The Bayesian Research Programme in the Methodology of Science, or Lakatos Meets Bayes*

Stephan Hartmann†

September 15, 2024

1 Introduction

In 1965, Imre Lakatos organised a famous International Colloquium in the Philosophy of Science in London. Of the four conference proceedings, the Lakatos–Musgrave volume *Criticism and the Growth of Knowledge* (1970) is perhaps the best known. This volume contains Lakatos's long essay 'Falsification and the Methodology of Scientific Research Programmes,' in which he develops and defends his response to Kuhn's challenge to the rationality of science. The theory that Lakatos develops in this essay is well known and is still taught today in introductory courses in the philosophy of science. It shifts the focus from the assessment of individual scientific theories to the assessment of whole research programmes. Research programmes are sequences of scientific theories, they have a positive and a negative heuristic, and they have a hard core (which should not be touched) and a protective belt (which can be modified without abandoning the whole research programme). Lakatos illustrates his ideas with many examples from the history

*To appear in: R. Frigg, J. Alexander, L. Hudetz, M. Rédei, L. Ross and J. Worrall (eds.), *The Continuing Influence of Imre Lakatos's Philosophy: A Reappraisal of his Philosophy on the Occasion of the Centenary of his Birth.* Berlin: Springer.

†Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany)

of science. In doing so, he provides rational reconstructions of important episodes and thus pursues (what he calls) an 'internal history of science.'[1]

Two years earlier, in 1968, another volume containing the proceedings of that colloquium had been published. This volume, *The Problem of Inductive Logic*, contains Lakatos's essay 'Changes in the Problem of Inductive Logic,' in which Lakatos attempts to show that Carnap's philosophical-mathematical research programme in inductive logic is not progressive but degenerative. Here is what he writes at the beginning of the paper:

> A successful research programme bustles with activity. There are always dozens of puzzles to be solved and technical questions to be answered; even if *some* of these – inevitably – are the programme's own creation. But this self-propelling force of the programme may carry away the research workers and cause them to forget about the problem background. They tend not to ask any more to what degree they have solved the original problem, to what degree they gave up basic positions in order to cope with the internal technical difficulties. Although they may travel away from the original problem with enormous speed, they do not notice it. Problem-shifts of this kind may invest research programmes with a remarkable tenacity in digesting and surviving almost any criticism.
>
> Now problem-shifts are regular bedfellows of problem-solving and especially of research programmes. One frequently solves very different problems from those which one has set out to solve. One may solve a more interesting problem than the original one. In such cases we may talk about a 'progressive problem-shift'. But one may solve some problems less interesting than the original one; indeed, in extreme cases, one may end up with solving (or trying to solve) no other problems but those which one has oneself created while trying to solve the original problem. In such cases we may talk about a '*degenerating problem-shift*'.
>
> I think that it can do only good if one occasionally stops problem-solving, and tries to recapitulate the problem background and assess the problem-shift. (Lakatos 1968, 316–317)

Lakatos then continues by applying these general considerations to Carnap's inductive logic:

[1] Nanay (2010) discusses Lakatos's idiosyncratic use of the terms 'internal' and 'external history of science' and how they relate to each other. See also Schindler (forthcoming).

> In the case of Carnap's vast research programme one may wonder what led him to tone down his original bold idea of an a priori, analytic inductive logic to his present caution about the epistemological nature of his theory; why and how he reduced the original problem of rational degree of belief in hypotheses (principally scientific theories) first to the problem of rational degree of belief in particular sentences, and finally to the problem of the probabilistic consistency ('coherence') of systems of beliefs. (Lakatos 1968, 317)

Lakatos then shows that Carnap's research programme is degenerative. Not so much because its predictions turned out to be false or because it did not predict new facts: a philosophical-mathematical research programme such as this cannot do that. It could, however, help to address new problems in the philosophy of science, but did not succeed in this respect. Also, and perhaps more importantly, Carnap's research programme failed according to Lakatos because it underwent a degenerative problem-shift by dealing with ever more specific internal problems and thereby moving further and further away from its original goals.

Lakatos might well be right in his assessment of Carnap's research programme. But what about its contemporary successor, Bayesianism? Can Lakatos's criticism also be levelled against it? To begin with, it is clear that Lakatos was not a Bayesian. However, at least two of his students – Colin Howson and Peter Urbach – became leading Bayesian philosophers of science, and another – John Worrall – sympathises at least somewhat with Bayesianism, as suggested in Worrall (2000), despite any objections he may have. It is doubtful, however, that the views of his students would have changed Lakatos's opinion in this regard. Nonetheless, I will argue below that Bayesianism, when properly understood, is a fine example of a Lakatosian research programme in the methodology of science. This research programme is progressive and can meet many challenges in an elegant way. It also has the capacity to address new and interesting problems in the methodology of science and it helps us to get answers to the big questions about the rationality and objectivity of science. Accordingly, I believe that Lakatos is wrong, at least with respect to contemporary Bayesianism, when he writes that

> [p]robabilism has never generated a programme of historiographical reconstruction; it has never emerged from grappling – unsuccessfully – with the very problems it created. As an epistemological programme it has been degenerating for a long time; as a historiographical programme it never even started. (Lakatos 1976, 20)

Contemporary Bayesianism is a progressive research programme, but not so much in the historiography of science. There were of course attempts to provide historical reconstructions, but I doubt that Lakatos would have been impressed by the Bayesian solution to the Duhem Problem proposed by John Dorling (1979) and popularised by Howson and Urbach (2006) and others. (Deborah Mayo somewhat pejoratively called it 'Dorling's Homework Problem' in her Lakatos Award–winning 1996 book *Error and the Growth of Experimental Knowledge*.) Rather, I will argue that Bayesianism is a progressive research programme in the methodology of science and that it is not only useful to analyse and reconstruct scientific reasoning, but that it also helps us to assess actual scientific reasoning at the frontier of science.

