The formal structure(s) of analogical inference*

Alexander Gebharter · Barbara Osimani

Abstract: Recently, Dardashti, Hartmann, Thébault, and Winsberg (2019) proposed a Bayesian model for establishing Hawking radiation by analogical inference. In this paper we investigate whether their model would work as a general model for analogical inference. We study how it performs when varying the believed degree of similarity between the source and the target system. We show that there are circumstances in which the degree of confirmation for the hypothesis about the target system obtained by collecting evidence from the source system goes down when increasing the believed degree of similarity between the two systems. We then develop an alternative model in which the direction of the variation of the degree of confirmation always coincides with the direction of the believed degree of similarity. Finally, we argue that the two models capture different types of analogical inference.

*This is a draft paper. Please do not cite or quote without permission. The final version of this paper forthcoming under the following bibliographical data: Gebharter, A., & Osimani, B. (forthcoming). The formal structure(s) of analogical inference. *Erkenntnis*.

1 Introduction

Scientists often rely on analogical inference of various kinds. This might be for several reasons. The dynamics of black holes, for example, cannot be observed directly for experimental and theoretical reasons. They might, however, be studied indirectly via investigating how similar (or analog) enough systems behave (cf. Dardashti, Thébault, & Winsberg, 2015). Direct evidence about climate change is available, but one has to wait for long periods of time for new evidence to come in. So there is also a lot of relevant evidence not accessible for practical reasons. This is why climate change is regularly studied on the basis of computer simulations.[1] A crucial assumption is, again, that these models capture the actual climate dynamics well enough. Finally, direct evidence might be inaccessible for moral reasons. An example would be how a new antiviral compound is studied in medicine and pharmacology. Before it is tested on humans, it is tested on a suitable model organism. Scientists typically rely on rats since their immune system functions similarly to the human immune system. There are even specific breeding programs aiming at making the immune system of rats even more similar to the human immune system in order to create even better model organisms (cf. Levy & Currie, 2015).

What all of these examples have in common is that scientists aim at establishing a hypothesis about a certain system of interest, the *target system*, but are not able to do so on the basis of direct evidence (alone). Instead, they (also) study another easier accessible system, the *source system*, for which they have good reasons to believe that it works similarly enough in order to justify an inference about the target system. Now one crucial question is how exactly these different kinds of analogical inference work. How can they be analyzed

[1] Whether computer simulations are a subspecies of analogical inference is still debated (cf. Boge, 2020; Winsberg, 2009).

and under which conditions are analogical inferences guaranteed to succeed? In this paper, we are concerned with questions like these as well as with a certain strategy to answer them. In particular, we focus on how these questions can be tackled in a Bayesian framework. Recently, Dardashti et al. (2019) proposed a Bayesian model for establishing Hawking radiation by analogical inference. Their model provides an answer to these questions for a particular case. They formulate several conditions which, if satisfied, guarantee that observing the evidence of the source qualitatively confirms the hypothesis about the target system. Though their model is intended as a model for a specific case of analogical inference aiming at establishing Hawking radiation,[2] Feldbacher-Escamilla and Gebharter (2020) argued that it has some merits as a general model for analogical inference as well.[3] In this paper, we follow this trend and further explore whether their model is suitable to cover analogical inference in general.

After introducing the formal details of Dardashti et al.'s (2019) model in section 2 and abstracting away from the specific case of Hawking radiation, we formulate two problems for the model as a general model for analogical inference. The first one will be a minor problem which can easily be solved. However, the motivation for the strategy to fix that minor problem will provide a relevant intuition on which we will heavily rely on later. The second problem points to more fundamental issues regarding analogical inference. It arises if one considers

[2] Crowther, Linnemann, and Wüthrich (2021) argue that Dardashti et al.'s (2019) model does not work in the specific case of Hawking radiation because no independent evidence for the analogy is provided. The role of independent support for the similarity of the source and the target system is important and we will come back to it later.

