4

Assessing Scientific Theories

The Bayesian Approach

RADIN DARDASHTI AND STEPHAN HARTMANN

4.1 Introduction

Scientific theories are used for a variety of purposes. For example, physical theories such as classical mechanics and electrodynamics have important applications in engineering and technology, and we trust that this results in useful machines, stable bridges, and the like. Similarly, theories such as quantum mechanics and relativity theory have many applications as well. Beyond that, these theories provide us with an understanding of the world and address fundamental questions about space, time, and matter. Here we trust that the answers scientific theories give are reliable and that we have good reason to believe that the features of the world are similar to what the theories say about them. But why do we trust scientific theories, and what counts as evidence in favor of them?

Clearly, the theories in question have to successfully relate to the outside world. But how, exactly, can they do this? The traditional answer to this question is that established scientific theories have been positively tested in experiments. In the simplest case, scientists derive a prediction from the theory under consideration, which is then found to hold in a direct observation. Actual experimental tests are, of course, much more intricate, and the evaluation and interpretation of the data is a subtle and by no means trivial matter. One only needs to take a look at the experiments at the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) near Geneva, and at the huge amount of workforce and data analysis involved there, to realize how nontrivial actual experiments can be. However, there is no controversy among scientists and philosophers of science about the proposition that the conduct and analysis of experiments is the (only?) right way to assess theories. We trust theories because we trust the experimental procedures to test them. Philosophers of science have worked this idea out and formulated theories of confirmation (or corroboration) that make the corresponding intuition more precise. We discuss some of these proposals in the next two sections.

While empirical testing is certainly important in science, some theories in fundamental science cannot (yet?) be tested empirically. String theory is a case in point. This arguably most fundamental scientific theory makes, so far, no empirically testable predictions, and even if it would make any distinct predictions, referring to predictions that are not also predictions of the theories it unifies (i.e., the Standard Model and general relativity), these predictions could not be confirmed in a laboratory because the required energies are too high. The question, then, is what we should conclude from this. Are fundamental theories such as string theory not really scientific theories? Are they only mathematical theories? Many people would refrain from this conclusion and argue that string theory does, indeed, tell us something about our world. But why should we believe this? Are there other ways of assessing scientific theories, ways that go beyond empirical testing?

Several well-known physicists think so (see, for example, Polchinski, this volume) and in a recent book, the philosopher Richard Dawid (2013) defends the view that there are non-empirical ways of assessing scientific theories. In his book, he gives a number of examples, including the so-called No Alternatives Argument, which we discuss in detail in Section 4.4. Another example is analogue experiments, examined in Section 4.5. There is, however, no agreement among scientists regarding the viability of these nonstandard ways of assessing scientific theories. In a similar vein, these new methods do not fit into traditional philosophical accounts of confirmation (or corroboration) such as Hempel's hypothetico-deductive (HD) model or Popper's falsificationism. Interestingly, these deductivist accounts exclude indirect ways of assessing scientific theories from the beginning as, in those accounts, the evidence for a theory under consideration must be a deductive consequence of it and it must be observed.

There is, however, an alternative philosophical framework available: Bayesian confirmation theory. We will show that it can be fruitfully employed to analyze potential cases of indirect confirmation such as the ones mentioned previously. This will allow us to investigate under which conditions indirect confirmation works, and it will indicate which, if any, holes in a chain of reasoning have to be closed if one wants to make a confirmation claim based on indirect evidence. By doing so, Bayesian confirmation theory helps the scientist (as well as the philosopher of science) better understand how, and under which conditions, fundamental scientific theories such as string theory can be assessed and, if successful, trusted.

The remainder of this chapter is organized as follows. Section 4.2 considers the traditional accounts of assessing scientific theories mentioned earlier and shows that they are inadequate to scrutinize indirect ways of assessing scientific theories. Section 4.3 introduces Bayesian confirmation theory and illustrates the basic mechanism of indirect confirmation. The following two sections present a detailed

Bayesian account of two examples, namely the No Alternatives Argument (Section 4.4) and analogue experiments (Section 4.5). Section 4.6 provides a critical discussion of the vices and virtues of the Bayesian approach to indirect theory assessment. Finally, Section 4.7 concludes with a summary of our main results.

4.2 Trusting Theories

Why do we trust a scientific theory? One reason for trusting a scientific theory is certainly that it accounts for all data available at the time in its domain of applicability. But this is not enough: We also expect a scientific theory to account for all future data in its domain of applicability. What, if anything, grounds the corresponding belief? What grounds the inference from the success of a theory for a finite set of data to the success of the theory for a larger set of data? After all, the future data are, by definition, not available yet, but we would nevertheless like to work with theories which we trust to be successful in the future.