The rest of this paper is organised as follows: Section 2 provides a brief introduction to standard Bayesianism and a list of three challenges it currently faces. Sections 3 and 4 discuss two of these challenges in more detail and show how they can be addressed within the Bayesian research programme. Section 5 discusses some further challenges and suggests what a Bayesian solution might look like in each case. In each of these cases, I will argue, one can speak of a 'progressive problem-shift'. Section 6 therefore concludes that the Bayesian research programme in the methodology of science is progressive. Nevertheless, it is imperative to continue exploring alternatives to it and to develop criteria for comparing competing research programmes.

2 Standard Bayesianism

Bayesianism is a philosophical theory about the statics and dynamics of (partial) beliefs. Its starting point is the psychological truism that we believe different (contingent) propositions more or less strongly, that is, we assign different degrees of belief (or credences) to them. To make the concept 'degree of belief' more precise, we need (i) a calculus for combining different degrees of belief, (ii) an algorithm for updating degrees of belief, and (iii) a (normative) justification for (i) and (ii). Bayesianism offers just this, providing a framework that can be applied to a variety of problems in philosophy, including epistemology and the philosophy of science.

Let's see how the justification of the static and the dynamic part of Bayesianism works. We begin with the static part. Here Bayesianism identifies degrees of belief with (subjective) probabilities, i.e., the (rational) degrees of belief of an agent at a certain time have to satisfy the axioms of proba-

bility theory (see also Weisberg 2011). But what justifies this identification? Bayesians present two types of arguments:

1. Pragmatic arguments ('Dutch book arguments'): these arguments show that an agent with incoherent degrees of belief (i.e., degrees of belief that do not respect the axioms of probability theory) will lose money in a corresponding betting scenario (see Pettigrew 2020 for details).

2. Epistemic arguments ('Epistemic Utility Theory'): these arguments show that identifying degrees of belief with probabilities makes sure that the *inaccuracy* of an agent's degrees of belief is minimised (see Pettigrew 2016 for a defense of this approach).

Let us now move on to the dynamic part of Bayesianism. Here we consider an agent who entertains the propositions $\mathrm{A}_1, \ldots, \mathrm{A}_n$. To proceed formally, one introduces an algebra $\mathcal{A}$ which comprises the propositional variables $A_1, \ldots, A_n$ with the values $\mathrm{A}_1, \neg\mathrm{A}_1$, etc.[2] over which a prior probability distribution P is defined. The agent then learns a piece of evidence, say, $\mathrm{E} = \mathrm{A}_1$. That is, the agent learns that proposition A_1 is true. This prompts her to switch from the prior probability distribution P to the posterior probability distribution P' which satisfies $P'(\mathrm{E}) = 1$. To make sure that her new degrees of belief are coherent (i.e., consistent with the probability calculus), she applies Bayes' Rule (or the *Principle of Conditionalisation*) to obtain, e.g., the new probability of a proposition A_i:

$$P'(\mathrm{A}_i) \;:=\; P(\mathrm{A}_i \,|\, \mathrm{E}) \;=\; \frac{P(\mathrm{E} \,|\, \mathrm{A}_i) P(\mathrm{A}_i)}{P(\mathrm{E})}.$$

There are pragmatic and epistemic arguments that justify the use of conditionalisation. These arguments are, however, more controversial than in the static case (Pettigrew 2020).

If the evidence is not fully certain and a further condition (the *Rigidity Condition*) holds, then conditionalisation generalises to Jeffrey conditionalisation (Jeffrey 2004). In that case, $P'(\mathrm{A}_i)$ is determined as follows:

$$P'(\mathrm{A}_i) \;:=\; P(\mathrm{A}_i \,|\, \mathrm{E}) \cdot P'(\mathrm{E}) \;+\; P(\mathrm{A}_i \,|\, \neg\mathrm{E}) \cdot P'(\neg\mathrm{E}).$$

[2]We use the convention of displaying propositional variables in italics and their values in roman script.

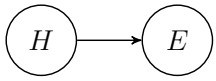

Figure 1: A Bayesian network representing the probabilistic relations between the hypothesis variable H and the evidence variable E.

The most important (though by far not the only) application of Bayesianism in philosophy is confirmation theory, which is concerned with the explication of the notion of 'confirmation': what does it mean that a piece of evidence E confirms a hypothesis H? In a typical scenario, the evidence E is direct, i.e., it is a deductive or inductive consequence of the tested hypothesis H. In this case, the probabilistic relationship between the corresponding propositional variables H and E can be represented by the Bayesian network in Figure 2 (with a probability distribution P defined over it).

Using Bayes' Rule, we can then calculate $P(\mathrm{H})$ and $P'(\mathrm{H}) := P(\mathrm{H} \,|\, \mathrm{E})$ (if the evidence becomes certain). We then say that E confirms H if $P'(\mathrm{H}) > P(\mathrm{H})$, E disconfirms H if $P'(\mathrm{H}) < P(\mathrm{H})$, and E is irrelevant for H if $P'(\mathrm{H}) = P(\mathrm{H})$. Confirmation, thus, means probability-raising.

Using the machinery of Bayesian networks, also more complicated testing scenarios (involving, e.g., various auxiliary hypotheses or partially reliable information sources) can be investigated (see, e.g., Bovens and Hartmann 2003 and Osimani and Landes 2023). Hájek and Hartmann (2010) discuss further epistemological applications of Bayesianism.

Despite these successes, Bayesianism faces a number of foundational problems (see, e.g., Glymour 1980 and Norton 2011; 2021). In my view, many of these problems are just *modelling challenges* (such as the problem of old evidence, which I will discuss below), while others (such as the possible failure to represent ignorance) may point to a better theory beyond Bayesianism.