[3] Feldbacher-Escamilla and Gebharter (2020) also point out several problems for a Dardashti et al.'s (2019) model if understood as a general model for analogical inference. Since they admit that these problems do not necessarily form the basis of a compelling argument against the model, we bracket them in this paper. For connections of the model to the literature on abduction and unification, see (Feldbacher-Escamilla & Gebharter, 2019; Glymour, 2019).

how the degree of confirmation provided by the evidence of the source system to the target system is influenced by another factor: the believed degree of similarity of the source system to the target system. We show that sometimes an increase in the believed degree of similarity of the two systems does not coincide with an increase in the degree of confirmatory impact the evidence of the source system has on the hypothesis about the target system, which goes against basic intuitions underlying many cases of analogical inference. Take the rat study case introduced earlier as an example: Intuitively, we expect that the more certain we become that the immune system of the rats used in the study resembles the human immune system, the more our findings on the model organism confirm the corresponding hypothesis about how the antiviral compound works on humans.

As a next step, we propose an alternative Bayesian model for analogical inference in section 3. We formulate constraints under which our alternative model guarantees that evidence about the source system qualitatively confirms the hypothesis about the target system. Next, we show that our alternative model can account for the problems that arise for Dardashti et al.'s (2019) model if understood as a general model: In the alternative model, the direction in a change of the degree of confirmation provided by the evidence of the source system for the corresponding hypothesis about the target system always goes hand in hand with the direction of a change in the believed degree of similarity between the two systems.

In section 4, we compare the Dardashti et al.'s (2019) original model and our alternative model. We argue that none of the two models is suitable as a general model for analogical inference, but that nonetheless each of them has its proper place in scientific reasoning: While the original model represents the case of analogical confirmation by means of a theoretical model or full-fledged theory

under whose domain both the source and the target system fall, our alternative model may be taken as a formal reconstruction of extrapolation via analogy. In the former case, the structure of the source system is assumed to reproduce the relevant theoretical features that should be sufficient to predict/reproduce the phenomena observed in the target system because both systems are governed by the same theoretical model or theory. In the latter case, the analogy is based upon knowledge of a set of shared features between the two systems. We argue that our proposal vindicates and provides a formalization of the distinct epistemic dynamics at work in the use of theoretical models and theories vs. extrapolations (e.g., based on model organisms) in scientific practice.

2 Analogical inference Bayesian style 1.0

To set the stage, let us think about analogical inference in somewhat more technical terms. Let s stand for the source system and let t stand for the target system. Next, let us introduce a few binary variables. Assume that H_t expresses the hypothesis to be established about the target system. In the rat study example introduced in section 1, $H_t = 1$ could be interpreted as the antiviral compound being efficacious (without leading to severe side effects) for humans and $H_t = 0$ as the negation of that claim. Let H_s represent the corresponding hypothesis about the source system. In the example, $H_s = 1$ and $H_s = 0$ would make the same claims as $H_t = 1$ and $H_t = 0$ make about the target system, but for rats instead of humans. Finally, let E_s represent the direct evidence about the source system. In our example, it could simply stand for the outcome of the rat study. $E_s = 1$ could stand for a positive and $E_s = 0$ for a negative outcome. The general structure of the setup is graphically illustrated in Figure 1. For convenience, we will write binary variables in italics, while upper case letters not written in italics will stand for the corresponding variable's values. The

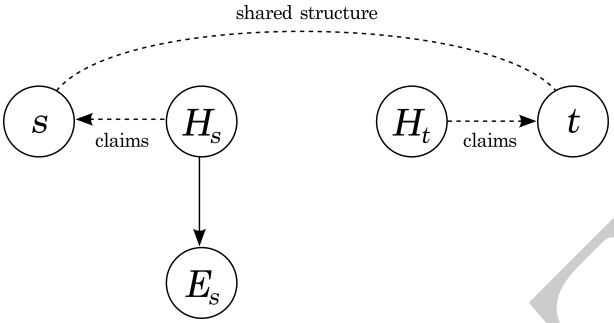

Figure 1: General structure of analogical inference. E_s confirms H_t by being direct evidence for H_s (represented by the continuous arrow), H_s making similar claims about the source system s as H_t makes about the target system t (represented by the dashed arrows), and s and t sharing structural features relevant for H_s and H_t (represented by the dashed line).

binary variable H_t, for example, can take the value H_t for $H_t = 1$ or the value $\bar{H}_t$ for $H_t = 0$.