Recalling Hume's problem of induction (Howson, 2000), Popper argues that there is no ground for this belief at all. Such inferences are not justified, so all we can expect from science is that it identifies those theories that do not work. We can only falsify a proposed theory if it contradicts the available data. Popper's theory is called *falsificationism* and it comes in a variety of versions. According to *naive falsificationism*, a theory T is corroborated if an empirically testable prediction E of T (i.e., a deductive consequence of T) holds. Note that this is only a statement about the theory's past performance, with no implications for its future performance. If the empirically testable prediction does not hold, then the theory is falsified and should be rejected and replaced by an alternative theory. As more sophisticated versions of falsificationism have the same main problem relevant for the present discussion as naive falsificationism, we do not have to discuss them here (see Pigliucci or Carroll, this volume). All that matters for us is the observation that, according to falsificationism, a theory can be corroborated only empirically. Hence, a falsificationist cannot make sense out of indirect ways of assessing scientific theories. These possible new ways of arguing for a scientific theory have to be dismissed from the beginning because they do not fit the proposed falsificationist methodology. We think that this is not a good reason to reject the new methods. It may well turn out that we come to the conclusion that none of them work, but the reason for this conclusion should not be that the new method does not fit our favorite account of theory assessment.

Hempel, an inductivist, argued that we have grounds to believe that a well-confirmed theory can also be trusted in the future. Here is a concise summary of the main idea of the hypothetico-deductive (HD) model of confirmation that Hempel famously defended:

> General hypotheses in science as well as in everyday use are intended to enable us to anticipate future events; hence, it seems reasonable to count any prediction that is borne out by subsequent observation as confirming evidence for the hypothesis on which it is based, and any prediction that fails as disconfirming evidence. (Hempel, 1945, p. 97)

Note that the HD model shares an important feature with Popper's falsificationism: Both are deductivist accounts; that is, the evidence has to be a deductive consequence of the tested theory. Thus indirect ways of confirmation also do not fit the HD model and must be dismissed for the same reason falsificationism has to dismiss them.

The HD model has a number of (other) well-known problems, which eventually led to its rejection in the philosophical literature (see, however, Schurz, 1991 and Sprenger, 2011). The first problem is the *tacking problem*: If E confirms T, then it also confirms $T \wedge X$. Note that X can be a completely irrelevant proposition (such as "pink dragons like raspberry marmalade"). This is counterintuitive, as we do not expect E to confirm the conjunction $T \wedge X$, but only T. The second problem has to do with the fact that the HD model cannot account for the intuition that some evidence may confirm a theory more than some other evidence. The HD model lacks an account of degrees of confirmation, but can justify only the qualitative inference that E confirms T (or not). Bayesian confirmation theory, discussed in the next section, accounts for these and other problems of more traditional accounts of confirmation. It also has the resources and the flexibility to model indirect ways of assessing scientific theories.

4.3 Bayesian Confirmation Theory

In this section, we give a brief introduction to Bayesian confirmation theory (BCT). For book-length introductions and discussions of the topic, we refer the reader to Earman (1992), Howson and Urbach (2006), and Sprenger and Hartmann (2019). For recent surveys of the field of Bayesian epistemology, see Hájek and Hartmann (2010) and Hartmann and Sprenger (2010).

Let us consider an agent (e.g., a scientist or the whole scientific community) who entertains the propositions T: "The theory under consideration is empirically adequate" and E: "The respective evidence holds" before a test of the theory is performed.[1] In this case the agent is uncertain as to whether the theory is empirically

[1] "Empirical adequacy" is a technical term made popular in the philosophical literature by Bas van Fraassen. In his book *The Scientific Image*, he writes that "a theory is empirically adequate exactly if what it says about the observable things and events in the world is true – exactly if it 'saves the phenomena'" (van Fraassen, 1980, p. 12). Note that empirical adequacy is logically weaker than truth. A true theory is empirically adequate, but not every empirically adequate theory is true. We could make our arguments also using the term "truth" instead of "empirical adequacy," but decided to stick to the weaker notion in the following discussion.

adequate; she also does not know that the evidence will hold. The easiest way to represent her attitude toward these two propositions is to assign a probability to them. Bayesians request that rational agents assign a prior probability distribution P over the propositional variables they consider. In our case the agent considers the binary propositional variables T (with the values T and ¬T) and E (with the values E and ¬E).[2]

Next we assume that a test is performed and that the evidence holds. As a result, the probability of E shifts from $P(\mathrm{E}) < 1$ to $P^*(\mathrm{E}) = 1$, where P^* denotes the new "posterior" probability distribution of the agent after learning the evidence. To make sure that P^* is coherent, meaning that it satisfies the axioms of probability theory, the agent has to adjust the other entries in the posterior probability distribution. But how can this be done in a rational way? For the situation just described, Bayesians argue that the posterior probability of T should be the old conditional probability:

