat{\theta}(y) - \theta_0) \\ & + \frac{1}{2}N(\hat{\theta}(y) - \theta_0)^T\left(\left(\frac{\partial^2}{\partial\theta^2}\log g_\theta(x)\right)(\theta_0)\right) \\ & (\hat{\theta}(y) - \theta_0) \end{aligned} \quad (10.10)$$

The matrix

$$J := -\frac{\partial^2}{\partial\theta^2} \log g_\theta(x)(\theta_0)$$

that also occurs in (10.10) is called the *Fisher information matrix* of the data. It plays a crucial role in an asymptotic approximation of the maximum likelihood estimator that holds under plausible regularity conditions:

$$\sqrt{N}(\hat{\theta}(y) - \theta_0) \to \mathcal{N}(0, J^{-1}).$$

This asymptotic normality of the maximum likelihood estimator can be used to simplify (10.10). The term

$$\sqrt{N}(\hat{\theta}(y) - \theta_0)^T (-J) \sqrt{N}(\hat{\theta}(y) - \theta_0) \tag{10.11}$$

is asymptotically χ^2-distributed with K degrees of freedom. Hence, the expectation of (10.11) is K. By taking a double expectation over x and y, we thus obtain that

$$\mathbb{E}_x\mathbb{E}_y \left[\frac{1}{2} N(\hat{\theta}(y) - \theta_0)^T \left(\left(\frac{\partial^2}{\partial\theta^2} \log g_\theta(x)\right)(\theta_0)\right)(\hat{\theta}(y) - \theta_0)\right] \approx \frac{K}{2} \tag{10.12}$$

Moreover, the linear term in (10.10) vanishes because the maximum likelihood estimate is an extremal point of the log-likelihood function. Thus, the mean of the first derivative is also zero:

$$\mathbb{E}_x\mathbb{E}_y \left[N \left(\left(\frac{\partial}{\partial\theta} \log g_\theta(x)\right)(\theta_0)\right)(\hat{\theta}(y) - \theta_0)\right] = 0 \tag{10.13}$$

Combining (10.10) with (10.12) and (10.13), we obtain for large samples that

$$\mathbb{E}_x\mathbb{E}_y \left[\log g_{\hat{\theta}(y)}(x)\right] \approx \mathbb{E}_x \left[\log g_{\theta_0}(x)\right] - \frac{K}{2}. \tag{10.14}$$

Repeating the Taylor expansion around the maximum-likelihood estimate and applying the same arguments once more gives us

$$\mathbb{E}_x\mathbb{E}_y \left[\log g_{\theta_0}(x)\right] \approx \mathbb{E}_y \left[\log g_{\hat{\theta}(y)}(y)\right] - \frac{K}{2}. \tag{10.15}$$

Finally, by combining (10.14) and (10.15) we obtain AIC as an estimate of "expected predictive accuracy":

$$\mathbb{E}_x\mathbb{E}_y \left[\log g_{\hat{\theta}(y)}(x)\right] \approx \mathbb{E}_y \left[\log g_{\hat{\theta}(y)}(y)\right] - K.$$

Variation 11: Scientific Objectivity

Scientific objectivity pertains in the first place to scientific method. It expresses the idea that the claims, methods and results of science are not, or should not be influenced by particular perspectives, value commitments, community bias or personal interests, to name a few relevant factors (for a survey, see Reiss and Sprenger, 2014). Objectivity contributes to the reliability of scientific research, conveys an image of epistemic authority and strengthens our trust in science. The 2009 "Climategate" affair, when climate scientists were charged with presenting data in a misleading way, and the widespread failure to replicate experimental results in psychology (Galak et al., 2012; Open Science Collaboration, 2015) illustrate how an apparent lack of objectivity weakens trust in scientific findings.

The ideal of objectivity has been criticized repeatedly in philosophy of science, questioning both its value and its attainability (e.g., Feyerabend, 1975; Kuhn, 1977b). This variation does not aim at defending it. Rather, we start from the assumption that *some* degree of objectivity is beneficial in scientific reasoning. Then we discuss what kind of objectivity can be delivered by, and is compatible with, Bayesian inference. This discussion will be focused on Bayesian statistics, which we see as the primary application of the subjective probability calculus in science. We investigate three challenges to the objectivity of Bayesian inference, and for each of them, we develop proposals how they can be rebutted, and how Bayesian reasoning in science can be defended.

This variation is divided into five sections. Section 11.1 presents some background on the role of objectivity in science. Section 11.2, 11.3 and 11.4 describe three major challenges to Bayesian inference. In our replies, we blend arguments from statistical methodology with recent results in the philosophical analysis of scientific objectivity (e.g., Douglas, 2004, 2011). The final Section 11.5 places our arguments into a broader perspective and concludes.

Forms of Scientific Objectivity

There are two principled ways of understanding scientific objectivity: we can relate it to the products of science (e.g., theories, models, laws) and their relation to the world, and we can regard it as a property of the process of scientific reasoning. In line with the approach of this book, we focus on **process objectivity** instead of **product objectivity** (Reiss and Sprenger, 2014): we are interested in the objectivity of the procedures that lead to the acceptance of scientific theories, not in whether the theories themselves provide a faithful image of reality. First of all, that latter question stands orthogonal to our discussion of Bayesian inference in science. Second, the traditional idea of objectivity as correspondence between theory and world has been thoroughly debunked: The historical case studies of Porter (1996) and Daston and Galison (2007), the systematic work on the theory-observation interaction by Thomas Kuhn (1962, 1977b) and Paul Feyerabend (1962, 1975), and finally, critiques from feminist philosophy of science and standpoint epistemology (e.g., Longino, 1990; Harding, 1991; Lloyd, 2005; Okruhlik, 2005) have led to the conclusion that the view of scientific objectivity as faithfulness to facts is problematic. If objectivity is supposed to be a meaningful cognitive value, it needs to be associated with peculiar ways of **scientific reasoning**.

Most analyses of objectivity in scientific reasoning proceed in one of the two following ways: (1) scientific reasoning is objective to the extent that it is free of non-cognitive (moral, social and political) values; (2) scientific reasoning is objective to the extent that personal biases are absent, or that they can be eliminated in a social process. These two forms of objectivity are related, e.g., personal bias is often expressed by endorsement of a particular non-cognitive value, such as a political ideology. To simplify the analysis, we will not draw a sharp distinction between them. Nevertheless, we would like to stress that these forms of objectivity cannot be reduced to each other: not all individual biases correspond to a particular non-cognitive value; not all non-cognitive values correspond to an individual bias.

Moreover, it is helpful to distinguish four stages at which non-cognitive values may affect science. They are: (i) the choice of a scientific research problem; (ii) the gathering of evidence in relation to the problem; (iii) the assessment and acceptance of a scientific hypothesis or theory as an

adequate answer to the problem on the basis of the evidence; (iv) the proliferation and application of scientific research results (Weber, 1904, 1917).

Most philosophers of science would agree that the role of values and bias in science is contentious only with respect to dimensions (ii) and (iii): the gathering of evidence and the **assessment of scientific theories**. It is almost universally accepted that the choice of a research problem is often influenced by the interests of individual scientists, funding parties, and society as a whole. This influence may make science more shallow and slow down its long-run progress, but it has benefits, too: scientists will focus on providing solutions to those intellectual problems that are considered urgent by society and they may actually improve people's lifes. Similarly, the proliferation and application of scientific research results is evidently affected by the personal values of journal editors and end users. The real debate is about whether or not the 'core' of scientific reasoning—the gathering of evidence and the assessment and acceptance scientific theories—is, and should be free of values and bias. And since Bayesian inference provides a theory of scientific inference rather than a theory of designing and conducting experiments, we will focus on stage (iii)—scientific theory choice— in particular.

We can now define objectivity in scientific inference according to the two conceptions mentioned above:

> **Value-Free Ideal (VFI):** Scientists should strive to minimize the influence of non-cognitive values on scientific reasoning in gathering evidence and assessing and accepting scientific theories.

It is mainly this ideal against which Bayesian inference and its main competitor in scientific reasoning, **frequentist inference**, will be assessed. Other relevant forms of objectivity, such as intersubjective agreement on which conclusions to draw from a body of data, are often related to the VFI. For instance, differences in the inferences that scientists draw may be tracked to the impact of non-cognitive values on the reasoning of the individual scientists.

Apparently, there is a **straightforward clash between the VFI and subjective Bayesian inference**: "a notion of probability as personalistic degree of belief [...], by its very nature, is not focused on the extraction and

presentation of evidence of a public and objective kind" (Cox and Mayo, 2010, 298). This view is echoed in writings of well-known statisticians and philosophers of science such as Fisher (1956), Mayo (1996), Popper (2002) and Senn (2011). By and large, the objectivity-related criticisms of subjective Bayesian inference come in three different forms. First, the lack of constraints on prior probabilities, second, the entanglement of statistical evidence and degree of belief, third, the apparent blindness to bias in experimental design. In the light of these challenges, one is tempted to conclude that Bayesian inference cannot produce objective knowledge, is not suitable for scientific communication and is therefore inferior to frequentist inference. Let us now review these challenges in detail.

Challenge 1: The Choice of the Prior Distribution

As explained in the introduction of this book, the core of Bayesian inference consists in representing degrees of belief by probabilities, in changing them by means of Bayesian Conditionalization, and in basing decisions on posterior probabilities. In particular, the posterior degree of belief in a hypothesis H upon learning evidence E can be written as follows:

$$p(\mathrm{H}|\mathrm{E}) = \frac{p(\mathrm{H})p(\mathrm{E}|\mathrm{H})}{p(\mathrm{E})} \tag{11.1}$$

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

Posterior probability depends on prior probability, and often, there is not sufficient background knowledge to establish consensus on the latter. Subjective Bayesians such as de Finetti (1972) have stressed that in principle, **any coherent prior probability distribution can be defended as rational**. This attitude seems to jeopardize any claims to objectivity that subjective Bayesians could possibly make. What kind of epistemic warrant does a Bayesian inference still provide? After all, the choice of the prior can hide all kind of pernicious values, e.g., financial interests of the experiment sponsor. This is particularly worrying in sensitive areas such as medicine, where the need for impartial inference methods is particularly

high, due to the manifest financial interests in clinical trials and the ethical consequences of wrong decisions. As the medical methodologist Lemuel Moyé writes:

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

In other words, Bayesians can bias the final result in their preferred direction by choosing an appropriate prior. The first challenge is thus based on the value-free ideal that the core business of scientific reasoning, namely evaluating evidence, assessing and accepting theories, should be free of non-cognitive values and individual biases—a requirement that Bayesian inference seems to violate blatantly. Adherence to the value-free ideal has, however, in one form or another, been upheld as a trademark of scientific objectivity (e.g., Lacey, 1999; Reiss and Sprenger, 2014), and for practitioners, it plays an even greater role due to regulatory constraints and conflicts of interests. Even if one doubts that the value-free ideal can be attained in practice, values should not be allowed to *replace* scientific evidence (Douglas, 2009b). How can Bayesian inference be safeguarded against this danger?

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

The second line of defense notes that prior probabilities are open to rational criticism. Whenever a prior distribution is used, be its shape conventional or peculiar, the researcher should justify her particular choice

and explain which considerations (theoretical and empirical ones) led her to this choice. We cannot justify an extreme posterior simply by choosing a suitably extreme prior because it is part of the Bayesian model of reasoning that also the prior needs to be justified. This is also explicit in regulations for medical trials, such as the guidelines for the use of Bayesian statistics, issued by the Food and Drug Administration of the United States:

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

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

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

The bottom line is that the choice of the prior is just like any other modeling assumption in science open to rational criticism. Indeed, if the prior were not varied and judged critically, there would be no corrective mechanism for gauging to what extent personal bias has influenced the results through the choice of the prior. But the same is true of scientific inference in general, and of competing statistical paradigms in particular. Invalidating a subjective Bayesian analysis with a biased prior is as easy or difficult as invalidating a non-Bayesian analysis with biased modeling

assumptions. Therefore, this challenge is not more fearsome for Bayesians than for any other framework of inductive inference. We now move on to the next challenge: that Bayesians mix up belief and evidence.