I'm not too worried about these difficulties. Bayesianism should be treated just like any other scientific theory (and nothing more), and since all scientific theories are facing problems and challenges, it's hard to expect things to be better in philosophy. At the same time, I think that the successful application of a (scientific or philosophical) theory to many cases speaks largely in favour of the theory in question. In short, when evaluating a philosophical theory, we should also consider its pragmatic utility.

Having said this, I will now identify three further challenges to the Bayesian research programme in the methodology of science and suggest

how to address them by extending or modifying the Bayesian research programme.

1. Indirect Evidence

   Standard Bayesianism, as presented so far, assumes that the evidence is direct in the sense that it is a deductive or inductive consequence of the scientific theory under consideration. However, this may not always be the case. Some evidence may be indirect in a sense I will soon make precise. Interesting examples of indirect evidence come from fundamental physics, which is a field where direct empirical evidence is scarce or even non-existent. Here are two examples:

   (a) The no alternatives argument: Does the observation that scientists have not yet found an alternative to string theory (despite a lot of effort and brain power) confirm the theory? Some authors, such as Dawid (2013), think so.

   (b) Analogue simulation: Is it possible to confirm a claim about an empirically unaccessible phenomenon (such as black hole Hawking radiation) by experimenting on a different physical system (e.g., a Bose–Einstein condensate)? Some authors, such as Dardashti, Thébault, and Winsberg (2017), think so.

   Occasionally, indirect evidence has also been called 'non-empirical evidence' (Dawid 2013). This term is somewhat misleading as in both cases an empirical observation is cited. In the case of the no alternatives argument, it is an observation about the respective scientific community (which has not yet found an alternative theory), and in the case of analogue simulation, it is an observation about another physical system. Hence, 'indirect evidence' seems to be the better term.

   In both cases, it would be very helpful to have other means than providing direct empirical evidence to test the respective theories. But is this really possible? Wouldn't it be too good to be true? Clearly, philosophical theories such as hypothetico-deductivism or Popper's falsificationism dismiss this alleged evidence from the outset. However, this inference may be too quick: While it may well turn out that the alleged examples of indirect evidence are not confirmatory, indirect evidence should not be disregarded because one's favourite theory of confirmation (or corroboration) only allows for deductive evidence. Such

theories are not useful for understanding the methodological development of contemporary science. Bayesianism, on the other hand, allows us to analyse confirmation scenarios involving indirect evidence. We will indeed see that indirect confirmation is in principle possible provided that certain conditions hold.

2. New Types of Evidence

   Standard Bayesianism assumes that the evidence is propositional. This is easy to see from an inspection of Bayes' Rule where one has to condition on a proposition representing the evidence. In a learning situation, the probability of this proposition shifts 'by hand' to 1 (in the case of conditionalisation) and the probabilities of all other propositions are in turn updated to make sure that the new probability distribution P' is coherent.

   However, there may also be evidence that does *not* lead to a probability shift of any of the propositions in the algebra: some evidence may be non-propositional. Here are two examples:

   (a) Structural evidence: The agent may learn, e.g., that the underlying causal network of a set of propositions is such and such. This will lead to an update of the probability distribution. But how could it be modelled? One could add meta-propositions to the algebra which make statements about the causal structure, but this does not seem to be very practicable.

   (b) Indicative conditionals: The agent may learn an *indicative conditional* of the form 'If A, then C.' Here the only way to proceed seems to be to condition on the corresponding material conditional, as it can be represented as a Boolean combination of the antecedent and the consequent proposition (and therefore is a proposition itself). But the material conditional is controversial. It is fraught with many problems (but see Williamson 2020 for a recent defense) and, most importantly, it is not at all clear that indicative conditionals are propositions at all (see Douven 2015 for a survey).

   We will show below that Bayesianism has the resources to model such learning experiences.

3. Genuinely New Evidence

   Standard Bayesianism assumes that the learned proposition is already on the agent's 'radar.' It is expected and is given a prior probability. However, this may not always be the case: some evidence may be genuinely new. Let me explain. In many cases, it is not plausible that agents have prior beliefs about each and every piece of evidence they may learn in the future. However, this is expected from a Bayesian agent. One can only update on a proposition which is already in one's algebra and which has a prior probability attached to it. The following examples from various fields of inquiry raise doubts that this is always possible.

   (a) Testimony: Someone told me that there is an excellent new ice cream parlour in my neighbourhood. I update the probability that I get some tasty ice cream today.

   (b) Argumentation: We are debating a policy issue and you make a new argument (based, e.g., on a recent scientific finding) which I didn't anticipate at all.

   (c) Scientific theory change: An old theory runs into problems and a new theory is proposed. This new theory was unexpected and no one assigned a prior to it. One way that has been proposed to deal with this problem is to argue that the new theory is part of the 'catch-all' of the old theory, i.e., is included in the negation of the old theory (Salmon 1990). In this case, however, nothing is known about the new theory, and in particular no prior probability is assigned to this new theory (since there may be many other theories in the 'catch-all' set). Accordingly, this proposal is unsatisfactory.

   These examples show that the standard Bayesian assumption that the algebra of propositions remains fixed is often a strong idealisation. Logical approaches, such as the AGM model of belief revision (Alchourrón, Gärdenfors, and Makinson 1985; Hansson 2022), on the other hand, are not confronted with this problem and can, at least in principle, deal with such cases. This problem for Bayesianism is well known and there is a literature in economics ('awareness') and philosophy (e.g., Bradley 2017) that deals with it (see also Williamson 2003 and de Canson 2024

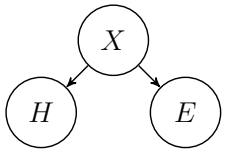

Figure 2: A 'common cause' Bayesian network representing the probabilistic relations between the propositional variables H, E and X.

for a recent discussion). However, this literature still awaits its application to problems from philosophy of science (such as the problem of theory change).