$$P^*(\mathrm{T}) = P(\mathrm{T}|\mathrm{E}). \tag{4.1}$$

This identification, which is sometimes called "Bayes' theorem" or "conditionalization," can be justified in various ways such as via Dutch Book arguments (Vineberg, 1997), epistemic utility theory (Pettigrew, 2016) and distance-minimization methods (Diaconis & Zabell, 1982; Eva & Hartmann, 2018), which we do not consider here. Once we accept Bayes' theorem as a diachronic norm, the right-hand side of Eq. (4.1) can be expressed differently using the definition of the conditional probability. As $P(\mathrm{T}|\mathrm{E})P(\mathrm{E}) = P(\mathrm{T},\mathrm{E}) = P(\mathrm{E},\mathrm{T}) = P(\mathrm{E}|\mathrm{T})P(\mathrm{T})$, we obtain

$$P^*(\mathrm{T}) = \frac{P(\mathrm{E}|\mathrm{T})\,P(\mathrm{T})}{P(\mathrm{E})}. \tag{4.2}$$

This equation expresses the posterior probability of T, $P^*(\mathrm{T})$, in terms of the *prior probability* of T, $P(\mathrm{T})$; the *likelihood* of the evidence, $P(\mathrm{E}|\mathrm{T})$; and the *expectancy* of the evidence, $P(\mathrm{E})$.

According to BCT, E *confirms* T if and only if $P^*(\mathrm{T}) > P(\mathrm{T})$, so that the observation of E raises the probability of T. Likewise, E *disconfirms* T if and only if $P^*(\mathrm{T}) < P(\mathrm{T})$, so that the observation of E lowers the probability of T. The evidence E is *irrelevant* for T if it does not change its probability, so that $P^*(\mathrm{T}) = P(\mathrm{T})$.

Note that in the standard account of BCT, the prior probability distribution P (and therefore also the posterior probability distribution P^*) is a *subjective* probability distribution, such that different agents may disagree on it. It may therefore

[2] We follow the notation of Bovens and Hartmann (2004) and denote propositional variables in italic script and their values in roman script. Note further that we sometimes use the letter "E" (in roman script) to refer to the evidence directly and not to the proposition it expresses, and likewise for the theory T. We submit that this does not cause any confusion.

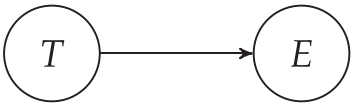

Figure 4.1 A Bayesian network representing the direct dependence between the variables E and T.

happen that one agent considers E to confirm T while another agent considers E to disconfirm T (or to be irrelevant for T). That is, in BCT, confirmation is defined only relative to a probability distribution, which implies that confirmation is not an objective notion in BCT.

Using BCT, especially Eq. (4.2), has a number of desirable features. Consider, for example, the situation where the evidence is a deductive consequence of the theory in question. In this case, $P(\mathrm{T}|\mathrm{E}) = 1$ and therefore $P^*(\mathrm{T}) = P(\mathrm{T})/P(\mathrm{E})$. As $P(\mathrm{E})$ expresses the expectancy of the evidence before a test is performed, a rational agent will assign a value $P(\mathrm{E}) < 1$. Hence, $P^*(\mathrm{T}) > P(\mathrm{T})$ and therefore E confirms T. This is in line with our expectations, and it is in line with the hypothetic-deductive (HD) model of confirmation mentioned in the last section. Note further that the more E confirms T, the lower the expectancy of E is. Again, this is in line with our intuition: More surprising evidence confirms a theory better than less surprising (or expected) evidence. As BCT is a quantitative theory, it allows us to account for this intuition, whereas qualitative theories such as the HD model and Popper's falsificationism do not have the resources to do so. They answer only the yes–no question of whether E confirms (or corroborates) T.

Let us develop BCT a bit further. In many cases, there is a *direct dependency* between T and E. We mentioned already the case where E is a deductive consequence of T. In other cases, there is a direct probabilistic dependency between E and T (because of the presence of various uncontrollable disturbing factors). In these cases, $P(\mathrm{T}|\mathrm{E}) < 1$, but $P^*(\mathrm{T})$ may, of course, still be greater than $P(\mathrm{T})$. The direct dependence between theory and evidence is depicted in the Bayesian network in Figure 4.1. Here the nodes E and T represent the respective propositional variables and the arc that connects them indicates the direct dependency.[3]

Note, however, that BCT can also deal with cases where the evidence is indirect – that is, when the evidence is not a deductive or inductive consequence of the theory in question. In these cases, the correlation between E and T is accounted for by a third ("common cause") variable X as depicted in the Bayesian network in Figure 4.2. Here is an illustration: We take it that having yellow fingers (E) is evidence of having heart disease (T). However, there is no direct dependence between the two

[3] For an introduction to the theory of Bayesian networks and their use in epistemology and philosophy of science, see Bovens and Hartmann (2004, Section 3.5).