Challenge 2: Belief vs. Evidence

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

> Advocates of inverse probabilities [=ascribing probabilities to scientific hypotheses] are forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies [=degrees of belief], theorems respecting which are useless for scientific purposes (Fisher, 1935, 6–7, our explanations in parentheses)

Royall (1997, 4) makes a similar distinction between three major questions in statistical analysis: "What should we believe?", "What should we do?" and "What is the evidence?". A good answer to one of them need not be a good answer to another question. The Bayesian answers the belief question by providing prior and posterior probabilities, and the decision question by means of its connection to rational choice theory (e.g., Savage, 1972), but what does a satisfactory response to the evidence question look like?

Underlying this challenge is the idea of "detached objectivity" (Douglas, 2009b, 459): claims to scientific knowledge should be detached from personal belief and wishful thinking. Bayesians also struggle to achieve "concordant objectivity" (Douglas, 2004, 462–463) which is expressed in intersubjectively agreed assessments of evidence. As Quine (1992, 5) stated:

"The requirement of intersubjectivity is what makes science objective." However, the "psychological tendencies" that correspond to personal degrees of belief do not fulfil this requirement.

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

> I have great sympathy for the Bayesian position [...] The problem in implementing Bayesian statistics for scientific publications, however, is that such analyses are inherently subjective, by definition [...] with no objectively right answer as to what priors are appropriate. I do not see that as useful scientifically.

In other words, it is unclear how Bayesians can separate personal belief from evidential support and achieve intersubjective agreement on levels of evidence.

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

$$B_{01}(E) \quad := \quad \frac{p(\mathrm{H}_0|\mathrm{E})}{p(\mathrm{H}_1|\mathrm{E})} \cdot \frac{p(\mathrm{H}_1)}{p(\mathrm{H}_0)} = \frac{\int_{\theta\in\Theta_0} p(\mathrm{E}|\theta)p(\theta)d\theta}{\int_{\theta\in\Theta_1} p(\mathrm{E}|\theta)p(\theta)d\theta} \tag{11.2}$$

It is important to note that the Bayes factor is not affected by $p(\mathrm{H}_0)$ and $p(\mathrm{H}_1)$ simpliciter. For two point hypotheses H_0 and H_1, it is even fully independent of the prior probability distribution: it is just the likelihood ratio $p(\mathrm{E}|\mathrm{H}_0)/p(\mathrm{E}|\mathrm{H}_1)$, indicating how much E favors H_0 over H_1. Nothing depends on personal belief.

For composite hypotheses (e.g., $\mathrm{H}_1 : \theta \neq \theta_0$), things are more complicated. The value of the Bayes factor depends on how likely the observed evidence is under the various components of H_0 and H_1, weighted with their relative prior probability. It is important to realize that this dependency is benign and not pernicious in the context of null hypothesis testing. Imagine the frequent case that we are testing the null hypothesis that

a certain intervention, e.g., taking vitamin C tablets as a cure for the common flu, has no effect at all: $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$, where θ is the variable denoting the effect size. Of course, it is implausible that the effect of the vitamin C intervention is *exactly* zero: the pill will cause *some* biochemical reaction in the human body. The test does not aim at ruling out this possibility, but at finding out whether we can use the null hypothesis as a simple and precise, but strictly speaking wrong idealization of a complex reality. In order to assess whether a finding is evidence for or against H_0, we need to know which effect sizes are plausible and clinically relevant. Only if this is clarified, we can state meaningfully that the observed results speak in favor of or against the null hypothesis. This conventional methodological wisdom is mirrored in the calculation of Bayes factors.

In fact, **also frequentist inference with null hypothesis significance tests (NHST) needs such plausibility judgments**. In NHST, the null hypothesis $H_0 : \theta = \theta_0$ of zero effect is pitched against the alternative hypothesis $H_1 : \theta \neq \theta_0$ that there is some effect ($\rightarrow$ Variation 9). While a type I error corresponds to erroneous rejection of the null hypothesis, a type II error stands for erroneous acceptance of the null, or more precisely, failure to reject the null. Conventionally, acceptable type I error rates are set at a level of 5%, 1% or 0.1%, dependent on the experiment. By choosing an appropriate sample size N, one tries to minimize type II error. In other words, we may be rational in following the test procedure because of its favorable long-run properties:

> [...] we shall reject H_0 when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H_0 sufficiently often when it is false. (Neyman and Pearson, 1967, 142)

In other words, the relative frequency of correct decisions will clearly exceed the relative frequency of wrong decision. This reliance on favorable long-run properties also explains the name frequentism.

However, when H_1 is a composite hypothesis (e.g., $\theta \neq \theta_0$), the power of an experiment has to be calculated relative to specific effect sizes. Usually, one would choose effect sizes which correspond to theoretical expectations and which imply a scientifically meaningful difference to the null hypothesis. Without a judgment on which effect sizes are likely to be expected, the choice of the sample size N is tantamount to groping in the

dark. One may collect much more evidence than needed or, in the opposite case, end up with a severely underpowered study. Therefore, the relative plausibility of the different alternatives and the initial plausibility of the null hypothesis affects the design and evaluation of experiments in frequentist inference, too.

The same dependency is even more evident when frequentist inference is used for supporting practical decisions. As already argued by Rudner (1953), choosing and balancing type I and type II error levels involves non-cognitive value judgments: we implicitly reveal how severe and how probable we find these errors. A decision about whether or not to accept/reject a hypothesis and to act on its basis must trade off plausibility with the utility of an act and the strength of the evidence. Hence, the view that degrees of belief must not play any role in assessing evidential support is taking the value-free ideal and the idea of detached objectivity one step too far. If applied rigorously, this would mean that we could stop making inferences from data to theory. Indeed, also Douglas (2004, 460) stresses that objectivity in scientific reasoning does not imply the elimination of personal perspective; this would actually be a gross misrepresentation of how science works. Therefore we conclude that the second challenge is misguided, too: scientific evidence cannot, and should not, be neatly separated from judgments of plausibility and degrees of belief.

Challenge 3: Neglect of Experimental Design

The third challenge to subjective Bayesianism concerns the problem of bias in trials with interim looks at the data. The problem can best be motivated with an example from medicine. Randomized Controlled Trials (RCTs) are currently the gold standard within evidence-based medicine. They are usually conducted as **sequential trials** allowing for monitoring for early signs of effectiveness or harm. In sequential trials, data are typically monitored as they accumulate. That is, we have interim looks at the data and we may decide to stop the trial before the planned sample size is reached. By terminating a trial when overwhelming evidence for the effectiveness or harmfulness of a new drug is available, the prohibitive costs of a medical trial can be limited and in-trial patients are protected against receiving inferior treatment.

Such truncated trials are often seen as problematic. In a review of

134 trials stopped early for benefit, Montori et al. (2005) point to an inverse correlation between sample size and treatment effect: the smaller the sample size achieved by the trial at the moment of stopping, the larger the estimate it provided for the effect. These findings are supported by a more recent study by Bassler et al. (2010) where truncated trials report significantly higher effects than trials that were not stopped early. While the authors of these studies do not object to monitoring and truncating trials in general, they advocate that results (e.g., effect size estimates) from such trials be treated with caution. Bayesian measures of evidence such as the Bayes factor do not depend on the sampling protocol or experimental design and evaluate truncated trials like fixed-sample trials. This seems to introduce a bias toward overestimating effect sizes.

Indeed, critics of Bayesian inference such as Deborah Mayo (1996) complain that decoupling statistical inference from the sampling protocol "can lead to a high probability of error, and [...] this high error probability is not reflected in the interpretation of data" (Mayo and Kruse, 2001). In the context of medical research, the Bayesian seems to provide *carte blanche* for implementing any design that favors the pursuit of certain non-cognitive values, such as the financial interests of the trial sponsor. For instance, we could sample on until a convincing result is reached, conduct a Bayesian analysis and submit the study for publication, without mentioning the biased sampling procedure. After all, whether the data were obtained by means of a biased sampling protocol, an unbiased protocol or no protocol at all affects neither the posterior probability nor the Bayes factor. Again, the perceived threat to the objectivity of Bayesian inference comes from the hidden intrusion of bias and non-cognitive values into statistical reasoning.

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

of their outcome. In that light, it is questionable whether the observed effect size difference between truncated and non-truncated trials is really problematic.

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

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

Fourth, the neglect of stopping rules follows immediately from a fundamental principle of Bayesian inference: the **Likelihood Principle**. According to that principle, all experimental evidence (=judgments of evidential support) about an unknown parameter θ is contained in the likelihood function $L_{\mathrm{E}}(\theta) = p(\mathrm{E}|\theta)$ for observed data E. Formally, the principle

is stated thus:

> **Likelihood Principle (LP):** Consider a statistical model $\mathcal{M}$ with a set of probability measures $p(\cdot|\theta)$ parametrized by $\theta \in \Theta$. Assume we conduct an experiment $\mathcal{E}$ in $\mathcal{M}$. Then, all evidence about θ generated by $\mathcal{E}$ is contained in the *likelihood function* $p(\mathrm{E}|\theta)$, where the observed data E are treated as a constant. (Birnbaum, 1962; Berger and Wolpert, 1984)

This principle is one of the cornerstones of Bayesian inference. As Birnbaum (1962) showed in a celebrated paper, it can be derived from two more foundational principles: the Sufficiency Principle and the Conditionality Principle.

We begin with the first one. A statistic (i.e., a function of the data X) $T(X)$ is *sufficient* if the distribution of the data X does not depend on the unknown parameter θ, conditional on T. In other words, sufficient statistics are compressions of the data set that do not lose any relevant information about θ. An example is an experiment about the bias of a coin. Assuming that the tosses are independent and identically distributed, the overall number of heads and tails is a sufficient statistics for an inference about the bias of the coin. Thus, we can neglect the precise order in which the results occurred. Formally, the **Sufficiency Principle** states that any two observations x_1 and x_2 are evidentially equivalent with regard to the parameter of interest θ as long as $T(x_1) = T(x_2)$ for a sufficient statistic T. Therefore, the principle is usually accepted by Bayesians and frequentists alike.

The **Conditionality Principle** is more controversial: it states that evidence gained in a probabilistic mixture of experiments is equal to the evidence in the actually performed experiment. In other words, if we throw a die to decide whether experiment $\mathcal{E}_1$ is conducted (in case the die comes up with an odd number) or experiment $\mathcal{E}_2$ (even number) and we throw a six, then the evidence from the overall experiment $\mathcal{E} = \mathcal{E}_1 \oplus \mathcal{E}_2$ is equal to the evidence from $\mathcal{E}_2$. Frequentists usually reject Conditionality since their measures of evidence take the entire sample space into account. A thorough discussion of these principles goes beyond the scope of this variation and can be found in, e.g., Mayo (2010) or Gandenberger (2015). To the extent that the Sufficiency and Conditionality Principle are found compelling requirements for objective, truth-directed inference that is free of

non-cognitive values, the Likelihood Principle constrains our judgments of evidential support in a way that is incompatible with frequentist inference, e.g., *p*-values.

In particular, the Likelihood Principle implies the **Stopping Rule Principle** (Berger and Berry, 1988, 34). Since the Likelihood Principle implies that only information contained in the likelihood function are evidentially relevant, and since the likelihood functions of the parameter values under different stopping rules are proportional to each other (proof omitted), stopping rules cannot have an evidential role. Indeed, Bayesians argue that

> The design of a sequential experiment is [...] what the experimenter actually *intended* to do. (Edwards et al., 1963, 239. See also Savage 1962, 76.)

In other words, since such intentions are "locked up in [the experimenter's] head" (ibid.), not verifiable for others, and apparently not causally linked to the data-generating process, they should not matter for sound statistical inference. Hence the dismissal of stopping rules in Bayesian judgments of evidential support.

This position has substantial practical advantages: if trials are terminated for unforeseen reasons, e.g. because funds are exhausted or because unexpected side effects occur, the observed data can be interpreted properly in a Bayesian framework, but not in a frequentist framework. Same for cases where the sampling protocol cannot be retrieved (e.g., historical records). As externally forced discontinuations of sequential trials frequently happen in practice, claims to the evidential relevance of stopping rules would severely compromise the proper interpretation of sequential trials.