Before we can present Dardashti et al.'s (2019) model, we need to introduce one more variable modelling the structural similarity between the source and the target system. This job is done by the binary variable X. In Dardashti et al.'s (2019) original model, X represents universality arguments that are established with much care.[4] Since we are not interested in Hawking radiation but in whether their model works as a general model for analogical inference, we will not follow their specific interpretation of the binary variable X. Instead, we interpret X as postulating a relevant similarity between the source and the target system and $\bar{X}$ as denying such a similarity. Thus, higher $Pr(X)$ values indicate that the source system and the target system are considered to more likely

[4]For further details on how X can be interpreted as representing universality arguments and on the conditions under which they can be built up to support analogical inference from the source to the target system, see (Field, 2021).

behave similarly w.r.t. the specific hypotheses and phenomena of interest. This is in accord with Dardasthi et al.'s interpretation as well as with the classical assumption from the analogical inference literature that the source system and the target system share relevant similarities (cf. Hesse, 1966). To further flesh out X, one can follow this tradition and interpret X as the claim that the source system s and the target system t are similar w.r.t. a certain feature and $\bar{\mathrm{X}}$ as the negation of that claim (cf. Feldbacher-Escamilla & Gebharter, 2020). This feature can be simple or complex consisting of several subfeatures. We also assume that it is relevant for the truth of both hypotheses H_s and H_t and that the two systems s and t are not similar in other relevant respects. This leaves room for the possibility that there are other features only realised by one of the two systems or not relevant for the truth of both H_s and H_t. We interpret $Pr(\mathrm{X})$ as the agent's *believed degree of similarity* of s and t w.r.t. that feature. Given this interpretation, $Pr(\mathrm{X}) = 1$ stands for belief in maximal similarity, $Pr(\mathrm{X}) = 0$ for belief in no similarity, and $0 < Pr(\mathrm{X}) < 1$ for belief in all the different degrees of similarity between the two extremes.[5]

Dardashti et al. (2019) build a Bayesian network[6] with the graph depicted in Figure 2. In addition, they formulate the following constraints for the prob-

[5]Under this interpretations $Pr(\mathrm{X})$ and $Pr(\bar{\mathrm{X}})$ function as weights steering the agent's believed degree of similarity between maximal similarity vs. no similarity w.r.t. a certain feature. If we assume that believed degree of similarity w.r.t. that feature can be measured by a function *sim* normalized to the interval $[0, 1]$, then $Pr(\mathrm{X})$ and $Pr(\bar{\mathrm{X}})$ are related to *sim* as follows:

$$sim = 1 \cdot Pr(\mathrm{X}) + 0 \cdot Pr(\bar{\mathrm{X}})$$

Since $0 \cdot Pr(\bar{\mathrm{X}}) = 0$, *sim* collapses to $Pr(\mathrm{X})$, which makes $Pr(\mathrm{X})$ perfectly suited to express the believed degree of similarity of s and t w.r.t. a certain feature.

[6]For a primer on Bayesian networks, see the Appendix.

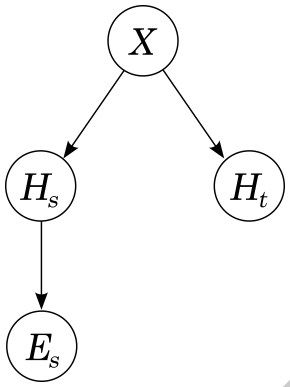

Figure 2: Bayesian network for modeling analogical inference

ability distribution Pr associated with this graph:

$$0 < Pr(\mathrm{X}) < 1 \tag{1}$$

$$\forall i \in \{s,t\} : Pr(\mathrm{H}_i|\mathrm{X}) > Pr(\mathrm{H}_i|\bar{\mathrm{X}}) \tag{2}$$

$$Pr(\mathrm{E}_s|\mathrm{H}_s) > Pr(\mathrm{E}_s|\bar{\mathrm{H}}_s) \tag{3}$$

Equation 1 requires the probability distribution over X to be non-extreme. Equation 2 says that structural similarity between the source and the target system renders the truth of the corresponding hypotheses more likely. Equation 3 reflects the assumption that E_s is considered to be positive evidence for H_s.