In the next two sections, I will address the first two challenges in turn. I will give a detailed answer to the third problem on another occasion.

3 Challenge 1: Indirect Evidence

The theory of Bayesian networks (Pearl 1988) is well suited to model confirmation scenarios where there is no direct link between the hypothesis variable H and the evidence variable E. For example, the correlation between H and E may be mediated by a 'common cause' variable X, as illustrated in Figure 3.[3]

To apply this idea to a concrete example, one has to find a variable X which (i) plays an *active role* in the reasoning of the agent and which (ii) plausibly screens off H from E. Such variables can indeed be found for the analysis of the no alternatives argument and for the problem of analysing reasoning with analogue simulations. Here is how it works for the no alternatives argument (NAA), which I first present in somewhat more detail than above.

Scientists often argue as follows (P_1 and P_2 are the premises and C is the conclusion of the argument):

P_1: Hypothesis H satisfies several desirable conditions (e.g., it incorporates various scientific principles, it coheres with other established theories, etc.).

[3]For a short introduction to the theory of Bayesian networks, see Hartmann 2021.

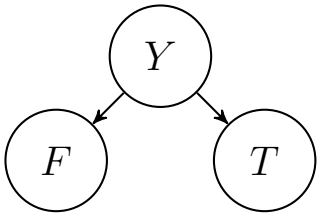

Figure 3: The Bayesian network representing the NAA.

P_2: Despite a lot of effort, the scientific community has not yet found an alternative to H.

C: H is confirmed.

This argument raises at least two questions: (i) How good are NAAs? And (ii) under what conditions, if any, do they work? To address these questions, I propose a Bayesian network model involving the following three variables:

1. The variable T has two values, viz., T: the hypothesis H is true, and ¬T: the hypothesis H is not true.

2. The variable F also has two values, viz., F: the scientific community has not yet found an alternative to H that accounts for the data $\mathcal{D}$ (if there are any) and satisfies the desired constraints $\mathcal{C}$, and ¬F: the scientific community has found an alternative to H that accounts for $\mathcal{D}$ and satisfies $\mathcal{C}$.

3. The variable Y has N values, viz., Y_i: there are exactly i hypotheses which explain $\mathcal{D}$ and fulfil $\mathcal{C}$. (H is one of them.)

Next, we assume that the conditional independencies represented in the Bayesian network in Figure 3 hold. More specifically, we assume that Y screens off T from F, i.e., once the value of Y is known, T and F are independent. I take this to be a plausible assumption.

With this, the following theorem holds. (For details and the proof, see Dawid, Hartmann, and Sprenger 2015.)

Theorem 1. *We set* $P(\mathrm{Y}_i) =: y_i$, $P(\mathrm{F} \mid \mathrm{Y}_i) =: f_i$ *and* $P(\mathrm{T} \mid \mathrm{Y}_i) =: t_i$. *If (a)* f_i *and* t_i *are monotonically decreasing in* i, *(b)* $y_i < 1$ *for all* i *and*

(c) there is at least one pair (i, j) *with* $j > i$ *such that* y_i, $y_j > 0$, $f_i > f_j$ *and* $t_i > t_j$, *then* $P(\mathrm{T} \mid \mathrm{F}) > P(\mathrm{T})$.

It is interesting to note that the NAA works under rather weak and largely plausible assumptions. Every prior probability distribution that satisfies the conditions stated in the theorem will result in the confirmation of T once F is observed. But what about the assumptions? Are they really plausible? Assumption (a) is plausible if we think of the confirmation situation in terms of a sampling scenario. Assumption (b) is perhaps the weakest. It says that the agent is uncertain about the number of alternatives to the theory under consideration. But doesn't the *underdetermination thesis* teach us that there are always infinitely many alternatives to a given theory that imply the given data (if there are any)? In this case, one should set $P(\mathrm{Y}_\infty) = 1$ so that the NAA would work. Clearly, a defender of the NAA has to respond to this worry (see Dawid 2013 for a response). Finally, assumption (c) is related to assumptions (a) and (b), to which it does not add much which could be controversial.

The case of analogue simulation can be analysed in a similar way (Dardashti et al. 2019). In general, the analysis of scenarios involving indirect evidence requires the specification of (i) at least one other 'active' variable (besides H and E) and of (ii) a causal structure which represents the conditional probabilistic independencies that hold amongst the variables. I conclude that Bayesianism (unlike deductive theories of confirmation or corroboration) has the resources to model and investigate scenarios involving indirect evidence. The changes or additions that need to be made are rather insignificant and at best concern the 'protective belt' of the Bayesian research programme.

4 Challenge 2: New Types of Evidence

Let us now explore how the learning of an indicative conditional can be modelled in Bayesianism. To start with, consider the following example (the 'Ski Trip Example' from Douven and Dietz 2011):

> Harry sees his friend Sue buying a skiing outfit. This surprises him a bit, because he did not know of any plans of hers to go on a skiing trip. He knows that she recently had an important exam and thinks it unlikely that she passed. Then he meets Tom, his best friend and also a friend of Sue, who is just on his way to

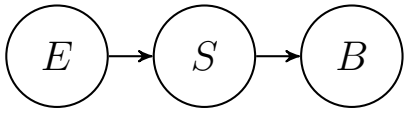

Figure 4: The Bayesian network for the Ski Trip Example.

> Sue to hear whether she passed the exam, and who tells him, 'If Sue passed the exam, then her father will take her on a skiing vacation.' Recalling his earlier observation, Harry now comes to find it more likely that Sue passed the exam.