In total, the claim that Bayesian inference in sequential trials contains an implicit bias or invalidates scientific reasoning can be soundly rebutted. The particular problem of sequential analysis and monitoring ongoing trials poses no challenge to Bayesian inference that is not equally pressing for its competitors, such as frequentist inference.

Discussion: A Digression on Scientific Objectivity

The concept of scientific objectivity is a notoriously difficult one, with various aspects and interpretations. It is a commonly shared view, though,

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

First, there is the idea that subjective Bayesian inference is particularly vulnerable to the intrusion of bias and non-cognitive values due to their dependence on prior probabilities, and the lack of restrictions on choosing them. However, prior degrees of belief can incorporate valuable expertise and background information and they can (and should!) be criticized like any statistical model assumption. Once these points are recognized, the challenge loses its bite. It can also be demonstrated that sensitivity analysis in Bayesian inference contributes to convergent objectivity in Douglas's sense: validation of a result from different independent perspectives.

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

Third, Bayesian inference is criticized for neglecting that certain experimental designs may lead to biased effect sizes. However, this criticism relies on a questionable description of evidence from truncated trials and on a failure to recognize how observations and prior belief are amalgamated in Bayesian inference. In addition, the Likelihood Principle provides a justification for why stopping rules, and aspects of experimental design more generally, should not affect post-experimental judgments of evidential support.

It is also worth glossing on aspects of objectivity that relate to the social dimension of science. Helen Longino (1990) has forcefully argued that scientific objectivity is not only about scientific reasoning itself, but also about the structure of scientific discourse: the possibility of openly criticizing each other's assumptions, providing a floor for the exchange of rational arguments, etc. In this respect, Bayesian inference has several important assets: it is honest and transparent about the assumptions it makes and clearly distinguishes between prior belief, evidence, and conclusions (=posterior belief). This points out clear avenues for model criticism and allows for a straightforward detection of inappropriate bias, such as prior assumptions that heavily favor a particular hypothesis. Moreover, subjective Bayesianism provides a rigorous description of what happens when the prior assumptions on a parameter value are varied. The transparency of the role of individual degrees of belief can be seen as a plus of subjective Bayesianism from the vantage point of scientific objectivity.

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

One caveat, though. We do not want to promote subjective Bayesianism as a one-size-fits-all solution for problems of statistical inference, and more generally, scientific inference. We have mentioned in the previous variation how difficult it can be to come up with meaningful subjective priors, and how computationally expensive the calculation of posteriors can be. Rather, we have argued that when an inference problem lends itself to a Bayesian analysis, the apparent lack of objectivity, due to the use of subjective degrees of belief, may be misleading. For subjective Bayesian inference, the objectivity problem is no more and less pressing than for any other inference method in science.

This is not meant to deny that there are many open challenges for Bayesian inference with respect to their objectivity. For example, the question of meta-analysis of experiments translates, on a Bayesian reading, into the aggregation of posterior probability distributions. How can such a pooling procedure be immunized against the intrusion of bias (e.g., by

manipulating the prior probability distributions in the individual studies)?

Second, in the light of the discussion about the replicability of psychological research and the reliability of statistical analysis (Ioannidis, 2005; Makel et al., 2012; Francis, 2014; Francis et al., 2014; Open Science Collaboration, 2015), some methodologists have called for radical conclusions. Taking issues with the NHST methdology which they see as highly problematic, Trafimow and Marks (2015) have banned p-values from the journal they are editing, the Journal of Basic and Applied Social Psychology (BASP). This is not to say that they are in favor of Bayesian inference: they see it as an alternative to frequentist inference that is viable in some cases, but often struggles to come up with meaningful prior distributions. Therefore, they recommend to conduct statistical analysis without inferential tools and to rely exclusively on descriptive statistics (e.g., effect sizes, correlation coefficients), which they see as more objective. Defending Bayesian inference against this minimalist approach is an exciting task for those who are convinced by the arguments in this variation.

Third and last, there are various varieties of "objective Bayesian inference", which try to find a middle ground between the two grand schools of statistical inference. With the exception of the Principle of Maximum Entropy, they have not received much attention from philosophers. Future research should take these approaches, explained below, more seriously and investigate whether they build a philosophically sound and practically viable bridge between Bayesian and frequentist inference that lives up to the ambitions of objectivity.

Objective Priors The method of objective priors in Bayesian inference tries to get the sting out of the first challenge (arbitrariness of priors) by giving up the static dimension of Bayesianism—that probabilities represent subjective degrees of belief. Instead, this method advocates priors that implement a sort of Principle of Indifference between the hypotheses under consideration, such as assigning equal probability to each parameter value. The problem with this approach is that the underlying Principle of Indifference is philosophically shaky (e.g., Hájek, 2011). However, statisticians have also worked on refinements of this approach on information-theoretic grounds (e.g., Jeffreys, 1961). José Bernardo (1979a,b, 2012) proposed so-called reference priors, motivated by invariance under 1:1-transformation of parameters of interest. See Sprenger (2012) for a discussion of their

philosophical implications.

The Principle Maximum Entropy This approach parts with the dynamic dimension of Bayesian inference: the use of Bayesian Conditionalization as a principle of belief revision. Rather, an agent's rational degrees of belief should satisfy three constraints (Jaynes, 1968; Williamson, 2010): they should (i) conform to the axioms of probability; (ii) satisfy empirically given constraints on our rational degrees of belief; and given these constraints, (iii) they should be **equivocal**, that is, as middling as possible. This amounts to maximizing the entropy of the probability function that represents an agent's rational degree of belief. If ω denotes the atoms of the relevant σ-algebra, the entropy is given by the term

$$H = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega). \tag{11.3}$$

While the Principle of Maximum Entropy is of great help in many practical problems in engineering, computer science, and related disciplines, it is hard to find a waterproof epistemic or decision-theoretic justification for why degrees of belief should in general be as middling as possible (Seidenfeld, 1979, 1986).

Conditioning on Evidence Strength Of the three approaches discussed, this is the least known one. The idea is to give a valid Bayesian interpretation to frequentist error probabilities (type I and type II error) by appropriate conditioning on the strength of the observed evidence, e.g., an observed p-value (Berger et al., 1994, 1997; Berger, 2003). It can be shown that in many cases, Bayesian and frequentist reasoners agree numerically in conditional inference; they just use different interpretations. Moreover, conditional inference is directly applicable to salient problems in the analysis of sequential trials in medicine (Nardini and Sprenger, 2013). These attempts to find a compromise between Bayesian and frequentist inference are, for the most part, still terra incognita from a philosophical point of view, but they strike us as original and worthy of further attention.

The Theme Revisited

This book is an attempt to analyze and to elucidate scientific reasoning by means of subjective Bayesian inference. Subjective Bayesians assign probabilities to scientific theories and interpret these values as personal degrees of belief. By the principle of Bayesian Conditionalization, or one of its generalizations, these degrees of belief are changed in the light of incoming evidence. While scientific reasoning is, of course, much broader and richer than what can be expressed by degrees of belief, subjective probability provides fruitful explications of several important concepts in science (such as confirmation, explanatory power, and causal effect), and it reconstructs several prominent argument patterns, such as the NAA and the NMA. The book brings together and unifies these Bayesian models in philosophy of science.

More precisely, we vary the Bayesian theme $p(\mathrm{E}|\mathrm{H}) = p(\mathrm{H})p(\mathrm{E}|\mathrm{H})/p(\mathrm{E})$ into three directions. The first set of variations investigates various confirmatory arguments in science, most of which are not straightforwardly captured in the Bayesian framework. Variation 1 provides a framework for learning conditionals (e.g., scientific hypotheses of the form "if A, then B") in a Bayesian framework that is immune to the drawbacks of previous proposals. Variation 2 presents and evaluates different proposals for quantifying degree of confirmation. Variation 3 deals with the Problem of Old Evidence—the problem of describing how learning a dependency between theory and evidence can boost confidence in a theory when the evidence itself is already known. In this variation, two novel solutions are provided. Variation 4 on the No Alternatives Argument shows how the failure to find alternatives to a theory can confirm a theory even in the absence of genuine empirical evidence. The argument also suggests how Inference to the Best Explanation can be justified within a confirmation-theoretic perspective. Variation 5 frames the famous No Miracles Argument in favor of scientific realism

in Bayesian terms and investigates its scope and limits according to different ways to frame and to model the argument. Taken together, these variations demonstrate that Bayesian confirmation theory extends beyond the standard case of evaluating predictions of a scientific hypothesis: it suits a remarkable variety of modes of scientific reasoning.

The second set of variations focuses on causal effect, explanatory power, and intertheoretic coherence. While Variations 6 and 7 provide axiomatic characterizations of causal effect and of explanatory power, Variation 8 demonstrates how the establishment of intertheoretic reductive relations can raise the probability of a scientific theory at the fundamental level.

The third set of variations is motivated by issues in statistical inference. Variation 9 closes a lacuna in the methodology of hypothesis testing by developing a probabilistic measure of corroboration—a task which is of utmost importance for scientific practice (e.g., for the interpretation of non-significant results). Variation 10 investigates the role of simplicity in Bayesian model selection. Finally, Variation 11 takes up various challenges to the objectivity of Bayesian inference and demonstrates that it is no less objective than its frequentist competitors.

These final variations also demonstrate the limits of Bayesian modeling—for example, in Variation 10, we see that popular model selection criteria fail to have a Bayesian justification, and that they should, despite being known as Bayesian criteria, rather be seen as Bayesian *heuristics*. Variation 9 argues that there cannot be a purely Bayesian, confirmation-theoretic explication of corroboration. Bayesian and non-Bayesian approaches have to be combined in order to measure the degree to which a hypothesis has stood up to severe tests. Finally, Variation 11 investigates scope and limits of objectivity in Bayesian reasoning.

It is also notable that there is a high degree of methodological similarity between the different variations, despite the divergent nature of their explicanda. For instance, Variation 6 on causal effect transfers techniques from Bayesian Confirmation Theory (Variation 2) almost one-to-one. Similar things can be said about Variation 7 (explanatory power). Variation 3 on the Problem of Old Evidence benefits from our improved account of learning conditional information in Variation 1. Finally, Variation 4 and 5 model the assessment of a scientific theory by means of including an additional propositional variable: the number of available alternatives.

We do not want to convince the reader that Bayesian modeling is a universal method or solution to all topics and problems in philosophy of science. This standpoint would be exposed to manifold criticism, e.g., see Norton (2003, 2011) for foundational criticisms of purely formal accounts of inductive inference. What we hope to have demonstrated is much less ambitious: that Bayesian inference is more than a simple and appealing theory for representing and updating degrees of belief. It is home to powerful models that can be applied to a surprising variety of problems in scientific reasoning.

The use of Bayesian inference as a model for explicating scientific values is also characteristic of our general methodology. Indeed, our approach can be characterized as "scientific philosophy"—not in the sense of the logical empiricists or naturalists such as W.V.O. Quine (1969), but in an understanding that is closer to the views of Hans Reichenbach (1951) and Hannes Leitgeb (2013). See also Hartmann and Sprenger (2012). The logical empiricists (e.g. Carnap, 1935) understood scientific philosophy as the task of refining, improving and laying the foundations for a language of science. Naturalists such as Quine saw philosophy as a branch of science—e.g., epistemology was thought to reduce to cognitive psychology. Recently, Maddy (2009) and Ladyman and Ross (2009) have tried to revive this style of philosophical theorizing. We do believe, however, that the epistemic problems of science are *genuinely philosophical problems* that cannot be reduced to purely scientific questions (above all, because of the normative character of many questions). Like Leitgeb, we believe that such questions can be addressed with the use of scientific tools, that is, formal modeling, case studies, experimentation and computer-based simulations, which can be fruitfully combined with conceptual analysis as a core methods of philosophical analysis.