Finally, we also need to say a few words about the particular structure of the Bayesian network in Figure 2: The edge $H_s \longrightarrow E_s$ indicates that E_s stands for the direct evidence for the hypothesis about the source system. This corresponds to the usual way how hypothesis and evidence are connected in Bayesian networks (cf. Bovens & Hartmann, 2003). Because it is assumed in analogical inference that information coming from the source system can reach the target system only due to shared structural features, the variable X modelling

these shared structural features needs to mediate between these two systems. This is represented by $H_s \longleftarrow X \longrightarrow H_t$. This representation guarantees that H_s-changes can lead to H_t-changes only due to X-changes.

Throughout the paper we use the ordinary Bayesian difference measure as a proxy for measuring confirmation:

$$c(\mathrm{H}_t; \mathrm{E}_s) = Pr(\mathrm{H}_t|\mathrm{E}_s) - Pr(\mathrm{H}_t)$$

As briefly mentioned in section 1, Dardashti et al. (2019) are mainly concerned with a qualitative notion of confirmation. For now, we follow their lead. We ignore the numerical details and rather focus on whether $c(\mathrm{H}_t; \mathrm{E}_s)$ is positive (confirmation), zero (no confirmatory impact), or negative (disconfirmation). One of the main achievements of their paper is to prove the following theorem:[7]

Theorem 2.1. *For every Bayesian network with the graph in Figure 1 whose probability distribution satisfies Equations 1–3:* $c(\mathrm{H}_t; \mathrm{E}_s) > 0$.

Theorem 2.1 establishes that under certain conditions analogical inference can be analyzed in terms of Bayesian updating.

Next, we will point at two problems with the model if understood as a general model for analogical inference. We start with the minor problem. This problem concerns the probabilistic constraint Equation 2. One might contest this assumption's plausibility. In words, it says that if the source and the target system are similar w.r.t. a certain feature, then the hypotheses H_s and H_t should both be more likely to be true. But why should the probability of both

[7] The other theorem Dardashti et al. (2019) prove is about the increase of confirmatory impact when several source systems are available. It can be used to further justify their universality arguments. Since we are interested in models for analogical inference in general and not particularly in building up universality arguments, we bracket further discussion of that theorem.

hypotheses increase rather than decrease? What one can actually learn about the truth of the hypotheses H_s and H_t by establishing that the source and the target system are similar w.r.t. a certain feature is that both hypotheses should be more likely to be true or false *together*. Accordingly, we can modify the original constraint Equation 2 as follows:

$$\forall i \in \{s,t\} : Pr(\mathrm{H}_i|\mathrm{X}) > Pr(\mathrm{H}_i|\bar{\mathrm{X}}) \text{ or } \forall i \in \{s,t\} : Pr(\mathrm{H}_i|\mathrm{X}) < Pr(\mathrm{H}_i|\bar{\mathrm{X}}) \quad (4)$$

Equation 4 captures the basic idea outlined above: Establishing that s and t are indeed analogue w.r.t. some relevant feature does not yet provide any information about whether the hypotheses about these systems are more likely to be true or false. What it gives us, however, is a good reason to believe that they are more likely to be true or false together. Since this basic intuition underlying the connection of X to H_s and H_t will play a major role later on in section 3, we shall keep it in mind.

Before we come to the second problem, let us briefly state that replacing Equation 2 by Equation 4 does indeed change nothing about the result that analogical inference can be analyzed in terms of Bayesian updating:

Theorem 2.2. *For every Bayesian network with the graph in Figure 1 whose probability distribution satisfies Equations 1, 3, and 4:* $c(\mathrm{H}_t; \mathrm{E}_s) > 0$.