To model Harry's learning experience, we first note that there are three propositional variables (B, E and S) with the following (positive) values involved here: (i) E: 'Sue passes the exam,' (ii) S: 'Sue is invited on a ski trip', and (iii) B: 'Sue buys a skiing outfit.' We assume that Harry has a prior probability distribution over these three propositional variables and then learns two items of information: (I_1) B and (I_2) 'If E, then S.' Conditionalising on B and the *material conditional* $\mathrm{E} \supset \mathrm{S} \equiv \neg \mathrm{E} \vee \mathrm{S}$, one can show that the probability of E increases under plausible conditions, which is what we – and Harry – expect (Eva, Hartmann, and Rad 2020). This becomes especially clear if one makes the additional assumption that E is probabilistically independent of B given S, or in more formal terms: $E \perp\!\!\!\perp B \,|\, S$. This suggests the 'chain structure' depicted in Figure 4.

So far, so good. However, representing the indicative conditional $\mathrm{A} \rightarrow \mathrm{C}$ by the material conditional $\mathrm{A} \supset \mathrm{C} \equiv \neg \mathrm{A} \vee \mathrm{C}$ has two problems. Firstly, it cannot handle *non-extreme* conditionals, i.e., when there are exceptions and when the conditional is not learnt with certainty. Interestingly, also Jeffrey conditionalising on the material conditional leads to counter-intuitive consequences in these cases. Secondly, it cannot deal with conditionals which are uttered by an only partially reliable information source. This, however, is typically the case and in line with the general idea behind Bayesianism that certainties are hard to find (see Jeffrey 1983).

The second problem is still unsolved (see Collins et al. 2020 for some preliminary ideas). To address the first problem, the *distance-based approach* to Bayesianism can be adapted (Diaconis and Zabell 1982). The idea behind this approach is that it is rational to change one's degrees of belief only minimally once one learns new information. Call this the *Principle of*

Conservativity, which is also used in other accounts of belief revision (such as the AGM model). More specifically, we consider an agent whose degrees of belief are represented by a prior probability distribution P. The agent then learns some new information that poses *probabilistic constraints* on the posterior probability distribution Q. For examples, if the agent learns that the evidence E obtains, then the corresponding constraint is $Q(\mathrm{E}) = 1$.

To work out this proposal, we need to choose a measure for the 'distance' between two probability distributions. For this, the following class of measures turns out to be especially useful:

Definition 1. *f-Divergence (Csiszár 1967). Let* $\mathrm{S}_1, \ldots, \mathrm{S}_n$ *be the possible values of a random variable* S *over which probability distributions* P *and* Q *are defined and let* f *be a convex function with* $f(1) = 0$, $p_i := P(\mathrm{S}_i)$ *and* $q_i := Q(\mathrm{S}_i)$. *Then*

$$D_f(Q \parallel P) \ := \ \sum_{i=1}^{n} p_i \cdot f(q_i/p_i).$$

Many well-known probabilistic divergences are f-divergences. For example, the Kullback–Leibler divergence (KL) obtains for $f(t) = t \log t$. The inverse KL-divergence, the χ^2-divergence and the Hellinger distance follow accordingly. Note that f-divergences are not necessarily symmetrical and that they may violate the triangle inequality. They are therefore not distance functions. And yet, f-divergences are particularly suitable for the present purpose because it can be shown that they yield (Jeffrey) conditionalisation if the agent learns a piece of propositional evidence (Diaconis and Zabell 1982; Eva, Hartmann, and Rad 2020).

Theorem 2. *An agent considers the propositional variables* H *and* E *and has a probability distribution* P *defined over them. She then learns that* $Q(\mathrm{E}) =: e' < 1$. *Minimising an f-divergence between* Q *and* P *taking this constraint into account yields* $Q(\mathrm{H}) = P(\mathrm{H} \,|\, \mathrm{E}) \cdot e' + P(\mathrm{H} \,|\, \neg\mathrm{E}) \cdot (1 - e')$. *This is Jeffrey conditionalisation.*

This is an important result, showing that all f-divergences imply Jeffrey conditionalisation (which I regard as a plausible learning rule) when the agent learns a piece of propositional evidence. (Interestingly, the *Rigidity Condition* is automatically satisfied in this case and does not need to be imposed as an additional constraint.) At the same time, it turns out that all f-divergences are indistinguishable in the case of learning propositional

evidence. It is therefore not necessary to decide in favour of a particular f-divergence.

If one learns the (strict) indicative conditional 'If A, then C' from a perfectly reliable source, then the probabilistic constraint on Q is simply $Q(\mathrm{C} \,|\, \mathrm{A}) = 1$. Nothing more is required. In particular, nothing needs to be said about the propositional status of an indicative conditional. One only needs to specify which probabilistic constraint applies to Q when learning an indicative conditional. Minimising an f-divergence between Q and P taking this constraint into account then yields the same new probability distribution for all f-divergences. The situation is therefore similar to learning a piece of propositional evidence. Interestingly, the new probability distribution is identical with the one which one obtains by conditioning on the corresponding material conditional: $Q = P'$. This is easy to see by noting that $Q(\mathrm{C} \,|\, \mathrm{A}) = 1$ if and only if $Q(\mathrm{A} \supset \mathrm{C}) = 1$ (provided that $Q(\mathrm{A}) > 0$). Hence, the distance-based approach to Bayesianism justifies the use of the material conditional if the learnt indicative conditional is strict and if the information source is perfectly reliable.