All these methods have a history in philosophy, some longer, some shorter. Conceptual analysis goes back to the very beginnings of philosophy, e.g., Plato's famous analysis of knowledge. Mathematical and logical analysis have a similarly rich history, going back to Aristotle's logic and the Medieval logicians. Interestingly, mathematics, and probability theory in particular, have been used less frequently for explicating philosophical arguments—Hume's Dialogues on Natural Religion and the famous 10th chapter of the *Enquiry on Human Understanding* ("Of Miracles") being among the notable exceptions. Case studies–already part of Bacon's

Novum Organon and the Descartes' *Discours de la Méthode* (Part V)—have been popular in philosophy of science since the 1960s and 1970s. They answered the need for calibrating general philosophical models of scientific reasoning, such as those provided by the logical empiricists, with the practice of science and have inspired philosophical theorizing ever since. For instance, the mechanistic model of explanation popularized by Machamer et al. (2000) and Craver (2007) heavily draws on case studies in cognitive and biological sciences. Experimental methods, by contrast, have a quite young history in philosophy, related to the emergence of experimental philosophy as a part of epistemology (e.g., Stich, 1988; Weinberg et al., 2001; Alexander and Weinberg, 2007), collaborations between cognitive scientists and philosophers of science (e.g., Crupi et al., 2008, 2013; Colombo et al., 2016a). Finally, over the last years, computational methods and agent-based simulations in particular have gained ground in philosophy of science. Often they are used to study the emergence and stability of social norms and contracts (e.g., Alexander, 2007; Skyrms, 2010; Muldoon et al., 2014), but sometimes they are also applied to modeling scientific progress and the communication structure of epistemic communities (e.g., Zollman, 2007; Weisberg and Muldoon, 2009; De Langhe and Rubbens, 2015; Heesen, 2016a). Of particular interest are those studies where probabilstic reasoning in science (e.g., NHST) interacts with rewards and biases in the scientific community (e.g., Romero, 2016). Notably, all these methods are rarely combined with each other, and doing so is perhaps one of the main innovations of this book.

Indeed, most variations in this book feature a majority of these methods. Conceptual analysis and formal modeling, the core methods of our explicative project, are used in almost any of the eleven variation. The final Variation 11, which evaluates the objectivity of Bayesian reasoning, is perhaps the only one that explicitly eschews formal modeling. Case studies play an important role in Variation 3 (confirmation by old evidence), Variation 4 (string theory as an example of the NAA), Variation 6 (measuring causal effect), Variation 8 (reduction in statistical mechanics), Variation 9 (corroboration and null hypothesis significant testing), and Variation 10 (evaluating different model selection criteria). These variations also address methodological problems in specific disciplines (e.g., Variation 4 for particle physics, Variation 6 for cognitive psychology and medical science, Variation 9 and 10 for statistics). Experimental evidence from psychol-

Method	Relevant Variation
Conceptual Analysis	all
Formal Modeling	1-10
Case Studies + Addressing Discipline-Specific Methodological Problems	3, 4, 6, 8, 9, 10
Experimental Evidence	1, 2, 7, 9
Computational Methods	3, 4, 5, 7, 8

Table 12.1: An overview of the methods used in the book.

ogy and cognitive science is cited in Variation 1 (learning conditionals), Variation 2 (judgments of confirmation), Variation 6 (causal induction), Variation 7 (judgments of explanatory power) and Variation 9 (scientists' use of null hypothesis significance tests). Finally, computational methods are used—sometimes behind the screens—in Variation 3 (comparing our assumptions to those proposed by Jeffrey & Co.), Variation 4 (degree of confirmation of the NAA), Variation 5 (scientrometic analysis of theoretical developments), Variation 7 (explanatory power vs. posterior probability) and Variation 8 (degree of confirmation for successful reductions). Table 12.1 gives a schematic overview.

We finish by sketching open questions for future research. For the reader's convenience, we recap three open research questions from each variation below.

Variation 1 Learning Conditional Information

- Applying different divergence measures to learning conditional information
- Transferring the analysis from indicative to subjunctive conditionals
- Developing a general theory for causal and evidential constraints on degree of belief

Variation 2 Confirmation

- Investigating Information-theoretic foundations of confirmation measures
- Applying confirmation theory to the diagnostic value of scientific tests (e.g., in medicine)
- Using confirmation judgments to explain phenomena in the psychology of reasoning

Variation 3 The Problem of Old Evidence (POE)

- Applying the POE solutions to the prediction/accommodation problem
- Integrating the POE with an analysis of explanatory reasoning in science
- Solving the POE in terms of learning conditional information ($\rightarrow$ Variation 1)

Variation 4 The No Alternatives Argument (NAA)

- Relating the NAA to eliminative inference
- Applying of the NAA to Inference to the Best Explanation
- Finding instances of NAA-based reasoning in diverse scientific fields

Variation 5 Scientific Realism and the No Miracles Argument (NMA)

- Studying parallels between the NAA and the NMA
- Conducting a scientometric analysis of theoretical stability in different scientific disciplines
- Extending the NMA towards the full realist thesis

Variation 6 Causal Effect

- Calculating causal effect complicated network structures
- Generalizing the causal effect measure to real-valued variables and integrating it with statistical effect size measures
- Proposing a unified probabilistic theory of causal strength and causal specificity

Variation 7 Explanatory Power

- Conducting experiments on the determinants of explanatory judgments
- Relating measures of explanatory power to measures of confirmation and causal effect
- Developing a normatively convincing Bayesian account of IBE and abductive inference

Variation 8 Intertheoretic Reduction

- Checking the robustness of the analysis under different confirmation measures
- Describing intertheoretic reduction as increasing the coherence of a set of theories
- Modeling the disconfirmation of the phenomenological theory by the fundamental theory

Variation 9 Hypothesis Testing and Corroboration

- Extending the corrorboration measure to more complicated statistical inference problems (nuisance parameters, hierarchial modeling, etc.)
- Conducting statistical meta-analysis with corroboration judgments
- Finding case studies for corroboration-based reasoning in the history of science

Variation 10 Simplicity

- Exploring the context-sensitivity of model selection criteria
- Working out the thesis of instrumental Bayesianism, and transferring it to other areas of statistical inference
- Comparing the role of simplicity in model selection to simplicity in causal and explanatory inference (e.g., IBE $\rightarrow$ Variation 5 and 7)

Variation 11 Objectivity

- Assessing whether Bayesian analysis is less vulnerable to replication failure than frequentist analysis (i.e., NHST)
- Building bridges between Bayesian and frequentist reasoning
- Investigating the philosophical foundations of objective Bayesian methods

Now, we proceed to describing five major research projects that fit into the Bayesian philosophy of science research program, and that bridge different topics discussed in this book.

An obvious direction into which our research program could be extended is the **integration of probabilistic, causal and counterfactual reasoning**. The analysis of learning conditionals and confirmation by old evidence in Variation 1 and 3 showed that causal considerations often constrain an agent's rational degrees of belief. Moreover, the evaluation of conditional degrees of belief proceeds counterfactually, and via the notion of an intervention, counterfactual considerations play an important role in evaluating causal relations, too. While we restrict ourselves to solving some concrete problems at the intersection of causal and probabilistic inference, future work should come up with an integrated theory of causal induction, Bayesian learning and conditional reasoning (e.g., building on Oaksford and Chater, 2000; Pearl, 2000; Douven, 2016; Over, 2016). Apart from theoretical pioneering work, we see a lot of promise in experiments that investigate the role of causal structure in reasoning with conditionals. The same holds true for experiments that predict the truth or acceptability of a conditional by the strength of the expressed causal effect ($\rightarrow$ Variation 6).

Second, Bayesian confirmation is intimately related to causal and explanatory considerations (Variations 2, 6 and 7). On a theoretical level, this calls for an **extended analysis of the relationship between measures of confirmation, causal effect and explanatory power**, similar to what Schupbach (2016) did for measures of explanatory power and posterior probability. This should lead to a sharper demarcation of these concepts and to a description of the conditions when the one is conducive to the other. On an empirical level, we propose experiments that uncover correlations and differences in judgments of explanatory power, causal strength, and probability, in order to reveal the determinants of explanatory judgments and to provide a more nuanced and descriptively appropriate view of explanatory reasoning. Since explanation is a concept which is loaded with causal and probabilistic connotations, such experiments strike us as highly valuable. Colombo et al. (2016a) and Colombo et al. (2016b) already made some steps in this direction, but there is still much work to be done. See also the survey by Sloman and Lagnado (2015). This research could also be related to the role of simplicity in scientific reasoning: our formal analysis of simplicity in model selection ($\rightarrow$ Variation 10) should, at some point, be complemented by an empirical investigation of how simplicity considerations affect scientific reasoning, similar to what Lombrozo (2007)

did for simplicity in explanatory judgments.

Third, the material in this book is an outstanding basis for a detailed investigation of the **scope and rationality of Inference to the Best Explanation**. Not only that one can assess IBE on the basis of diverse measures of explanatory power (→ Variation 7), it is also possible to relate IBE to other argument patterns that we explicated in this book: the No Alternatives Argument (NAA, → Variation 4) and the No Miracles Argument (NMA, → Variation 5). As we argued in those Variations, both arguments are essentially abductive in arguing that the empirical adequacy of a theory T is the best explanation for the absence of viable alternatives (→ NAA) and the predictive success of T (→ NMA).

Fourth, this book does not include a variation on **unification and its role in reduction and explanation** (→ Variation 7 and 8)—mainly because there is not much literature on this topic from a Bayesian perspective. Unification is traditionally regarded as an important cognitive value in scientific reasoning (e.g. McMullin, 1982; Douglas, 2013), as a value that counts for most scientists as a reason to accept a theory and to pursue it further. Based on the pioneer work done by Myrvold (2003, 2016) and Schupbach (2005), it seems plausible to explicate unification by means of confirmation-theoretical or information-theoretic models, to explain its role in intertheoretic reductions and explanatory reasoning, and to describe unification in important case studies, such as Bayesian cognitive science (Colombo and Hartmann, 2016).

Fifth and last, Bayesian methods can provide **better foundations for hypothesis testing in science**. It has been frequently noted that the current method of hypothesis testing, essentially based on p-values, is not only at odds with the very principles of Bayesian reasoning, but also a danger for the reliability of scientific inquiry (e.g., Berger and Sellke, 1987; Goodman, 1999a; Cumming, 2012, 2014). It is therefore important to integrate Bayesian reasoning into hypothesis tests and to reconcile both paradigms (Wetzels et al., 2009; Wetzels and Wagenmakers, 2012; Lee and Wagenmakers, 2013; Morey et al., 2014, 2016). However, doing so is often far from straightforward due to the different motivations that feed Bayesian confirmation theory (→ Variation 2) and hypothesis testing in the tradition of Popper and Fisher. Variation 9 makes an attempt to quantify the degree of corroboration of a hypothesis and Variation 6 axiomatizes various measures of causal effect that could result from RCTs or case-control studies.

These projects need to be pursued further in order to obtain a full Bayesian account of hypothesis testing.

This brings us, finally, to a wider perspective on Bayesian philosophy of science. First, this book has neglected the **social dimension of science**. We have, so far, focused on the perspective of an individual scientist (or a homogenous research team) who does experiments, analyzes data and assesses theories. Future work could link the issues covered in this book to questions about merging opinions and the role of experts in science (for survey articles, see Dietrich and List, 2016; Martini and Sprenger, 2016). For a yet different research program in the social epistemology of science that can be tackled by Bayesian models, consider the exploration of epistemic landscapes and the credit reward system in science (e.g., Zollman, 2007; Weisberg and Muldoon, 2009; Heesen, 2016a,b).

Second, we could tighten the link between Bayesian reasoning in philosophy and **Bayesian reasoning in science** (e.g., Bayesian statistics). In this book, we have only scratched the surface, but there is a fascinating and largely unexplored set of questions how philosophical insights about Bayesian reasoning and hypothesis testing should translate into practical statistical reasoning (e.g., Gallistel, 2009; Bernardo, 2012; Sprenger, 2013b). The fifth research project listed above, concerned with scientific hypothesis testing and Bayesian Inference, falls into this domain. But there are also more general methodological questions. For example, Gelman and Shalizi (2012, 2013) suggest that Bayesian inference is very convenient at the micro-level of statistical inference within a given class of models, but that the proper task of model testing and evaluation rather follows a hypothetico-deductive rationale. Moreover, there has not yet been a systematic investigation and comparison of the philosophical foundations of the different objective Bayesian approaches, and the conceptions of objectivity that they endorse. Given the centrality of claims to objectivity in the evaluation of research findings in modern science, this strikes us as a highly worthwhile endeavor.