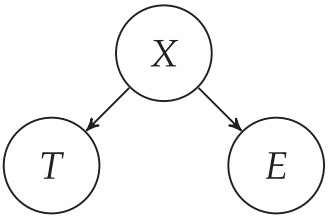

Figure 4.2 A Bayesian network representing the (indirect) dependence between the variables E and T, mediated by a common cause variable X.

variables E and T: For example, painting your fingers green will (unfortunately) not lower your probability of having heart disease.[4] Instead, the positive correlation between E and T and the fact that the observation of E confirms T result from the presence of a common cause X, the possible tobacco use (smoking) of the respective agent. Note that X and T and X and E are positively correlated, such that observing E confirms X, which in turn confirms T. Hence, for the Bayesian network depicted in Figure 4.2, E confirms T.[5]

To elaborate a bit further on this, the common cause X (of E and T) has the following property: If the value of X is not known, then E and T are dependent. But once the value of X is known (i.e., once the variable X is instantiated), E and T become independent. One also says that X *screens off* E from T.

We will see later that the common cause structure can be used to model cases of indirect confirmation. Here, similar to the previous example, the evidence probes a third variable, which in turn probes the theory in question. The powerful formal machinery of Bayesian networks (and conditional independence structures) makes sure that all of this can be made precise. It is important to note that deductivist accounts of confirmation lack something analogous to a common cause structure, which is their major drawback.

In the next two sections, we show in detail how these ideas can be applied to analyze cases of indirect confirmation.

4.4 Illustration 1: The No Alternatives Argument

Scientists have been defending and developing string theory for several decades, despite the lack of direct empirical confirmation. What is more, no one expects

[4] For more on this, see Woodward (2005).

[5] Note that the confirmation relation is symmetric: E confirms T if and only if T confirms E. Note further that the confirmation relation is not necessarily transitive. Consider, for example, the situation where, in addition to the arcs in the Bayesian network in Figure 4.2, there is an arc from T to E (or vice versa) that represents a sufficiently strong negative correlation between E and T. This negative correlation can then outweigh the positive correlation from the path via X.

this situation to change in the foreseeable future. This raises a question: Why do scientists put so much trust in a theory that is not (and perhaps cannot be) assessed by empirical data? What grounds this enormous trust in string theory? In his recent book *String Theory and the Scientific Method*, Dawid (2013) provides a rationale for three non-empirical ways of arguing for a scientific theory that lack direct empirical support. One of them is the No Alternatives Argument (NAA), which was subsequently analyzed in the framework of BCT by Dawid, Hartmann, and Sprenger (2015). Our presentation follows this discussion.

The NAA relies on an observation at the meta-level; it is not a prediction of the theory itself, but rather an observation about the status of the theory. It is only in this sense non-empirical. The evidence, which is supposed to do the confirmatory work, is the observation that the scientific community has, after considerable effort, not yet found an alternative to a hypothesis H that solves some scientific problem **P**. Let us denote the corresponding proposition by $\mathrm{F_A}$ ("the scientific community has not yet found an alternative to H"). Let us furthermore define T as the proposition that the hypothesis H is empirically adequate. The propositional variables T and F_A take as values the previously described propositions and their respective negations.

To show that the meta-level observation $\mathrm{F_A}$ confirms T in the Bayesian sense, one has to show that

$$P(\mathrm{T}|\mathrm{F_A}) > P(\mathrm{T}). \tag{4.3}$$

Now, as $\mathrm{F_A}$ is neither a deductive nor an inductive consequence of T, there can be no direct probabilistic dependence between the two variables. Following the strategy suggested in the last section, we therefore look for a common cause variable that facilitates the dependence. But which variable could that be? Here Dawid, Hartmann, and Sprenger (2015) introduce the multi-valued variable Y, which has the values

$\mathrm{Y_k}$**:** There are k distinct alternative theories that solve **P**,

where k runs from 0 to some maximal value N. $\mathrm{Y_k}$ is a statement about the *existing* number of theories able to solve the scientific problem in question. It is easy to see that Y screens off F_A from T: Once we know the value of Y, learning $\mathrm{F_A}$ does not tell us anything new about the probability of T. To assess it, all that matters is that we know how many equally suitable candidate theories there are. Y facilitates the probabilistic dependence between F_A and T, since if there *are* only a small number of alternative theories, this would provide an explanation for why the scientists have not yet found any; that is, it would explain $\mathrm{F_A}$. In addition, if there *are* only a few alternative theories, that should probabilistically impact our trust in the available theory. This relies on an innocuous assumption, namely that there is at least one theory which is empirically adequate. If this is the case, the number of alternatives

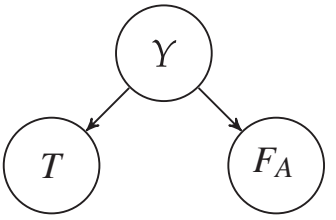

Figure 4.3 A Bayesian network depicting the No Alternatives Argument.

will probabilistically influence our trust in the theory. After the introduction of the variable Y and its inclusion in the Bayesian network depicted in Figure 4.3, one can show, given certain reasonable assumptions, that eq. (4.3) holds.[6]