Let us now consider non-strict indicative conditionals (from a perfectly reliable information source), which are, as I stated already, much more natural from a Bayesian point of view. In this case the constraint is $Q(\mathrm{C} \,|\, \mathrm{A}) < 1$ and one finds that different f-divergences yield different new probability distributions. We therefore have to 'put our money' on one specific f-divergence if we want to model these cases. But on which? To proceed, we have the following three options: First, one can accept the additional epistemic norm *Minimising Inaccuracy* (as in Epistemic Utility Theory) along with the Principle of Conservativity. Then it can be shown that the inverse KL-divergence is the unique probabilistic divergence (Eva, Hartmann, and Rad 2020). Second, one can try to identify other diachronic norms which (hopefully) restrict the class of admissible divergences. Third, one can explore *empirically* which f-divergence is best. However, the answer to this question may vary with the respective context. In any case, it is still too early to decide which of these options is the right one. And so it is currently best to continue investigating all three options.

As should be clear by now, I do not think that conditionalisation ('Bayes' Rule') or Jeffrey conditionalisation are in the Lakatosian hard core of the Bayesian research programme. The Principle of Conditionalisation often leads to the right results (in particular when the evidence learned is propositional), but it should not be considered one of the central elements of

Bayesianism – at least if we want the scope of Bayesianism to extend beyond the learning of propositional evidence. There are many other types of evidence an agent may learn, and the corresponding updating can often not be modelled as an instance of conditionalisation, as we have seen for non-strict conditionals. Modeling the learning of structural evidence also requires that one use a different updating rule, which makes sure that the new probability distribution satisfies various probabilistic conditional independencies. Besides, even if we learn a proposition, there may be other relevant propositional variables involved in the reasoning situation whose probability assignment we might want to consider fixed across the update. Such additional constraints cannot be taken into account when using conditionalisation.

The distance-based approach, on the other hand, justifies (Jeffrey) conditionalisation (if it can be applied) and is more general and accordingly worthy of further investigation. I therefore suggest that the *Principle of Conservativity* for updating is in the Lakatosian hard core of the Bayesian research programme in the methodology of science. It needs to be spelled out in detail, in a given context, by choosing a specific probabilistic divergence. Which one of these divergences is best will probably depend on the context.

5 Further Challenges

Bayesianism faces a number of further challenges. Here are some of them.

1. The Problem of Old Evidence

   If the agent assigns a prior probability of 1 to the evidence, i.e., if $P(\mathrm{E}) = 1$, then E cannot be learnt (because the probability of E does not change) and it therefore makes no sense to apply an updating rule. Consequently, so-called old evidence (i.e., evidence to which the agent assigns already a prior probability of 1) cannot confirm a hypothesis. This contradicts the practice of science, as Glymour (1980) has pointed out. In response, Bayesians have suggested two ways in which the respective hypothesis can be given a probability increase in scenarios with old evidence:

   (a) Work with a counterfactual probability function that assigns a prior probability of less than 1 to E (e.g., Howson 1991).

   (b) Argue that the agent learns something else than the old evidence E. For example, Garber (1983) suggested that the agent

learns that E is a logical consequence of H, and argued that one should therefore condition on the new proposition X: $H \rightarrow E$.

Glymour (1980) has already anticipated, insightfully discussed, and largely rejected both ways to address the anomaly. The main problem with the first way out is that the proposal seems rather ad hoc and leaves open many questions (such as: how far should we go back in time?). The problems with the second way out are questions regarding the possibility of logical learning and shortcomings of the specific models that have been suggested (see Sprenger and Hartmann 2019 for a discussion). I favour a solution which replaces 'H logically implies E' by X: 'H adequately explains E,' and by introducing another proposition Y: 'The best competitor of H adequately explains E.' One can then formulate a number of plausible conditions under which X confirms H (see Hartmann and Fitelson 2015 and Eva and Hartmann 2020 for details).

Lakatos might have judged that the problem of old evidence is a problem 'one has oneself created while trying to solve the original problem' (to repeat a quote from the beginning of this paper). However, it should be noted that the problem of old evidence is an important one to solve, and the way in which it can actually be solved not only represents an internal progress, but also helps us to better understand how scientists (should) reason.

2. Scientific Theory Change

Earman (1992) and Salmon (1990) (see also Worrall 2000) have discussed Bayesian accounts of Kuhn's influential theory of theory change. They were not entirely successful. This is not least due to the fact that they have not considered all aspects of Kuhn's theory. Farmakis (2008), for example, has noted that they have left out the incommensurability issue. But perhaps a full Bayesian account of Kuhn's theory is not necessary. Kuhn may well be right that there is no 'algorithm' that helps us decide once and for all when a particular theory should be abandoned. Feyerabend also made this point in his response to Lakatos when he wrote, 'if you are permitted to wait, why not wait a little longer?' (1970, 215). And even Lakatos argued that there is no 'instant rationality' and that we can provide a rational and objectivist account

of theory change only in retrospect, when the internal history is available in the form of a rational reconstruction. Nevertheless, I would like to argue that Bayesianism can help us in everyday scientific reasoning and argumentation, e.g., when we reason about whether we should abandon a theory or research programme and look for an alternative instead.

Bayesianism lends itself here because theorising takes place in the realm of uncertainty, and scientists, like all of us in everyday life, have to make decisions all the time. These decisions should be rational, and Bayesian decision theory provides a useful and justifiable framework for achieving this while still allowing for subjective judgements by scientists. For example, an individual scientist may be faced with the decision of whether to maintain and continue researching the current theory. Perhaps this will lead to a major discovery? And perhaps an observed anomaly can be explained after all. (Remember that Lakatos taught us that every research programme evolves in an 'ocean of anomalies.') One does not know with certainty in advance. A reconstruction of the decision situation that makes explicit the different propositions that the scientist considers and how they are related, together with the corresponding (subjective) probability distribution, can help the agent to make better-reasoned decisions. For example, in the case mentioned above, consider how likely the agent thinks it is that a model can be found within the given research programme (or paradigm) that explains the evidence. Perhaps the agent initially assigns a fairly high probability to this proposition, which she then updates in the light of her (possibly unsuccessful) attempts to find such a model. At some point she will give up, and if many other scientists do the same, the theory (or research programme) will eventually be replaced by another. This thought process can be modelled, wherein Bayesianism proves useful without promising more than it can deliver, which is what one should expect from a progressive research programme.