Next, let us come to the more severe problem announced in section 1. It arises if we take the believed degree of similarity between the source and the target system and the evidence for that similarity into account. That the similarity required to get the analogical inference going cannot simply be postulated a priori from the armchair, but needs some independent support, has been argued for in detail by Crowther et al. (2021) for the specific case of Hawking radiation. We agree with Crowther et al. that independent support for the

analogy is essential for analogical inference, but leave it open here of what kind this support has to be.[8] We allow for support in the form of empirical evidence, but also for support in the form of testimony, expert assessments, or other forms of evidence compatible with an orthodox Bayesian understanding of evidence.[9] One way to add such independent evidence to the model that suggests itself is to use reliability models, which have been investigated in detail in (Bovens & Hartmann, 2003).[10] These models allow us to represent independent evidence as well as to manipulate the believed degree of similarity between the source and the target system by varying the believed degree of reliability of the evidence's source.

A suitable reliability model would be a Bayesian network with the graph depicted in Figure 3 that satisfies the probabilistic constraints in Table 1, where $a > 0$. E_X represents the independent evidence about the structural similarity expressed by X, while R_X stands for the reliability of this evidence's source. As before, E_X stands for positive evidence and $\bar{\mathrm{E}}_X$ for negative evidence for X. R_X stands for the source being reliable, while $\bar{\mathrm{R}}_X$ stands for the source not being reliable. That both R_X and X are root variables reflects the assumption that the reliability of the source is not biased (i.e., influenced by the degree of similarity between s and t).

The probabilistic constraints in Table 1 allow for modelling the believed de-

[8] For non-empirical support for the universality arguments used by Dardashti et al. (2019), see (Field, 2021).

[9] See, for example, Bayesian formalizations of the no alternatives and the no miracles argument (cf. Dawid, Hartmann, & Sprenger, 2014; Sprenger, 2015; Sprenger & Hartmann, 2016).

[10] For other applications of reliability models see, for example, (de Pretis Francesco, Landes, & Osimani, 2019; Henderson & Gebharter, 2021; Landes, Osimani, & Poellinger, 2017; Osimani & Landes, 2020).

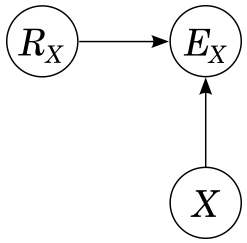

Figure 3: Reliability model for the similarity hypothesis X

X	R_X	$Pr(\mathrm{E}_X \mid X, R_X)$
0	0	a
0	1	0
1	0	a
1	1	1

Table 1: Constraints on the conditional probabilities $Pr(\mathrm{E}_X|X, R_X)$ constituting a reliability model (where $a > 0$)

gree of similarity on the basis of the evidence E_X and the believed degree of reliability of its source. If the source is believed to be unreliable, then the probability to find the evidence is a, regardless of whether the similarity hypothesis X is true or false. If the evidence's source is believed to be reliable, then the truth of the similarity hypothesis does with certainty produce the evidence E_X and its falsity does with certainty produce the counter-evidence $\bar{\mathrm{E}}_X$. As a result, assigning probability 0 to R_X renders the posterior probability of X equal to the prior probability of X, and assigning probability 1 to R_X results in X's posterior probability being 1 or 0, depending on whether E_X or $\bar{\mathrm{E}}_X$ is observed. The full range of posterior probabilities of X between these two extremes can be captured by assigning probabilities greater than 0 and smaller than 1 to R_X. Since these posterior probabilities will represent the believed degree of similarity, all logically possible believed degrees of similarity can be modelled by varying the

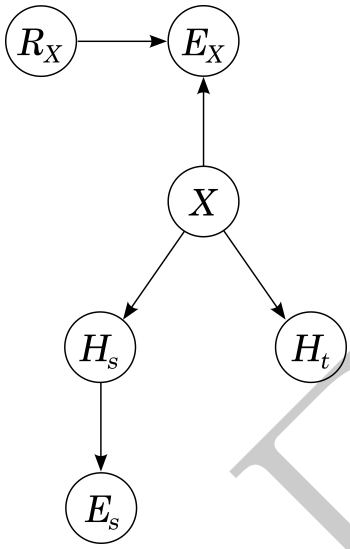

Figure 4: Model for analogical inference taking independent evidence for the similarity hypothesis as well as the reliability of its source into account

probability distribution over R_X.[11]