This may suggest a very simple method to confirm scientific theories. However, this method relies on how well F_A functions as a probe of Y. Note that several complicating factors can arise. First, there might be another explanation for why the scientific community has not yet found an alternative. For instance, Dawid, Hartmann, and Sprenger (2015) introduce an additional node into the Bayesian network representing the difficulty of the scientific problem **P**. The observation of $\mathrm{F_A}$ then may provide support only for the difficulty of the problem. However, the difficulty of the problem is probabilistically independent of the empirical adequacy of the hypothesis, indicated by the variable T, and hence the observation of $\mathrm{F_A}$ may not confirm the theory in question. Second, our argument relies on it being possible to establish $\mathrm{F_A}$ in the first place. However, it is a nontrivial task to find agreement among the members of the scientific community about the existence or nonexistence of alternative solutions to some scientific problem **P**. Even if there is agreement, this is only probative of the existing number of alternatives, indicated by the value of the variable Y, provided that the scientific community has attempted to explore the space of alternative theories (see Oriti, this volume) and have considered all the problems one may encounter in doing so (see Dardashti, this volume). This may have as a normative consequence a requirement to change the way physics is practiced. Theory proliferation is not a common practice, but it may be required for a successful application of non-empirical theory assessment.

4.5 Illustration 2: Analogue Experiments

When scientists are concerned with black holes, neutron stars, or the whole universe, experiments are hard to come by. In these cases it has been suggested that

[6] One of these assumptions is that the agent is uncertain about the number of alternatives. If the agent were certain (i.e., if she knew the value of Y), then T and F_A would be probabilistically independent and F_A would not confirm T. For example, an anti-realist who adopts the underdetermination thesis to show that the number of alternatives is infinite (and therefore sets $P(\mathrm{Y}_\infty) = 1$) will not find the NAA convincing.

one may be able to use so-called analogue experiments.[7] The idea is to model the experimentally inaccessible *target system* via a table-top experiment, the *source system*, that has specifically been built to model the equations that are assumed to hold in the target system. Among the choices of source systems, one finds fluid systems, Bose–Einstein condensates (BECs), and optical lattices. The underlying physical laws in these source systems are, therefore, significantly different from the laws governing the inaccessible target system. This raises the question of what one can learn about a target system from analogue experiments (see Thébault, this volume). More specifically, one may ask whether the evidence obtained from manipulating the source system is also confirmatory evidence for the inaccessible target system.

The most-discussed examples of analogue experiments concern black hole Hawking radiation. The thermal radiation of black holes was predicted in the 1970s by Hawking. It has played an important role in foundational debates ever since (see Wüthrich [this volume] for a critical discussion), but lacks any direct empirical support. This motivated a first proposal of an analogue model of black hole Hawking radiation based on a fluid system by Unruh in the early 1980s. This model and many of the subsequently proposed analogue models[8] are very difficult to implement experimentally. After several partial successes in the last decade, Jeff Steinhauer (2016) finally announced that he has observed quantum Hawking radiation in a BEC analogue model.

Steinhauer's claim, however, goes even further. He stated that his findings provide "experimental confirmation of Hawking's prediction about the thermodynamics of the black hole" (Steinhauer in Haaretz, August 2016). Thus the evidence obtained in the experiment is taken to be evidence not only for the existence of Hawking radiation in BECs, but also of black hole Hawking radiation. This is in stark contrast to theoretical physicist Daniel Harlow's attitude regarding Steinhauer's experiment, namely that it is "an amusing feat of engineering that won't teach us anything about black holes" (*Quanta Magazine*, November 8, 2016). This example illustrates that scientists disagree about whether the observation of Hawking radiation in a BEC confirms black hole Hawking radiation. We submit that a Bayesian analysis can shed some light on this question.

To do so, we follow the analysis given in Dardashti, Hartmann, Thébault, and Winsberg (2018) where the technical details can be found. As a first step, we identify the relevant propositional variables and specify the probabilistic dependencies that hold between them. Let us denote by T_{BH} the binary propositional variable that takes the following values:

[7] See Barceló et al. (2011) for a comprehensive review article on analogue experiments of gravity.
[8] See Barceló et al. (2011, Ch. 4) for a discussion of classical and quantum analogue systems.

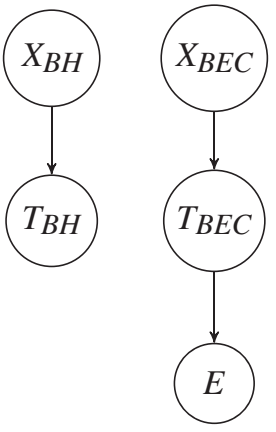

Figure 4.4 The Bayesian network of the BEC model and the black hole model (independent case).

T_{BH}**:** The model provides an empirically adequate description of the black hole system $\mathcal{T}$ within a certain domain of conditions.