Both projects can also join forces. The social sciences, psychology in particular, are undergoing a **replication crisis**, that is, difficulties to reproduce findings from published experiments (e.g., Galak et al., 2012; Makel et al., 2012; Francis, 2014; Francis et al., 2014; Open Science Collaboration, 2015). How much can we trust scientific method if research results are so often unstable? Bayesian methods may provide an answer to this question:

they have been used to explain why the current publication culture promotes unreliable findings and to point out how such biases can be cured (e.g., Ioannidis, 2005; Ioannidis and Trikalinos, 2007). The work in this book, especially in Variation 2, 9 and 11, provides a starting point for further philosophical investigation of these problems. Working along these lines would combine Bayesian philosophy of science with the practice of Bayesian statistics and a social perspective of the scientific enterprise.

All in all, it should be clear by now that there is an exciting and inexhaustible set of unanswered research questions in Bayesian philosophy of science. Therefore, we predict that the Bayesian research program will have a bright and fascinating future—in philosophy of science and beyond.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic*, 50(2):510–530.

Alexander, J. and Weinberg, J. M. (2007). Analytic Epistemology and Experimental Philosophy. *Philosophy Compass*, 2(1):56–80.

Alexander, J. M. (2007). *The structural evolution of morality*. Cambridge University Press, Cambridge.

Allais, M. (1953). Le Comportement de l'Homme Rationnel Devant le Risque: Critique des Postulats et Axiomes de l'École Américaine. *Econometrica*, 21:503–546.

Allais, M. and Hagen, O. (1979). *Expected Utility Hypotheses and the Allais Paradox*. Reidel, Dordrecht.

Aquinas, T. (1945). *Basic Writings of St. Thomas Aquinas*. Random House, New York.

Atkinson, D. (2012). Confirmation and Justification: A Commentary on Shogenji's Measure. *Synthese*, 184:49–61.

Baker, A. (2003). Quantitative parsimony and explanation. *British Journal for the Philosophy of Science*, 54:245–259.

Baker, A. (2010). Simplicity. In *The Stanford Encyclopedia of Philosophy*.

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

Batterman, R. W. (2002). *The devil in the details: asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press, Oxford.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418.

Beckers, S. and Vennekens, J. (2016). A Principled Approach to Defining Actual Causation.

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

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science*, 18:1–32.

Berger, J. O. and Berry, D. A. (1988). The relevance of stopping rules in statistical inference. In Berger, J. O. and Gupta, S. S., editors, *Statistical decision theory and related topics: Vol. IV*, pages 29–72. Springer, New York.

Berger, J. O., Boukai, B., and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science*, 12(3):133–160.

Berger, J. O., Brown, L. D., and Wolpert, R. L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, 22(4):1787–1807.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association*, 82:112–122.

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward/CA.

Bernardo, J. M. (1979a). Expected information as expected utility. *The Annals of Statistics*, 7:686–690.

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

Bernardo, J. M. (1999). Nested hypothesis testing: The Bayesian reference criterion. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6: Proceedings of the Sixth Valencia Meeting on Bayesian Statistics*, pages 101–130 (with discussion). Oxford University Press, Oxford.

Bernardo, J. M. (2012). Integrated objective Bayesian estimation and hypothesis testing. In *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*, pages 1–68 (with discussion). Oxford University Press, Oxford.

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

Berry, S. M., Carlin, B. P., and Connor, J. (2010). Bias and Trials Stopped Early for Benefit. *Journal of the American Medican Association*, 304(2):156.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. The MIT Press, Cambridge, MA.

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

Blackwell, D. and Dubins, L. (1962). Merging of Opinions with Increasing Information. *The Annals of Mathematical Statistics*, 33(3):882–886.

Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*. Oxford University Press, New York.

Boyd, R. (1981). Scientific Realism and Naturalistic Epistemology. In Asquith, P. and Giere, R., editors, *PSA 1980*, volume II, pages 613–662. Philosophy of Science Association, East Lansing, MI.

Boyd, R. (1983). On the Current Status of the Issue of Scientific Realism. *Erkenntnis*, 19:45–90.

Boyd, R. (1984). The Current Status of Scientific Realism. In Leplin, J., editor, *Scientific Realism*, pages 41–82. University of California Press, Berkeley/CA.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Bromberger, S. (1965). An Approach to Explanation. In Butler, R. J., editor, *Analytical Philosophy*, pages 72–105. Oxford University Press, Oxford.

Brush, S. G. (1989). Prediction and Theory Evaluation: The Case of Light Bending. *Science*, 246:1124–1129.

Brössel, P. (2013). The problem of measure sensitivity redux. *Philosophy of Science*, 80(3):378–397.

Brössel, P. (2016). *Rethinking Bayesian Confirmation Theory*. Springer, Berlin.

Brössel, P. and Huber, F. (2015). Bayesian Confirmation: A Means with No End. *The British Journal for the Philosophy of Science*, 66:737–749.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods Research*, 33:261–304.

Buss, D. M. (1998). Sexual strategies theory: Historical origins and current status. *Journal of Sex Research*, 35(1):19–31.

Buss, D. M. and Schmitt, D. P. (1993). Sexual Strategies Theory: An Evolutionary Perspective on Human Mating. *Psychological Review*, 100(2):204–232.

Bylander, T., Allemang, D., Tanner, M. C., and Josephson, J. R. (1991). The computational complexity of abduction. *Artificial Intelligence*, 49(1):25–60.

Callender, C. (2001). Taking Thermodynamics Too Seriously. *Studies in History and Philosophy of Modern Physics*, 32:539–553.

Carnap, R. (1935). *Philosophy and Logical Syntax*. Kegan Paul, Trench, Trubner & Co., London.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Cartwright, N. (1979). Causal Laws and Effective Strategies. *Noûs*, 13(4):419–437.

Chakravartty, A. (2011). Scientific Realism. In *The Stanford Encyclopedia of Philosophy*.

Chase, W. and Brown, F. (2000). *General Statistics*. Wiley, New York.

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review*, 104(2):367–405.

Chockler, H. and Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, pages 93–115.

Christensen, D. (1999). Measuring Confirmation. *Journal of Philosophy*, 96(9):437–461.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge University Press, Cambridge.

Churchland, P. M. (1985). Reduction, Qualia, and the Direct Introspection of Brain States. *The Journal of Philosophy*, 82:8–28.

Cohen, M. (2016a). Explanatory Justice: The Case of Disjunctive Explanations.

Cohen, M. P. (2015). On Schupbach and Sprenger's Measures of Explanatory Power. *Philosophy of Science*, 82:97–109.

Cohen, M. P. (2016b). On Three Measures of Explanatory Power with Axiomatic Representations. *British Journal for the Philosophy of Science*.

Colombo, M. (2016). Experimental Philosophy of Explanation Rising. The case for a plurality of concepts of explanation. *Cognitive Science*.

Colombo, M., Bucher, L., and Sprenger, J. (2016a). Determinants of Explanatory Judgment.

Colombo, M. and Hartmann, S. (2016). Bayesian Cognitive Science, Unification and Explanation. *The British Journal for the Philosophy of Science*.

Colombo, M., Postma, M., and Sprenger, J. (2016b). Explanatory Value, Probability and Abductive Inference. In *Proceedings of the CogSci2016*.

Colombo, M. and Wright, C. (2016). Explanatory Pluralism: An Unrewarding Prediction Error for Free Energy Theorists. *Brain and Cognition*.

Colyvan, M. (2001). *The Indispensability of Mathematics*. Oxford University Press, New York.

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

Cox, D. and Mayo, D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In Mayo, D. G. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, chapter 2, pages 276–304. Cambridge University Press, Cambridge.

Cox, R. (1946). Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14:1–10.

Craver, C. F. (2007). *Explaining the Brain*. Oxford University Press, Oxford.

Crupi, V. (2013). Confirmation. In *The Stanford Encyclopedia of Philosophy*.

Crupi, V., Chater, N., and Tentori, K. (2013). New Axioms for Probability and Likelihood Ratio Measures. *British Journal for the Philosophy of Science*, 64(1):189–204.

Crupi, V., Fitelson, B., and Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, 14:182–199.

Crupi, V. and Tentori, K. (2012). A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems). *Philosophy of Science*, 79(3):365–385.

Crupi, V. and Tentori, K. (2013). Confirmation as partial entailment: A representation theorem in inductive logic. *Journal of Applied Logic*, 11:364–372.

Crupi, V. and Tentori, K. (2014). Measuring information and confirmation. *Studies in the History and Philosophy of Science*, 47:81–90.

Crupi, V., Tentori, K., and Gonzalez, M. (2007). On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues. *Philosophy of Science*, 74:229–252.

Crupi, V., Tentori, K., and Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, 116:971–985.

Cumming, G. (2012). *Understanding the New Statistics*. Routledge, New York.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25:7–29.

Daston, L. and Galison, P. (2007). *Objectivity*. Cambridge University Press, Cambridge, MA.

Davies, H. T., Crombie, I. K., and Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal*, 316(7136):989–991.

Dawid, R. (2006). Underdetermination and Theory Succession from the Perspective of String Theory. *Philosophy of Science*, 73:298–322.

Dawid, R. (2009). On the Conflicting Assessments of the Current Status of String Theory. *Philosophy of Science*, 76:984–996.

Dawid, R., Hartmann, S., and Sprenger, J. (2015). The No Alternatives Argument. *The British Journal for the Philosophy of Science*, 66:213–234.

de Finetti, B. (1937). La Prévision: ses lois logiques, ses sources subjectives. In *Annales de l'Institut Henri Poincaré*.

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

de Finetti, B. (1974). *Theory of Probability*. John Wiley & Sons, New York.

de Finetti, B. (2008). *Philosophical Lectures on Probability*. Springer, Berlin.

De Langhe, R. and Rubbens, P. (2015). From Theory Choice to Theory Search: The Essential Tension Between Exploration and Exploitation in Science. In Devlin, W. J. and Bokulich, A., editors, *Kuhn's Structure of Scientific Revolutions—50 Years On*, pages 105–114. Springer.

de Regt, H. and Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, 144:137–170.

Deeks, J. (1998). When can odds ratios mislead? *British Medical Journal*, 317:1155.

Diaconis, P. and Zabell, S. L. (1982). Updating Subjective Probability. *Journal of the American Statistical Association*, 77(380):822–830.

Dietrich, F. and List, C. (2016). Probabilistic Opinion Pooling. In *Oxford Handbook of Probability and Philosophy*. Oxford University Press, Oxford.

Dietrich, F. and Moretti, L. (2005). On Coherent Sets and the Transmission of Confirmation. *Philosophy of Science*, 72:403–424.

Diez, J. (2011). On Popper's strong inductivism (or strongly inconsistent anti-inductivism). *Studies in History and Philosophy of Science A*, 42:105–116.

Dizadji-Bahmani, F., Frigg, R., and Hartmann, S. (2010). Who's afraid of Nagelian reduction? *Erkenntnis*, 73:393–412.

Dizadji-Bahmani, F., Frigg, R., and Hartmann, S. (2011). Confirmation and Reduction: A Bayesian Account. *Synthese*, 179:321–338.

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

Douglas, H. (2009a). Reintroducing Prediction to Explanation. *Philosophy of Science*, 76:444–463.

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

Douglas, H. (2011). Facts, Values, and Objectivity. *Ian Jarvie and Jesús Zamora*, pages 283–306.

Douglas, H. (2013). The Value of Cognitive Values. *Philosophy of Science*, 80(5):796–806.

Douven, I. (2011). Abduction. In *The Stanford Encyclopedia of Philosophy*.

Douven, I. (2012). Learning Conditional Information. *Mind & Language*, 27:239–263.

Douven, I. (2016). *The Epistemology of Indicative Conditionals*. Cambridge University Press, Cambridge.