We are now able to combine our Bayesian model for analogical inference and the reliability model for X. We thus arrive at the Bayesian network whose graph is depicted in Figure 4. Finally, we can assess the confirmatory impact of E_s on H_t under variations of the believed degree of similarity of the source and the target system by computing the confirmatory impact of E_s on H_t conditional

[11] The probability distribution over R_X is interpreted similarly as the one over X: $Pr(\mathrm{R}_X)$ represents the *believed degree of reliability*. Accordingly, if believed degree of reliability is measured by a function *bdr* normalized to the interval $[0, 1]$, then

$$bdr = 1 \cdot Pr(\mathrm{R}_X) + 0 \cdot Pr(\bar{\mathrm{R}}_X)$$

reduces to $Pr(\mathrm{R}_X)$, which makes $Pr(\mathrm{R}_X)$ perfectly suitable to capture the believed degree of reliability.

on the evidence about the similarity E_X as follows:

$$c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) = Pr(\mathrm{H}_t|\mathrm{E}_s, \mathrm{E}_X) - Pr(\mathrm{H}_t|\mathrm{E}_X)$$

With all these assumptions in place, we can now make the following observations about our Bayesian model for analogical inference:[12]

Theorem 2.3. *For every Bayesian network with the graph in Figure 4 whose probability distribution Pr satisfies Equations 1, 3, and 4 as well as the constraints in Table 1:*

(a) $c(\mathrm{H}_t; \mathrm{E}_s) > 0$.

(b) If $Pr(\mathrm{R}_X) = 1$, *then* $c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) = 0$.

Theorem 2.4. *For some Bayesian networks with the graph in Figure 4 whose probability distribution Pr satisfies Equations 1, 3, and 4 as well as the constraints in Table 1 there exist distributions* Pr^* *resulting from Pr by increasing* $Pr(\mathrm{R}_X)$ *while keeping all other parameters fixed such that:* $Pr^*(\mathrm{R}_X) < 1$ *and* $c^*(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) < c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X)$.

Let us briefly have a closer look at what these observations mean for our model as a general Bayesian model for analogical inference and why they are

[12] For Theorems 2.3 and 2.4, we manipulated the degree of similarity between the source and the target system by varying the reliability of the source represented by $Pr(\mathrm{R}_X)$ conditional on the independent evidence E_X. As is easy to see from the proofs, similar theorems can be proven by directly manipulating $Pr(\mathrm{X})$. We decided to state our theorems on the basis of the expanded model shown in Figure 4 for three reasons: It (i) reflects Crowther et al.'s (2021) observation that some sort of evidence is essential for analogical inference, (ii) shows that taking such independent evidence into account cannot solve the problems expressed by Theorems 2.3 and 2.4, and (iii) allows us to treat our model for analogical inference based on Dardashti et al.'s (2019) model as well as the one we will develop in section 3 in a unified way, which requires a technical device like the reliability model since the probability distribution over X cannot be manipulated directly in the latter model.

problematic. We start with Theorem 2.3(a), which restates Theorems 2.1 and 2.2 for the expanded model in Figure 4. The expanded model shows that this result is problematic. Theorem 2.3(a) tells us that even if the evidence E_X about the structural similarity between the source and the target system is not considered at all, the evidence about the source system E_s has some confirmatory impact on the hypothesis about the target system H_t. This contradicts scientific practice: If one does not have independent support for the structural similarity, may it be in the form of empirical evidence as Crowther et al. (2021) demand or in other forms, analogical inference should not be licensed. A supporter of the model can accommodate this observation by restricting its domain of application, in particular, by requiring that the model is only suitable to capture analogical inference if independent evidence E_X is available and taken into account. The model is simply not applicable otherwise.