$\neg T_{BH}$**:** The model does not provide an empirically adequate description of the black hole system $\mathcal{T}$ within a certain domain of conditions.

The domain of conditions encodes the conditions necessary for the respective context, namely the application of the model to derive Hawking radiation. Analogously, we define the binary propositional variable T_{BEC} that is associated with the BEC system and takes the following values:

T_{BEC}**:** The model provides an empirically adequate description of the BEC system $\mathcal{S}$ within a certain domain of conditions.

$\neg T_{BEC}$**:** The model does not provide an empirically adequate description of the BEC system $\mathcal{S}$ within a certain domain of conditions.

Furthermore, we introduce the propositional variable E, which has the values E: "The empirical evidence for the source system holds" and ¬E: "The empirical evidence for the source system does not hold."

To better understand how the probabilistic dependence between the two systems can occur, we need to introduce another layer of variables related to the domain of conditions mentioned previously. The domain of conditions depends on the various background assumptions involved in developing the model. The empirical adequacy of each model in the respective context therefore depends on whether the various background assumptions involved hold. These background assumptions rely on the knowledge of the modeler and involve both theoretical and experimental knowledge, both implicit and explicit. For instance, I may assume in my description of

the fluid model that the fluid is inviscid and barotropic. If the experiment turns out to agree with the predictions of that model, then that outcome not only supports the proposition regarding the empirical adequacy of the model, but also suggests that the assumptions were justified. Note that this does not entail that the model itself is actually viscosity-free or that barotropicity is a realistic assumption, but only that it was a justified assumption within the respective domain of conditions. Let us denote the variables representing the set of background assumptions by X_{BH} and X_{BEC}, respectively. They take the following values:

$\mathrm{X_{BH}}/\mathrm{X_{BEC}}$**:** The background assumptions are satisfied in the model of system $\mathcal{S}/\mathcal{T}$.

$\neg\mathrm{X_{BH}}/\neg\mathrm{X_{BEC}}$**:** The background assumptions are not satisfied in the model of system $\mathcal{S}/\mathcal{T}$.

If the background assumptions were probabilistically independent of the other variables, the relevant Bayesian network would be represented by Figure 4.4. In that case, there would be no probabilistic dependence, which does not seem to be unreasonable given that the assumptions in the context of black hole physics, and whether they are justified, seem to be independent of the assumptions involved in the context of BEC physics. However, it has been argued by Unruh and Schützhold (2005) that Hawking radiation may, under certain conditions, be considered a universal phenomenon. Here "universal" is meant in the sense that the phenomenon does not depend on the degrees of freedom at very high energies. The possible universality of the Hawking phenomenon now relates directly to one of the elements of the background assumptions involved in the black hole system, namely the possible influence of the trans-Planckian physics on the thermality of the radiation. The semi-classical derivation of Hawking radiation assumes that the trans-Planckian physics, whatever that might be, does not have an effect on whether Hawking radiation occurs. As the phenomenon relies on the physics at very high frequencies, a domain where the semi-classical approach is not applicable, the assumption is considered to be problematic. It is referred to as the "trans-Planckian problem." Note, however, that the analogue model contains a similar assumption regarding its independence from the underlying high energy theory. Based on these considerations, we introduce an additional variable U corresponding to the universality claim (see Figure 4.5). It has the following values:

U**:** Universality arguments in support of common background assumptions hold.

¬U**:** Universality arguments in support of common background assumptions do not hold.

One can now see how U can play the role of the common cause variable mentioned in Section 4.3: If the universality assumption is true, then that will directly impact

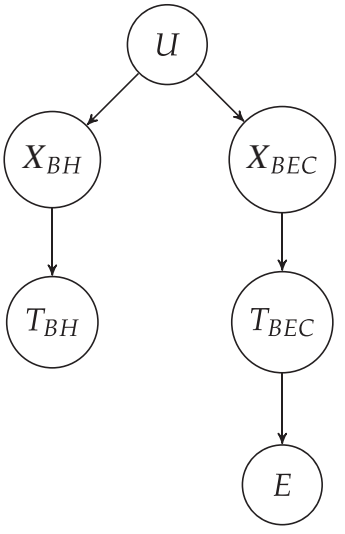

Figure 4.5 The Bayesian network of the BEC model and the black hole model (with universality).

the legitimacy of the corresponding background assumptions. Also, evidence for each analogue model of black hole Hawking radiation will provide empirical support for the universality claim itself, as each analogue model has a different high energy theory. Using the Bayesian network from Figure 4.5 one can then show that, under certain reasonable conditions, E confirms T_{BH}: $P(T_{BH}|E) > P(T_{BH})$.

Together with a number of plausible assumptions about the prior distribution, the Bayesian network depicted in Figure 4.5 provides a *possible* rationale for Steinhauer's strong claim about the empirical confirmation of black hole Hawking radiation. At the same time, it illustrates the more general problem of analogue experiments: To probe an inaccessible target system via analogue experiments, additional arguments establishing the probabilistic dependence need to be established. If these cannot be established, the analogue experiment cannot provide inductive support for the black hole system and may in this sense be just an "amusing feat of engineering." This, of course, does not rule out other (e.g., heuristic) roles it may have (see Thébault, this volume).