Douven, I. and Dietz, R. (2011). A Puzzle about Stalnaker's Hypotehsis. *Topoi*, 30:31–37.

Douven, I. and Romeijn, J. W. (2011). A new resolution of the Judy Benjamin problem. *Mind*, 120(479):637–670.

Douven, I. and Schupbach, J. N. (2015a). Probabilistic Alternatives to Bayesianism: The Case of Explanationism. *Frontiers in Psychology*, 6.

Douven, I. and Schupbach, J. N. (2015b). The role of explanatory considerations in updating. *Cognition*, 142:299–311.

Dowe, D. L., Gardner, S., and Oppy, G. (2007). Bayes not Bust! Why Simplicity is no Problem for Bayesians. *British Journal for the Philosophy of Science*, 58:709–754.

Dowe, P. (2000). *Physical Causation*. Cambridge University Press, Cambridge.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P., editors (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. The MIT Press, Cambridge/MA.

Duhem, P. (1914). *La Théorie Physique: Son Objet, Sa Structure*. Vrin, Paris.

Dupré, J. (1984). Probabilistic Causality Emancipated. *Midwest Studies in Philosophy*, 9(1):169–175.

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

Easwaran, K. (2011a). Bayesianism I: Introduction and Arguments in Favor. *Philosophy Compass*, 6:312–320.

Easwaran, K. (2011b). Bayesianism ii: Applications and criticisms. *Philosophy*, 6:321–332.

Easwaran, K. (2011c). The Varieties of Conditional Probability. In Bandyopadhyay, P. S. and Forster, M. R., editors, *Handbook of the Philosophy of Statistics*, pages 137–148. Elsevier, Amsterdam.

Edgington, D. (1995). On Conditionals. *Mind*, 104:235–329.

Edgington, D. (2014). Indicative Conditionals. In *The Stanford Encyclopedia of Philosophy*.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press, Cambridge.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242.

Eells, E. (1985). Problems of Old Evidence. *Pacific Philosophical Quarterly*, 66:283.

Eells, E. (1990). Bayesian Problems of Old Evidence. In Savage, C. W., editor, *Scientific Theories*, pages 205–223. University of Minnesota Press, Minneapolis.

Eells, E. (1991). *Probabilistic causality*. Cambridge University Press, Cambridge.

Eells, E. and Fitelson, B. (2000). Measuring Confirmation and Evidence. *The Journal of Philosophy*, 97(12):663–672.

Eells, E. and Fitelson, B. (2002). Symmetries and Asymmetries in Evidential Support. *Philosophical Studies*, 107(2):129–142.

Eiter, T. and Gottlob, G. (1995). The complexity of logic-based abduction. *Journal of the ACM*, 42(1):3–42.

Ellsberg, D. (1961). Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*, 75:643–669.

Ellsberg, D. (2001). *Risk, Ambiguity and Decision*. Routledge, New York.

Fahrbach, L. (2009). The pessismistic meta-induction and the exponential growth of. In Hieke, A. and Leitgeb, H., editors, *Reduction and elimination in philosophy and the sciences. Proceedings of the 31th international Wittgenstein symposium.*

Fahrbach, L. (2011). How the growth of science ends theory change. *Synthese*, 108:139–155.

Festa, R. (2012). For unto every one that shall be given. Matthew properties for incremental confirmation. *Synthese*, 184:89–100.

Feyerabend, P. (1962). Explanation, Reduction and Empiricism. In Feigl, H. and Maxwell, G., editors, *Explanation, Reduction and Empiricism*, pages 28–97. University of Minnesota Press, Minneapolis.

Feyerabend, P. (1975). *Against Method*. Verso, London.

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

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

Fitelson, B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, 66:S362–S378.

Fitelson, B. (2001a). A Bayesian Account of Independent Evidence with Applications. *Philosophy of Science*, 68:S123–S140.

Fitelson, B. (2001b). *Studies in Bayesian Confirmation Theory*. PhD thesis, University of Wisconsin–Madison.

Fitelson, B. (2008). A Decision Procedure for Probability Calculus with Applications. *The Review of Symbolic Logic*, 1(1):111–125.

Fitelson, B. (2015). Earman on Old Evidence and Measures of Confirmation.

Fitelson, B., Easwaran, K., and McCarthy, D. (2016). *Coherence*. Oxford University Press, Oxford.

Fitelson, B. and Hartmann, S. (2016). A New Garber-Style Solution to the Problem of Old Evidence. *Philosophy of Science*.

Fitelson, B. and Hawthorne, J. (2011). How Bayesian confirmation theory handles the paradox of the ravens. In Fetzer, J. H. and Eells, E., editors, *The Place of Probability in Science*, pages 247–275. Springer, New York.

Fitelson, B. and Hitchcock, C. (2011). Probabilistic Measures of Causal Strength. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 600–627. Oxford University Press, Oxford.

Forber, P. (2011). Reconceiving eliminative inference. *Philosophy of Science*, 78:185–208.

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

Forster, M. R. (1999). Model selection in science: The problem of language variance. *The British Journal for the Philosophy of Science*, 50:83–102.

Forster, M. R. (2000). Key Concepts in Model Selection: Performance and Generalizability. *Journal of Mathematical Psychology*, 44(1):205–231.

Forster, M. R. (2002). Predictive Accuracy as an Achievable Goal of Science. *Philosophy of Science*, 69:S124–S134.

Forster, M. R. and Sober, E. (2010). AIC Scores as Evidence—A Bayesian Interpretation. In Forster, M. R. and Bandyopadhyay, P. S., editors, *The Philosophy of Statistics*, pages 535–549. Kluwer, Dordrecht.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, 21(5):1180–1187.

Francis, G., Tanzman, J., and Matthews, W. J. (2014). Excess success for psychology articles in the journal science. *PloS ONE*, 9:e114255.

Friedman, M. (1974). Explanation and Scientific Understanding. *The Journal of Philosophy*, 71(1):5–19.

Frigg, R. (2008). A field guide to recent work on the foundations of statistical mechanics. In Rickles, D., editor, *The Ashgate companion to contemporary philosophy of physics*, pages 99–196. Ashgate, London.

Frigg, R. and Hartmann, S. (2012). Models in science. In *The Stanford Encyclopedia of Philosophy*.

Gaifman, H. and Snir, M. (1982). Probabilities Over Rich Languages, Testing and Randomness. *The Journal of Symbolic Logic*, 47(3):495–548.

Galak, J., LeBoeuf, R. A., Nelson, L. D., and Simmons, J. P. (2012). Correcting the Past: Failures to Replicate Psi. *SSRN Electronic Journal*.

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

Gandenberger, G. (2015). A new proof of the likelihood principle. *British Journal for the Philosophy of Science*, 66:475–503.

Garber, D. (1983). Old Evidence and Logical Omniscience in Bayesian Confirmation Theory. In Earman, J., editor, *Testing Scientific Theories*, pages 99–132. University of Minnesota Press, Minneapolis.

Gelman, A. and Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics in the social sciences. In Kincaid, H., editor, *Oxford Handbook of the Philosophy of the Social Sciences*. Oxford University Press, Oxford.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38.

Gemes, K. (1993). Hypothetico-Deductivism, Content and the Natural Axiomatisation of Theories. *Philosophy of Science*, 60:477–487.

Gemes, K. (1998). Hypothetico-deductivism: the current state of play; the criterion of empirical significance: endgame. *Erkenntnis*, 49(1):1–20.

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press, Princeton.

Glymour, C. (2015). Probability and the explanatory virtues. *The British Journal for the Philosophy of Science*, 66:591–604.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14:107–114.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society. Series B*, 22:319–331.

Good, I. J. (1961a). A Causal Calculus (I). *British Journal for the Philosophy of Science*, 11(44):305–318.

Good, I. J. (1961b). A Causal Calculus (II). *British Journal for the Philosophy of Science*, 12(45):43–51.

Good, I. J. (1967). The white shoe is a red herring. *The British Journal for the Philosophy of Science*, 17(4):322.

Good, I. J. (1968a). Corrigendum: Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(1):203.

Good, I. J. (1968b). Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor. *British Journal for the Philosophy of Science*, 19(2):123–143.

Good, I. J. (1971). 46656 Varieties of Bayesianism. *American Statistician*, 25:62–63.

Good, I. J. (1975). Explicativity, Corroboration, and the Relative Odds of Hypotheses. *Synthese*, 30:39–73.

Good, I. J. (2009). *Good Thinking*. Dover, Mineola, NY.

Goodman, S. (1999a). Toward Evidence-Based Medical Statistics. 1: The P value Fallacy. *Annals of Internal Medicine*, 130:995–1004.

Goodman, S. (1999b). Toward Evidence-Based Medical Statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130:1005–1013.

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

Goodman, S., Berry, D., and Wittes, J. (2010). Bias and Trials Stopped Early for Benefit. *Journal of the American Medical Association*, 304:157.

Gould, S. J. and Lewontin, R. C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205:581–598.

Griffiths, P., Pocheville, A., Calcott, B., Stotz, K., Kim, H., and Knight, R. (2015). Measuring Causal Specificity. *Philosophy of Science*, 82(4):529–555.

Gyenis, Z., Hofer-Szabó, G., and Rédei, M. (2016). Conditioning using conditional expectations: The Borel-Kolmogorov Paradox. *Synthese*.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge University Press, Cambridge.

Hacking, I. (1983). *Representing and Intervening*. Cambridge University Press, Cambridge.

Hahn, U. and Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review*, 114:704–732.

Hailperin, T. (1984). Probability logic. *Notre Dame Journal of Formal Logic*, 25:198–212.

Hailperin, T. (1996). *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. Lehigh University Press, Bethlehem/PA.

Hájek, A. (2003). What Conditional Probability Could Not Be. *Synthese*, 137:273–323.

Hájek, A. (2008). Arguments For—Or Against—Probabilism? *The British Journal for the Philosophy of Science*, 59:793–819.

Hájek, A. (2011). Interpretations of Probability. In *The Stanford Encyclopedia of Philosophy*.

Hájek, A. and Hartmann, S. (2010). Bayesian Epistemology. In Dancy, J., editor, *A Companion to Epistemology*, pages 93–106. Blackwell.

Halpern, J. Y. and Hitchcock, C. (2016). Graded causation and defaults. *The British Journal for the Philosophy of Science*.

Halpern, J. Y. and Pearl, J. (2005a). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4):843–887.

Halpern, J. Y. and Pearl, J. (2005b). Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *British Journal for the Philosophy of Science*, 56(4):889–911.

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors. *Journal of the American Statistical Association*, 96:1122–1132.

Harding, S. (1991). *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Cornell University Press, Ithaca.

Harman, G. H. (1965). The Inference to the Best Explanation. *Philosophical Review*, 74:88–95.

Hartmann, S. (1999). Models and Stories in Hadron Physics. In Morgan, M. and Morrison, M., editors, *Models as Mediators: Perspectives on Natural and Social Science*, pages 326–346. Cambridge University Press, Cambridge.

Hartmann, S. and Rafiee Rad, S. (2016). Updating on Conditionals. *Mind*.

Hartmann, S. and Sprenger, J. (2010). Bayesian Epistemology. In Pritchard, D., editor, *Routledge Companion to Epistemology*, pages 609–620. Routledge, London.

Hartmann, S. and Sprenger, J. (2012). The future of philosophy of science: Introduction. *European Journal for Philosophy of Science*, 2(2):157–159.

Hawthorne, J. (2005). Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions. *Mind*, 114:277–320.

Hawthorne, J. and Fitelson, B. (2004). Re-Solving Irrelevant Conjunction with Probabilistic Independence. *Philosophy of Science*, 71:505–514.

Heckerman, D. (1988). An Axiomatic Framework for Belief Updates. In J.F. Lemmer and L.N. Kanal, editor, *Uncertainty in Artificial Intelligence 2*, pages 11–22, Amsterdam. North-Holland.

Heesen, R. (2016a). Communism and the incentive to share in science.

Heesen, R. (2016b). When journal editors play favorites.

Hempel, C. G. (1960). Inductive inconsistencies. *Synthese*, 12(4):439–469.