Let us now turn to Theorem 2.3(b). It says that if we have perfect evidence (conditioning on E_X and assigning probability 1 to R_X) that the source and the target system are similar with respect to features relevant for H_s and H_t, then observing the evidence about the source system E_s results in no confirmatory impact on the hypothesis about the target system at all.[13] In many cases involving extrapolation this stands in stark contrast to our intuitions about scientific practice. In such cases one would think that the more similar the scientist considers the two systems under consideration to be, the more reliable the analogical inference should become. Recall, for example, the rat study case: Assuming that we believe that the similarity of the immune system of the rats used in the study and the human immune system is maximal when it comes to the response of the kind of antiviral compound we want to test should give us as

[13] Note that this observation is not in tension with the assumption that the prior probability of X is non-extreme (Equation 1).

much confidence as we can hope for when inferring that the compound will work in the same way on humans as it does on rats. Though this move would seem a bit ad hoc, a supporter of the model could accommodate also this observation by further restricting the domain of application: What Theorem 2.3(b) ultimately shows is that the model does not work for the extreme case in which the source of the independent evidence E_X is believed to be perfectly reliable.

Finally, the probably most cumbersome observation is expressed in Theorem 2.4. It says that there are not only cases in which the believed degree of similarity of the source and the target system going up (represented by increasing the prior probability of R_X while conditioning on E_X) goes hand in hand with the degree of the confirmatory impact E_s has on H_t, but also cases in which the degree of confirmatory impact E_s has on H_t goes down while the believed degree of similarity goes up. An example is provided in Figure 5.[14] This, again, stands in stark contrast to our intuitions and scientific practice, especially in contexts involving extrapolation. It conflicts, for example, with the attempt to breed strands of rats whose immune system behaves as similarly as possible to the human immune system. This problem seems not to be easily resolvable by further narrowing down the domain of application since it does not only plague extreme cases like the other problems expressed by Theorem 2.3. Theorem 2.4 covers a wide range of non-extreme probability assignments to R_X. Even when ignoring the extreme case where the scientist believes in maximal structural similarity, an increase in the believed degree of similarity between the source and the target system should continuously boost the confirmatory impact the evidence E_s has on H_t.

In the next section, we will suggest a modification of the model that results

[14] For the model's parameters used to produce this example, see the proof of Theorem 2.4 in the Appendix.

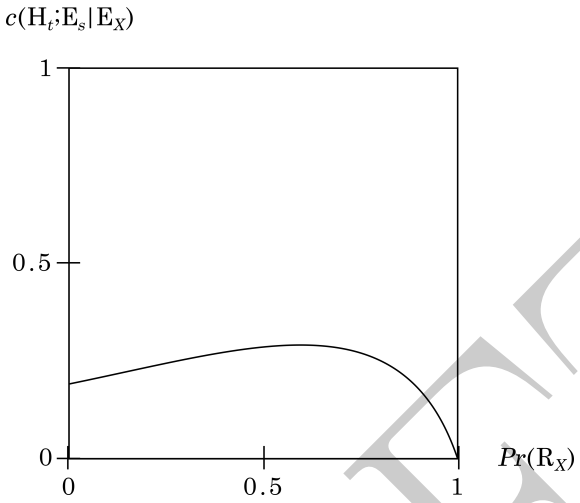

Figure 5: Graph showing that an increase in the believed degree of similarity can coincide with a decrease in confirmatory impact even in non-extreme cases

in an alternative Bayesian model for analogical inference that takes into account the basic intuitions conflicting with Theorems 2.3 and 2.4 we discussed above. In section 4 we will come back to the problems discussed above and argue that though they arise for a certain type of analogical inference, there is another type actually better covered by the network structure and the formal assumptions made by Dardashti et al. (2019): confirmation via a theoretical model or theory.