4.6 Discussion

In the previous two sections, we have shown how BCT can be used to evaluate ways of indirectly assessing scientific theories. Constructing Bayesian network models of the corresponding testing scenarios allowed us to explore the conditions under which these new ways of assessing scientific theories can be successful and

establish trust in a theory. In this section, we discuss a number of concerns related to the Bayesian approach and suggest how a Bayesian can respond to them.

1. The Bayesian approach is too flexible.

   The advantages that come with the flexibility of the Bayesian approach also bear the danger of too much flexibility ("anything goes"). Is it not always possible to come up with a complicated Bayesian network that does what one wants it to do? If this were so, then more needs to be done to make a convincing claim in favor of the confirmatory value of some piece of indirect evidence.

   In response, the following can be said: First, a model is always only as good as its assumptions. Sometimes called the GIGO principle ("garbage in, garbage out"), this concept holds for scientific models as well as for philosophical models. Clearly, the assumptions of a model have to be justified: We did, indeed, make sure that the crucial independence assumption (i.e., that the common cause variable screens off the other two variables) holds. We also discussed the possibility of including other variables (such as the difficulty of the problem in the NAA example) that may influence the conclusion. Second, by doing so, the Bayesian approach makes the reasoning transparent. It forces the modeler to put all assumptions on the table and it shows what can be concluded from a certain set of assumptions. The Bayesian approach provides a reasoning tool that may point to a hole in a putative argument.
2. Confirmation is an absolute, not a relative notion, and indirect confirmation cannot provide it.

   Rovelli (this volume, Chapter 7) claims that the common-sense understanding of confirmation means "very strong evidence, sufficient to accept a belief as reliable" and then goes on to claim that "only empirical evidence can grant 'confirmation' in the common sense."

   In response, we note that Rovelli conflates "confirmation" and "acceptance." While confirmation is a relation between theory and evidence, acceptance is not. According to BCT, a theory is confirmed if its probability increases after the evidence holds. In this case, the evidence is *one* reason in favor of the theory and confirmation is, therefore, an *epistemic notion*. For example, a theory is confirmed if its probability increases from 1% to 5%. Acceptance, by contrast, is a *pragmatic notion*. For example, one would not accept a theory if its probability is only 5%. Instead, the threshold for acceptance should be greater than 50%; that is, it must be more likely that the theory is true than that it is false.[9] Rovelli (this volume), Chapter 7 would set the threshold for acceptance even as high as 95%. If one sticks to this (Bayesian) use of the terms "confirmation" and

[9] Interestingly, this intuitive way of explicating acceptance is not without problems. For a discussion, see Leitgeb (2017).

"acceptance," then there is no problem with the claim that indirect evidence can confirm a theory (even if it does not lead to a high posterior probability).

3. The evidence provided by indirect confirmation is negligible.

   Related to the previous issue, indirect evidence is always much less effective than direct empirical evidence. Hence, we should not take indirect evidence seriously.

   In response, one may agree with the claim that direct evidence is typically better than indirect evidence. The observation of a new phenomenon that can be accounted for using string theory will provide us with much more confirmation than the NAA. However, if there is no such evidence (yet), then the search for indirect evidence for the various candidate theories is a means to better navigate in the epistemic landscape. Having discriminating indirect evidence, however small it is, in favor of string theory may provide a reason to keep working on the program and developing it, while having no evidence whatsoever does not support this claim (or justify the effort that is put into this research program).

4. BCT is subjective.

   BCT assumes that scientists assign subjective degrees of belief to their hypotheses, which they subsequently update in the light of evidence. The resulting posterior probabilities are also subjective, which in turn implies that confirmation is subjective. As a result, BCT does not provide an objective account of the relation between theory and evidence.

   In response, one may first point to the serious problems that more objective accounts such as HD confirmation or Popper's falsificationism face (see our discussion in Section 4.2). These problems led to the development of the Bayesian approach. Furthermore, Bayesians admit that there is an ineliminable subjective element in scientific reasoning and stress that the Bayesian approach makes it explicit. By putting on the table where subjective judgments come in (and where they do not), it becomes possible to assess the reasoning of a rational agent and to criticize it (if appropriate). Again, the Bayesian approach has the advantage of being transparent.[10]

4.7 Conclusion

The difficulty of experimentally probing certain energy regimes or reaching certain target systems have persuaded many physicists to seriously explore the possibility of alternative methods of theory assessment. At the same time, the stronger reliance on alternative methods has led to profound disapproval from members of the scientific community who question the viability of these methods. This makes

[10] Much more can be said on this and other objections to the Bayesian approach. See, for example, Howson and Urbach (2006).

it necessary to rigorously analyze the argumentative strategies employed when applying these methods, to identify the assumptions involved, and to investigate whether these methods can provide confirmation in the same way experiments can.