3. Collective Reasoning and Argumentation

   Standard Bayesianism is a philosophical theory in which a single agent is at the centre. This agent maintains a set of propositions that she believes more or less strongly and updates in the light of new evidence according to a particular rule. As we have seen, this simple approach

can be used to analyse a wide range of issues in the philosophy of science.[4] However, it turns out that science happens in a social context, which should be taken into account if Bayesianism is to critically accompany current science. For example, scientists try to convince each other and then update their individual probability distributions by taking into account the information coming from other scientists. Or a committee chair (debating environmental policy measures, for example) may consult scientific experts to make the best decision on the issue based on the experts' probabilistic judgements. There may also be situations where we want to assign a probability distribution to a group, e.g., a scientific community. Something like this could be helpful, for example, if we want to further reconstruct Kuhn's philosophy of science in Bayesian terms. It will be interesting to address these questions and many others in future work. There is no reason why the Bayesian research programme in the philosophy of science should not be further developed in this direction, especially since much work has already been done on which one can build. This again underlines the main point I want to make in this paper, namely that the Bayesian research programme in the philosophy of science is progressive.

6 Conclusion

Bayesianism is a progressive scientific research programme in the methodology of science. It is closely related to other Bayesian research programmes, as Bayesianism is not only flourishing in philosophy, but also in cognitive science ('the new paradigm'), neuroscience ('the Bayesian brain,' 'the free energy principle') and artificial intelligence. Lakatos's philosophy of science is useful in reconstructing these Bayesian research programmes. However, I have argued that it is more plausible to place the Principle of Conservativity at the hard core of the Bayesianism research programme in the methodology of science, rather than conditionalisation ('Bayes' rule') or Jeffrey conditionalisation. I have argued that this principle (if, as suggested, it is specified using f-divergences) justifies (Jeffrey) conditionals and allows updating on

[4] In discussing the NAA, we were dealing with an issue that the scientific community is concerned about. However, we did not model the probability functions of the individual scientists, but considered an external agent who assigns a probability function to the scientific community and updates it accordingly.

the basis of other types of evidence (such as indicative conditionals). The relevant research programme is progressive in that it successfully addresses various anomalies (such as the problem of old evidence) and is able to solve new problems. Many other problems are still open and await a Bayesian treatment.

Despite these successes of the Bayesian research programme in the methodology of science, it is important to also investigate alternative approaches, such as imprecise probabilities (e.g., Augustin et al. 2014) or ranking theory (Spohn 2012), and to develop criteria for how to evaluate and compare the results. For a similar plea in relation to Bayesian cognitive science, see Colombo, Elkin, and Hartmann (2021).

Acknowledgements

I would like to thank Christopher von Bülow and an anonymous reviewer for several suggestions for improvement. Thanks also go to my co-authors Benjamin Eva, Richard Dawid, Soroush Rafiee Rad and Jan Sprenger, with whom some of the results discussed here were found.

I dedicate this essay to the memory of Colin Howson – a much-missed fellow Bayesian and friend (and not least a source of entertaining stories about Lakatos).

References

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson. 1985. "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions." *Journal of Symbolic Logic* 50:510–530.

Augustin, Thomas, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, eds. 2014. *Introduction to Imprecise Probabilities*. Chichester NY: Wiley.

Bovens, Luc, and Stephan Hartmann. 2003. *Bayesian Epistemology*. Oxford: Clarendon Press.

Bradley, Richard. 2017. *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Collins, Peter J., Karolina Krzyżanowska, Stephan Hartmann, Gregory Wheeler, and Ulrike Hahn. 2020. "Conditionals and Testimony." *Cognitive Psychology* 122 (November): 101329.

Colombo, Matteo, Lee Elkin, and Stephan Hartmann. 2021. "Being Realist about Bayes, and the Predictive Processing Theory of Mind." *The British Journal for the Philosophy of Science* 72 (1): 185–220 (March).

Csiszár, Imre. 1967. "Information-Type Measures of Difference of Probability Distributions and Indirect Observation." *Studia Scientiarum Mathematicarum Hungarica* 2:229–318.

Dardashti, Radin, Stephan Hartmann, Karim Thébault, and Eric Winsberg. 2019. "Hawking Radiation and Analogue Experiments: A Bayesian Analysis." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 67 (August): 1–11.

Dardashti, Radin, Karim P. Y. Thébault, and Eric Winsberg. 2017. "Confirmation via Analogue Simulation: What Dumb Holes Could Tell Us about Gravity." *The British Journal for the Philosophy of Science* 68 (1): 55–89 (March).

Dawid, Richard. 2013. *String Theory and the Scientific Method*. Cambridge: Cambridge University Press.

Dawid, Richard, Stephan Hartmann, and Jan Sprenger. 2015. "The No Alternatives Argument." *The British Journal for the Philosophy of Science* 66 (1): 213–234 (March).

de Canson, Chloé. 2024. "The Nature of Awareness Growth." *Philosophical Review* 133 (1): 1–32 (January).

Diaconis, Persi, and Sandy L. Zabell. 1982. "Updating Subjective Probability." *Journal of the American Statistical Association* 77 (380): 822–830.

Dorling, Jon. 1979. "Bayesian Personalism, the Methodology of Scientific Research Programmes, and Duhem's Problem." *Studies in History and Philosophy of Science Part A* 10 (3): 177–187 (September).

Douven, Igor. 2015. *The Epistemology of Indicative Conditionals*. Cambridge: Cambridge University Press.

Douven, Igor, and Richard Dietz. 2011. "A Puzzle about Stalnaker's Hypothesis." *Topoi* 30:31–37.