Hempel, C. G. (1965). Aspects of Scientific Explanation. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, pages 331–496. Free Press, New York.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15(2):135–175.

Herfeld, C. and Doehne, M. (2016). The diffusion of scientific innovations: A role typology.

Hitchcock, C. and Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, 106(11):587–612.

Hitchcock, C. and Sober, E. (2004). Prediction versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science*, 55(1):1–34.

Hobbs, J. R., Stickel, M., Martin, P., and Edwards, D. (1988). Interpretation as abduction. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pages 95–103. Association for Computational Linguistics.

Howson, C. (1984). Bayesianism and support by novel facts. *British Journal for the Philosophy of Science*, 35:245–251.

Howson, C. (1985). Some Recent Objections to the Bayesian Theory of Support. *British Journal for the Philosophy of Science*, 36:305–309.

Howson, C. (1991). The 'Old Evidence' Problem. *British Journal for the Philosophy of Science*, 42:547–555.

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

Howson, C. (2008). De Finetti, Countable Additivity, Consistency and Coherence. *The British Journal for the Philosophy of Science*, 59:1–23.

Howson, C. (2013). Exhuming the No-Miracles Argument. *Analysis*, 73:205–211.

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

Huber, F. (2005). What is the point of confirmation? *Philosophy of Science*, 72(5):1146–1159.

Hume, D. (1739). *A Treatise of Human Nature*. Clarendon Press, Oxford.

Ioannidis, J. P. and Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3):245–253.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2:e124.

Jaynes, E. T. (1968). Prior Probabilities. In *IEEE Transactions on Systems Science and Cybernetics (SSC-4)*, pages 227–241.

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

Jeffrey, R. (1971). *The logic of decision*. University of Chicago Press, Chicago.

Jeffrey, R. C. (1983). Bayesianism with a Human Face. In Earman, J., editor, *Testing scientific theories*, pages 133–156. University of Minnesota Press, Minneapolis, minnesota edition.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, 3rd editio edition.

Jensma, F. (2014). Marokkaanse afkomst heeft met criminaliteit niets van doen. NRC Handelsblad.

Joyce, J. (2008). Bayes' Theorem. In *The Stanford Encyclopedia of Philosophy*.

Joyce, J. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Huber, F. and Schmidt-Petri, C., editors, *Degrees of Belief*. Springer, Berlin.

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, 65:575–603.

Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1999). *Rethinking the Foundations of Statistics*. Cambridge University Press, Cambridge.

Kaiserman, A. (2016a). Causal Contribution. *Proceedings of the Aristotelian Society*.

Kaiserman, A. (2016b). Partial Liability.

Karni, E. (2005). Subjective Expected Utility Theory without States of the World.

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

Kelly, K. (1996). *The Logic of Reliable Inquiry*. Oxford Un, Oxford.

Kemeny, J. G. (1955). Fair Bets and Inductive Probability. *Journal of Symbolic Logic*, 20:263–273.

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

Kieseppä, I. A. (1997). Akaike Information Criterion, Curve-Fitting, and the Philosophical Problem of Simplicity. *The British Journal for the Philosophy of Science*, 48(1):21–48.

King, N. B., Harper, S., and Young, M. E. (2012). Use of relative and absolute effect measures in reporting health inequalities: structured review. *British Medical Journal*, 345:5774.

Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science*, 48(4):507–531.

Knobe, J. and Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral psychology*, 2:441–448.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin.

Korb, K. B., Nyberg, E. P., and Hope, L. (2011). A New Causal Power Theory. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 628–652. Oxford University Press, Oxford.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Kuhn, T. S. (1977a). Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension*, pages 320–339. University of Chicago Press, Chicago.

Kuhn, T. S. (1977b). *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago University Press, Chicago.

Kyburg, H. E. (1961). *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown/CT.

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

Ladyman, J. and Ross, D. (2009). *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press, Oxford.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48:19–48.

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

Leitgeb, H. (2013). Scientific Philosophy, Mathematical Philosophy, and All That. *Metaphilosophy*, 44:267–275.

Leitgeb, H. (2014). The Stability Theory of Belief. *Philosophical Review*, 123:131–171.

Leitgeb, H. and Pettigrew, R. (2010a). An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science*, 77:201–235.

Leitgeb, H. and Pettigrew, R. (2010b). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*, 77:236–272.

Levi, I. (1963). Corroboration and Rules of Acceptance. *The British Journal for the Philosophy of Science1*, 13:307–313.

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70:556–567.

Lewis, D. (1976). Probabilities of Conditionals and Conditional Probabilities. *Philosophical Review*, 85:297–315.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Noûs*, 13:455–476.

Lewis, D. (1986). *Philosophical Papers, Volume 2*. Oxford University Press, Oxford.

Lewis, D. (1999). *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge.

Lipton, P. (2001). Is Explanation a Guide to Inference? A Reply to Wesley C. Salmon. In Hon, G. and Rakover, S. S., editors, *Explanation: Theoretical Approaches and Applications*, pages 93–120. Kluwer Academic, Dordrecht.

Lipton, P. (2004). *Inference to the Best Explanation*. Routledge, New York, 2nd edition.

Lloyd, E. A. (2005). *The Case of the Female Orgasm: Bias in the Science of Evolution*. Harvard University Press., Cambridge, MA.

Lombrozo, T. (2006). The Structure and Function of Explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

Lombrozo, T. (2007). Simplicity and Probability in Causal Explanation. *Cognitive Psychology*, 55:232–257.

Lombrozo, T. (2009). Explanation and Categorization: How 'Why?' Informs 'What?'. *Cognition*, 110:248–253.

Lombrozo, T. (2011). The Instrumental Value of Explanations. *Philosophy Compass*, 6(8):539–551.

Lombrozo, T. (2012). Explanation and Abductive Inference. In *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, Oxford.

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

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, 67(1):1–25.

Mackie, J. L. (1974). *The Cement of the Universe: a study in Causation*. Clarendon Press, Oxford.

Maddy, P. (2009). *Second Philosophy: A Naturalistic Method*. Oxford University Press, Oxford.

Magnani, L. (2001). *Abduction, reason and science*. Springer, New York.

Magnus, P. and Callender, C. (2004). Realist Ennui and the Base Rate Fallacy. *Philosophy of Science*, 71:320–338.

Maher, P. (2002). Joyce's Argument for Probabilism. *Philosophy of Science*, 69:73–81.

Maher, P. (2010). What is Probability? Unfinished book manuscript.

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

Makinson, D. (1985). How to give it up: A survey of some formal aspects of the logic of theory change. *Synthese*, 62(3):347–363.

Martini, C. and Sprenger, J. (2016). Opinion Aggregation and Individual Expertise. In Boyer-Kassem, T., Mayo-Wilson, C., and Weisberg, M., editors, *Scientific Collaboration and Collective Knowledge*. Oxford University Press, New York.

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

Mayo, D. G. (2010). An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle. In Mayo, D. G. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, chapter 3, pages 305–314. Cambridge University Press, Cambridge.

Mayo, D. G. and Kruse, M. (2001). Principles of Inference and their Consequences. In *Foundations of Bayesianism*. Kluwer Academic Publishers, Netherlands.

McGrew, T. (2003). Confirmation, Heuristics, and Explanatory Reasoning. *British Journal for the Philosophy of Science*, 54:553–567.

McMullin, E. (1982). Values in Science. In *Proceedings of the Biennal Meeting of the PSA*, pages 3–28.

McMullin, E. (2008). The Virtues of a Good Theory. In Curd, M. and Psillos, S., editors, *Routledge Companion to Philosophy of Science*, pages 499–508. Routledge, London.

Meadows, A. (1974). *Communication in science*. Butterworths, London.

Miller, G. (1998). A review of sexual selection and human evolution: How mate choice shaped human nature. *Evolution and Human Behavior: Ideas, Issues, and Applications*, pages 87–129.

Miller, G. F. (2000). *The mating mind: How sexual choice shaped the evolution of human nature*. Anchor Books, New York.

Milne, P. (1996). log[P(h/eb)/P(h/b)] is the One True Measure of Confirmation. *Philosophy of Science*, 63:21–26.

Monton, B. and Mohler, C. (2012). Constructive Empiricism. In *The Stanford Encyclopedia of Philosophy*.

Montori, V. M., Devereaux, P. J., Adhikari, N. K. J., Burns, K. E. A., Eggert, C. H., Briel, M., Lacchetti, C., Leung, T. W., Darling, E., Bryant, D. M., and Others (2005). Randomized trials stopped early for benefit: A systematic review. *JAMA*, 294(17):2203.

Moretti, L. (2007). Ways in which coherence is confirmation conducive. *Synthese*, 157:309–319.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23:103–123.

Morey, R. D., Rouder, J. N., Verhagen, J., and Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: a comment on Cumming (2014). *Psychological science*, 25(6):1289–1290.

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

Muldoon, R., Lisciandra, C., Bicchieri, C., Hartmann, S., and Sprenger, J. (2014). On the emergence of descriptive norms. *Politics, Philosophy and Economics*, 13:3–22.

Myrvold, W. (2016). On the evidential import of unification. *Philosophy of Science*.

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

Myrvold, W. C. (2015). You Can't Always Get What You Want: Some Considerations Regarding Conditional Probabilities. *Erkenntnis*, 80(3):573–603.

Nagel, E. (1961). *The Structure of Science*. Routledge, London.

Nagel, E. (1979). *Teleology Revisited and Other Essays in the Philosophy of Science*. Columbia University Press, New York.

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

Neyman, J. and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A*, 231:289–337.

Neyman, J. and Pearson, E. S. (1967). *Joint Statistical Papers*. University of California Press, Berkeley/CA.

Nicod, J. (1961). *Le problème logique de l'induction*. Presses Universitaires de France, Paris.

Niiniluoto, I. (1983). Novel Facts and Bayesianism. *British Journal for the Philosophy of Science*, 34(4):375–379.

Niiniluoto, I. (1999). *Critical Scientific Realism*. Oxford University Press, Oxford.

Nolan, D. (1997). Quantitative parsimony. *British Journal for the Philosophy of Science*, 48:329–343.

Norton, J. D. (2003). A Material Theory of Induction. *Philosophy of Science*, 70:647–670.

Norton, J. D. (2011). Challenges to Bayesian Confirmation Theory. In Bandyopadhyay, P. S. and Forster, M. R., editors, *Handbook to the Philosophy of Statistics*, pages 391–439. Elsevier, Amsterdam.

Norton, J. D. (2016). A Demonstration of the Incompleteness of Calculi of Inductive Inference.

Oaksford, M. and Chater, N. (2000). *Bayesian Rationality*. Oxford University Press, Oxford.

Oddie, G. (1986). *Likeness to Truth*. Reidel, Dordrecht.

Okasha, S. (2000). Van Fraassen's Critique of Inference to the Best Explanation. *Studies in the History and Philosophy of Science*, 31(4):691–710.

Okruhlik, K. (2005). Gender Bias in the Biological and Social Sciences. *Canadian Journal of Philosophy*, 20:21–42.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349.

Over, D. (2016). Causation and the probability of causal conditionals. In Waldmann, M., editor, *Oxford Handbook of Causal Reasoning*. Oxford University Press, Oxford.

Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge.

Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainy in Arifical Intelligence*, pages 411–420.

Peirce, C. S. (1931). *The Collected Papers of Charles Sanders Peirce*, volume I-VI. Harvard University Press, Cambridge, Mass.

Pettigrew, R. (2015). Epistemic utility arguments for probabilism. In *The Stanford Encyclopedia of Philosophy*.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press, Oxford.

Popper, K. R. (1954). Degree of confirmation. *The British Journal for the Philosophy of Science*, 5:143–149.

Popper, K. R. (1957). A Second Note on Degree of Confirmation. *The British Journal for the Philosophy of Science*, 7(28):350–353.

Popper, K. R. (1958). A third note on degree of corroboration or confirmation. *British Journal for the Philosophy of Science*, 8(32):294–302.

Popper, K. R. (1963). *Conjectures and Refutations: The growth of scientific knowledge*. Routledge, London.

Popper, K. R. (1979). *Objective Knowledge—An Evolutionary Approach*. Clarendon Press, Oxford.