We have argued that standard qualitative accounts of confirmation (or corroboration) do not provide the tools for an analysis of these methods, as they are restricted to direct (deductive) consequences of the theories only. After having introduced Bayesian confirmation theory, we discussed how this powerful methodological framework provides the appropriate flexibility to rationally reconstruct various indirect ways of theory assessment. More specifically, we used the theory of Bayesian networks to demonstrate the possibility of indirectly confirming a scientific theory via a common cause structure. Here the evidence confirms the theory under certain conditions via a mediating common cause propositional variable. This methodology was then illustrated by two examples, the No Alternatives Argument and analogue experiments.

The crucial task in evaluating these indirect ways of theory assessment in the Bayesian framework is to identify a variable, if there is one at all, that plays the role of the common cause. This is a nontrivial task, and there is no real heuristic to find it. In the case of the No Alternatives Argument, the evidence being used does not probe the theory directly but rather addresses the space of theories as a whole, more specifically the number of alternatives available in that space. By assessing how constrained the theory space itself is, it can indirectly provide confirmation for a theory. In our second illustration, we argued that analogue experiments of black hole Hawking radiation can indirectly provide confirmation by probing directly a universality claim (which, in turn, probes the theory in question). In both cases, BCT provides the tools to reconstruct the argumentative strategies, to make transparent all assumptions involved, and to identify the normative consequences for the practicing scientists. The possibility of indirect confirmation, relying on nontrivial Bayesian network structures, opens up the possibility of analyzing other alternative ways of theory assessment.

Section 4.6 considered several possible objections to BCT and provided some replies. There is much more to say about the vices and virtues of indirect ways of assessing scientific theories within the Bayesian paradigm, but we hope to have shown that BCT is a good framework with which to start to critically and rationally assess the various methods, and a coherent framework for the various exciting new scientific methodologies.

Acknowledgment

We would like to thank our collaborators Richard Dawid, Jan Sprenger, Karim Thébault, and Eric Winsberg for useful discussions and feedback as well as the Alexander von Humboldt Foundation for financial support.

References

Barceló, C., S. Liberati, M. Visser, et al. (2011). Analogue Gravity. *Living Reviews in Relativity 14*(3).

Bovens, L., and S. Hartmann. (2004). *Bayesian Epistemology*. Oxford: Oxford University Press.

Dardashti, R., S. Hartmann, K. Thébault, and E. Winsberg. (2018). Hawking Radiation and Analogue Experiments: A Bayesian Analysis. Preprint available at https://arxiv.org/abs/1604.05932.

Dawid, R. (2013). *String Theory and the Scientific Method*. Cambridge: Cambridge University Press.

Dawid, R., S. Hartmann, and J. Sprenger. (2015). The No Alternatives Argument. *British Journal for the Philosophy of Science 66*(1), 213–34.

Diaconis, P., and S. L. Zabell. (1982). Updating Subjective Probability. *Journal of the American Statistical Association 77*(380), 822–30.

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

Eva, B., and S. Hartmann. (2018). Bayesian Argumentation and the Value of Logical Validity. *Psychological Review* 125(5), 806–821.

Hájek, A., and S. Hartmann. (2010). Bayesian Epistemology. In: J. Dancy et al. (eds.), *A Companion to Epistemology*. Oxford: Blackwell, pp. 93–106.

Hartmann, S., and J. Sprenger. (2010). Bayesian Epistemology. In: S. Bernecker and D. Pritchard (eds.), *Routledge Companion to Epistemology*. London: Routledge, pp. 609–20.

Hempel, C. G. (1945). Studies in the Logic of Confirmation II. *Mind 54*(214), 97–121.

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

Howson, C., and P. Urbach. (2006). *Scientific Reasoning: The Bayesian Approach*. London: Open Court Publishing.

Leitgeb, H. (2017). *The Stability of Belief: How Rational Belief Coheres with Probability*. Oxford: Oxford University Press.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Schurz, G. (1991). Relevant Deduction. *Erkenntnis 35*(1–3), 391–437.

Sprenger, J. (2011). Hypothetico-Deductive Confirmation. *Philosophy Compass 6*(7), 497–508.

Sprenger, J., and S. Hartmann. (2019). *Bayesian Philosophy of Science*. Oxford: Oxford University Press.

Steinhauer, J. (2016, August). Observation of Quantum Hawking Radiation and Its Entanglement in an Analogue Black Hole. *Nature Physics 12*(10), 959

Unruh, W. G., and R. Schützhold. (2005). Universality of the Hawking Effect. *Physical Review D 71*(2), 024028.

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

Vineberg, S. (1997). Dutch Books, Dutch Strategies and What They Show About Rationality. *Philosophical Studies 86*(2), 185–201.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.