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

Eva, Benjamin, and Stephan Hartmann. 2020. "On the Origins of Old Evidence." *Australasian Journal of Philosophy* 98 (3): 481–494.

Eva, Benjamin, Stephan Hartmann, and Soroush Rafiee Rad. 2020. "Learning from Conditionals." *Mind* 129 (514): 461–508 (April).

Farmakis, Lefteris. 2008. "Did Tom Kuhn actually Meet Tom Bayes?" *Erkenntnis* 68 (1): 41–53.

Feyerabend, Paul. 1970. "Consolations for the Specialist." In Lakatos and Musgrave 1970, 197–230.

Garber, Daniel. 1983. "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory." In *Testing Scientific Theories*, edited by John Earman, Volume 10 of *Minnesota Studies in the Philosophy of Science*, 99–131. Minneapolis MN: University of Minnesota Press.

Glymour, Clark. 1980. "Why I Am Not a Bayesian." In *Theory and Evidence*, 63–93. Princeton NJ: Princeton University Press.

Hájek, Alan, and Stephan Hartmann. 2010. "Bayesian Epistemology." In *A Companion to Epistemology*, edited by Jonathan Dancy, Ernest Sosa, and Matthias Steup, second edition, 93–106. Oxford: Blackwell.

Hansson, Sven Ove. 2022. "Logic of Belief Revision." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2022 edition.

Hartmann, Stephan. 2021. "Bayes Nets and Rationality." In *The Handbook of Rationality*, edited by Markus Knauff and Wolfgang Spohn, 253–264. Boston MA: MIT Press.

Hartmann, Stephan, and Branden Fitelson. 2015. "A New Garber-Style Solution to the Problem of Old Evidence." *Philosophy of Science* 82 (4): 712–717 (October).

Howson, Colin. 1991. "The 'Old Evidence' Problem." *The British Journal for the Philosophy of Science* 42 (4): 547–555 (December).

Howson, Colin, and Peter Urbach. 2006. *Scientific Reasoning: The Bayesian Approach*. Third edition. La Salle, IL: Open Court.

Jeffrey, Richard. 1983. "Bayesianism with a Human Face." In *Testing Scientific Theories*, edited by John Earman, Volume 10 of *Minnesota Studies*

in the Philosophy of Science, 133–156. Minneapolis MN: University of Minnesota Press.

———. 2004. *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press.

Lakatos, Imre. 1968. "Changes in the Problem of Inductive Logic." In *The Problem of Inductive Logic*, edited by Imre Lakatos, 315–416. Amsterdam: North Holland.

———. 1976. "History of Science and its Rational Reconstructions." In *Method and Appraisal in the Physical Sciences: The Critical Background to Modern Science, 1800–1905*, edited by Colin Howson, 1–39. Cambridge: Cambridge University Press.

Lakatos, Imre, and Alan Musgrave, eds. 1970. *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science*. Cambridge: Cambridge University Press.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago and London: University of Chicago Press.

Nanay, Bence. 2010. "Rational Reconstruction Reconsidered." *Monist* 93 (4): 598–617 (October).

Norton, John D. 2011. "Challenges to Bayesian Confirmation Theory." In *Philosophy of Statistics*, edited by Prasanta S. Bandyopadhyay and Malcolm R. Forster, Volume 7 of *Handbook of the Philosophy of Science*, 391–440. Amsterdam: Elsevier.

———. 2021. *The Material Theory of Induction*. BSPS Open Series. Calgary, Canada: University of Calgary Press.

Osimani, Barbara, and Jürgen Landes. 2023. "Varieties of Error and Varieties of Evidence in Scientific Inference." *The British Journal for the Philosophy of Science* 74 (1): 117–170 (March).

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

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

———. 2020. *Dutch Book Arguments*. Cambridge: Cambridge University Press.

Salmon, Wesley C. 1990. "Rationality and Objectivity in Science, *or* Tom Kuhn Meets Tom Bayes." In *Scientific Theories*, edited by C. Wade Savage, Volume 14 of *Minnesota Studies in the Philosophy of Science*, 175–204. Minneapolis MN: University of Minnesota Press.

Schindler, Samuel. Forthcoming. "Beyond Footnotes: Lakatos's Meta-Philosophy and the History of Science." In *The Continuing Influence of Imre Lakatos's Philosophy: A Reappraisal of his Philosophy on the Occasion of the Centenary of his Birth*, edited by Roman Frigg, J. McKenzie Alexander, Laurenz Hudetz, Miklós Rédei, Lauren N. Ross, and John Worrall. Berlin: Springer.

Spohn, Wolfgang. 2012. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford: Oxford University Press.

Sprenger, Jan, and Stephan Hartmann. 2019. *Bayesian Philosophy of Science*. Oxford: Oxford University Press.

Weisberg, Jonathan. 2011. "Varieties of Bayesianism." In *Inductive Logic*, edited by Dov M. Gabbay, Stephan Hartmann, and John Woods, Volume 10 of *Handbook of the History of Logic*, 477–551. Amsterdam: Elsevier.

Williamson, Jon. 2003. "Bayesianism and Language Change." *Journal of Logic, Language and Information* 12 (1): 53–97 (December).

Williamson, Timothy. 2020. *Suppose and Tell: The Semantics and Heuristics of Conditionals*. Oxford: Oxford University Press.

Worrall, John. 2000. "Kuhn, Bayes and 'Theory-Choice': How Revolutionary is Kuhn's Account of Theoretical Change?" In *After Popper, Kuhn and Feyerabend: Recent Issues in Theories of Scientific Method*, edited by Robert Nola and Howard Sankey, Volume 15 of *Australasian Studies in History and Philosophy of Science*. Dordrecht: Springer.