Popper, K. R. (1983). *Realism and the Aim of Science*. Rowman & Littlefield, Towota, NJ.

Popper, K. R. (2002). *The Logic of Scientific Discovery*. Routledge, London. Reprint of the revised English 1959 edition. Originally published in German in 1934 as "Logik der Forschung".

Popper, K. R. and Miller, D. (1983). A proof of the impossibility of inductive probability. *Nature*, 302(5910):687–688.

Porter, T. (1996). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton.

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

Psillos, S. (2009). *Knowing the structure of nature: Essays on realism and explanation*. Palgrave Macmillan, London.

Putnam, H. (1975). *Mathematics, Matter, and Method*, volume I of *Philosophical Papers*. Cambridge University Press, Cambridge.

Quine, W. V. O. (1969). Epistemology Naturalized. In *Ontological Relativity and other Essays*. Columbia University Press, New York.

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

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

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press, Cambridge/MA.

Rédei, M. and Gyenis, Z. (2016). Measure theoretic analysis of consistency of the principal principle.

Reichenbach, H. (1951). *The Rise of Scientific Philosophy*. University of California Press, Berkeley/CA.

Reichenbach, H. (1956). *The Direction Of Time*. University, Berkeley and Los Angeles.

Reiss, J. and Sprenger, J. (2014). Scientific Objectivity. In *The Stanford Encyclopedia of Philosophy*.

Renyi, A. (1970). *Foundations of Probability*. Holden-Day, San Francisco.

Rips, L. J. (1989). Similarity, typicality, and categorization. In *Similarity and analogical reasoning*, pages 21–59. Cambridge University Press, Cambridge.

Roche, W. (2014). A Note on Confirmation and Matthew Properties. *Logic and Philosophy of Science*, 12:91–101.

Romeijn, J.-W., van de Schoot, R., and Hoijtink, H. (2012). One size does not fit all: derivation of a prior-adapted BIC. In Dieks, D., Gonzales, W., Hartmann, S., Uebel, T., and Weber, M., editors, *Probabilities, Laws and Structures*, pages 87–106. Springer, Berlin.

Romeijn, J. W. and Wenmackers, S. (2016). A New Theory About Old Evidence. *Synthese*.

Romero, F. (2016). Can the Behavioral Sciences Self-Correct? A Socio-Epistemic Assessment. *Studies in the History and Philosophy of Science*.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenkrantz, R. (1981). *Foundations and Applications of Inductive Probability*. Ridgeview Press, Atascadero/CA.

Rowbottom, D. P. (2008). The Big Test of Corroboration. *International Studies in the Philosophy of Science*, 22(3):293–302.

Rowbottom, D. P. (2011). *Popper's Critical Rationalism: A Philosophical Investigation*. Routledge, London.

Rowbottom, D. P. (2012). Popper's Measure of Corroboration and P(h | b). *British Journal for the Philosophy of Science*, 64:739–745.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 30:1–6.

Sakamoto, Y., Ishiguro, M., and Kitigawa, G. (1986). *Akaike Information Criterion Statistics*. Reidel, Dordrecht.

Salmon, W. C. (1971). Statistical Explanation and Statistical Relevance. In Colodny, R., editor, *The Nature and Function of Scientific Theories*, pages 173–231. Pittsburgh University Press, Pittsburgh.

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

Salmon, W. C. (2001). Explanation and Confirmation: A Bayesian Critique of Inference to the Best Explanation. In Hon, G. and Rakover, S. S., editors, *Explanation: Theoretical Approaches and Applications*, pages 61–91. Kluwer Academic, Dordrecht.

Savage, L. J. (1972). *The Foundations of Statistics*. Wiley, New York, 2nd edition. Originally published in 1954.

Schaffner, K. (1976). Reductionism in Biology: Prospects and Problems. In *PSA 1974 Special Edition—Boston Studies in the Philosophy of Science*, volume 32, pages 613–632. Springer, New York.

Schaffner, K. F. (1967). Approaches to Reduction. *Philosophy of Science*, 34:137–147.

Schaffner, K. F. (1969). The Watson-Crick Model and Reductionism. *British Journal for the Philosophy of Science*, 20:325–348.

Schaffner, K. F. (1977). Reduction, reductionism, values, and progress in the biomedical sciences. In Colodny, R., editor, *Logic, Laws and Life*, pages 143–171. Pittsburgh University Press, Pittsburgh.

Schaffner, K. F. (1993). *Discovery and explanation in biology and medicine*. Chicago University Press, Chicago.

Schippers, M. (2016). A representation theorem for absolute confirmation.

Schupbach, J. (2005). On a Bayesian Analysis of the Virtue of Unification. *Philosophy of Science*, 72:597–607.

Schupbach, J. N. (2011a). Comparing Probabilistic Measures of Explanatory Power. *Philosophy of Science*, 78(5):813–829.

Schupbach, J. N. (2011b). *Inference to the Best Explanation, Cleaned Up and Made Respectable*. PhD thesis, University of Pittsburgh.

Schupbach, J. N. (2016). Inference to the Best Explanation, Cleaned Up and Made Respectable. In *Best Explanations: New Essays on Inference to the Best Explanation*. Oxford University Press, Oxford.

Schupbach, J. N. and Sprenger, J. (2011). The Logic of Explanatory Power. *Philosophy of Science*, 78(1):105–127.

Schurz, G. (1991). Relevant deduction. *Erkenntnis*, 35:391–437.

Schwan, B. and Stern, R. (2016). A Causal Understanding of When and When Not to Jeffrey Conditionalize.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6:461–464.

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

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

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

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ.

Shannon, C. (1949). Communication Theory of Secrecy Systems. *Bell Systems Technical Journal*, 28:656–715.

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

Sistrom, C. L. and Garvan, C. W. (2004). Proportions, Odds, and Risk. *Radiology*, 230:12–19.

Skyrms, B. (2000). *Choice & Chance*. Wadsworth, Belmont, CA, 4th edition.

Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford University Press, Oxford.

Sloman, S. A. and Lagnado, D. (2015). Causality in Thought. *Annual Review of Psychology*, 66(1):223–247.

Sober, E. (2002). Instrumentalism, Parsimony, and the Akaike Framework. *Philosophy of Science*, 69:S112—-S123.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press, Cambridge.

Sober, E. (2009). Absence of evidence and evidence of absence: Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies*, 143(1):63–90.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B*, 64:583–639.

Spiegelhalter, D. J. and Smith, A. F. M. (1980). Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society. Series B*, 42:213–220.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. Springer, New York, 2nd edition.

Spohn, W. (2012). *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford University Press, Oxford.

Sprenger, J. (2009). Evidence and Experimental Design in Sequential Trials. *Philosophy of Science*, 76:637–649.

Sprenger, J. (2011). Hypothetico-Deductive Confirmation. *Philosophy Compass*, 6(7):497–508.

Sprenger, J. (2012). The Renegade Subjectivist : José Bernardo's Reference Bayesianism. *Rationality, Markets and Morals*, 3:1–13.

Sprenger, J. (2013a). A Synthesis of Hempelian and Hypothetico-Deductive Confirmation. *Erkenntnis*, 78:727–738.

Sprenger, J. (2013b). Testing a Precise Null Hypothesis: The Case of Lindley's Paradox. *Philosophy of Science*, 80:733–744.

Sprenger, J. (2013c). The Role of Bayesian Philosophy within Bayesian Model Selection. *European Journal for Philosophy of Science*, 2:101–114.

Sprenger, J. (2015). A Novel Solution of the Problem of Old Evidence. *Philosophy of Science*, 82:383–401.

Sprenger, J. (2016a). Conditional Degree of Belief.

Sprenger, J. (2016b). Confirmation and Induction. In Humphreys, P. W., editor, *Handbook of Philosophy of Probability*. Oxford University Press, Oxford.

Sprenger, J. (2016c). Foundations for a Probabilistic Theory of Causal Effect.

Sprenger, J. (2016d). Two impossibility results for Popperian corroboration. *British Journal for the Philosophy of Science*.

Sprenger, J. and Stegenga, J. (2016). Three Arguments for Absolute Outcome Measures. *Philosophy of Science*.

Stalnaker, R. (1968). A Theory of Conditionals. In *Studies in Logical Theory: American Philosophical Quarterly Monograph Series, No. 2*. Blackwell, Oxford.

Stalnaker, R. (1970). Probability and Conditionals. *Philosophy of Science*, 37:64–80.

Stalnaker, R. (1975). Indicative Conditionals. *Philosophia*, 5(3):269–286.

Stanford, K. (2006). *Exceeding our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press, Oxford.

Stegenga, J. (2015). Measuring effectiveness. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 54:62–71.

Stich, S. (1988). Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity. *Synthese*, 74(3):391–413.

Strevens, M. (2009). *Depth*. Harvard University Press, Cambridge, MA.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland, Amsterdam.

Teller, P. (1973). Conditionalization and Observation. *Synthese*, 26:218–258.

Tentori, K., Crupi, V., Bonini, N., and Osherson, D. (2007a). Comparison of Confirmation Measures. *Cognition*, 103:107–119.

Tentori, K., Crupi, V., and Osherson, D. (2007b). Determinants of Confirmation. *Psychonomic Bulletin & Review*, 14(5):877–883.

Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences*, 12:435–502.

Trafimow, D. and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37:1–2.

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

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

van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford University Press, New York.

van Riel, R. and Van Gulick, R. (2014). Scientific Reduction. In *The Stanford Encyclopedia of Philosophy*.

Vassend, O. B. (2016). Goals and the Informativeness of Prior Probabilities.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3):426–432.

Walton, D. (1995). *Arguments from Ignorance*. Penn State University Press, Philadelphia.

Wasserman, L. (2004). *All of Statistics*. Springer, New York.

Waters, C. K. (2007). Causes That Make a Difference. *Journal of Philosophy*, 104(11):551–579.

Weber, M. (1904). Die Objektivität sozialwissenschaftlicher und sozialpolitischer Erkenntnis. In *Gesammelte Aufsätze zur Wissenschaftslehre*, pages 146–214. UTB, Tübingen. Reprint edition, 1988.

Weber, M. (1917). Der Sinn der Wertfreiheit der soziologischen und ökonomischen Wissenschaften. In *Gesammelte Aufsätze zur Wissenschaftslehre*, pages 451–502. UTB, Tübingen. Reprint edition, 1988.

Weber, M. (2006). The Central Dogma as a Thesis of Causal Specificity. *History and Philosophy of the Life Sciences*, 28:595–609.

Weinberg, J. M., Nichols, S., and Stich, S. (2001). Normativity and Epistemic Intuitions. *Philosophical Topics*, 29:429–460.

Weisberg, J. (2009). Varieties of Bayesianism. In Gabbay, D. M., Hartmann, S., and Woods, J., editors, *Handbook of the History of Logic*, volume 10: Induct, pages 477–551. North-Holland, Amsterdam.

Weisberg, M. (2007). Who is a Modeler? *The British Journal for the Philosophy of Science*, 58:207–233.

Weisberg, M. (2012). *Simulations and Similarity: Using Models to Understand the World*. Oxford University Press, Oxford.

Weisberg, M. and Muldoon, R. (2009). Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science*, 76:225–252.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, 16:752–760.

Wetzels, R. and Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, pages 1057–1064.

Whewell, W. (1847). *Philosophy of the Inductive Sciences, Founded Upon Their History*. Parker, London.

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

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford University Press, Oxford.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.

Woodward, J. (2012). Causation and Manipulability. In *The Stanford Encyclopedia of Philosophy*.

Woodward, J. (2014). Scientific Explanation. In *The Stanford Encyclopedia of Philosophy*.

Worrall, J. (1989). Structural Realism: The Best of Both Worlds? *Dialectica*, 43:99–124.

Zhao, J., Crupi, V., Tentori, K., Fitelson, B., and Osherson, D. (2012). Updating: Learning versus Supposing. *Cognition*, 124:373–378.

Zhao, J., Shah, A., and Osherson, D. (2009). On the provenance of judgments of conditional probability. *Cognition*, 113(1):26–36.

Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science*, 74:574–587.

List of Figures

List of Tables