3 Analogical inference Bayesian style 2.0

Here are two possibilities of how the original model could be modified in order to avoid the problematic observations made in section 2. One could reconsider (i) the structure of the Bayesian network in Figure 4 or (ii) the constraints on the Bayesian network's probability distribution expressed in Equations 1, 3, and 4. Sticking to the original graphical structure and only modifying some of the probabilistic constraints does, however, not seem too promising. Equation 1

says that one's believed degree of similarity between s and t should not be extreme. This nicely fits scientific practice: We do not want to exclude that the believed degree of similarity can be changed later on.[15] Equation 3 is a standard assumption made in Bayesian epistemology (cf. Bovens & Hartmann, 2003). It simply expresses that E_s is evidence for H_s. Finally, as we already saw in section 2, also Equation 4 is plausible: The structural similarity of the source and the target system renders H_s and H_t more likely to be true or false together. Without this assumption it could not be guaranteed that the outcome of the rat study, for example, means that a similar outcome could be expected if the experiment would be repeated on humans.

For the reasons mentioned above, we opt for option (i): changing the graph of the Bayesian network. Let us have a brief look at the different structural elements of the original graph. The edge between H_s and E_s corresponds to the standard representation in Bayesian epistemology of the assumption that E_s (or $\bar{\mathrm{E}}_s$) is direct evidence for H_s (or $\bar{\mathrm{H}}_s$). Observing E_s (or $\bar{\mathrm{E}}_s$) can have a probabilistic effect on other variables only through H_s. This is captured by the fact that the edge between H_s and E_s is the only edge connecting E_s to one of the other variables in the graph. Thus, this edge seems justified. As we already saw in section 2, a path between H_s and H_t going through X (without any other connection between H_s and H_t) is also required. This reflects the idea that any probabilistic influence between the source and the target system can only be due to their similarity modelled by X. However, $H_s \longleftarrow X \longrightarrow H_t$ is not the only structure that can do this job. All in all, there are four possible candidate structures:

(a) $H_s \longleftarrow X \longrightarrow H_t$

(b) $H_s \longrightarrow X \longrightarrow H_t$

[15] An extreme distribution over X cannot be changed by conditioning on other variables.

(c) $H_s \longleftarrow X \longleftarrow H_t$

(d) $H_s \longrightarrow X \longleftarrow H_t$

Each one of these structures can be combined with different orientations of the edge between H_s and E_s. Sticking to (a) and flipping the arrow $H_s \longrightarrow E_s$ would not allow for any change of H_t's probability distribution after observing E_s's value. Since the concatenation of (a) and $H_s \longrightarrow E_s$ and the concatenation of (b) and $H_s \longrightarrow E_s$ are probabilistically indistinguishable (Definition 6.5 in the Appendix), replacing (a) by (b) would still result in the problematic observations made in section 2. The same goes for combining (b) with a flipping of the arrow $H_s \longrightarrow E_s$. Replacing (a) by (c) would not help either. The concatenation of (c) and $H_s \longrightarrow E_s$ would also result in a graph that is probabilistically indistinguishable from the concatenation of (a) and $H_s \longrightarrow E_s$. Again, flipping the arrow $H_s \longrightarrow E_s$ would not help, since the concatenation of (c) and $H_s \longleftarrow E_s$ would not allow for any change of H_t's probability distribution after observing E_s's value. Finally, only structure (d) is left. The structure resulting from (d) by adding $H_s \longrightarrow E_s$ and the one resulting from (d) by adding $H_s \longleftarrow E_s$ are probabilistically indistinguishable. Since the arrow $H_s \longrightarrow E_s$ is closer to the original model which we already investigated in section 2, we focus on this structure and explore its merits in the remainder of this section.

First of all, let us expand the concatenation of (d) and $H_s \longrightarrow E_s$ by a reliability model for X. The graph of the resulting Bayesian network is depicted in Figure 6. As before, we need to consider relevant constraints for the probability distributions to be associated with this graph. Again, we consider E_s (or $\bar{\mathrm{E}}_s$) to be direct evidence for H_s (or $\bar{\mathrm{H}}_s$). Thus, we stick to the original Equation 3. In our alternative model we have the two variables H_s and H_t replacing X as the only root variable. We follow Dardashti et al. (2019) in assuming that the probability distributions over our core model's root variables are non-extreme.

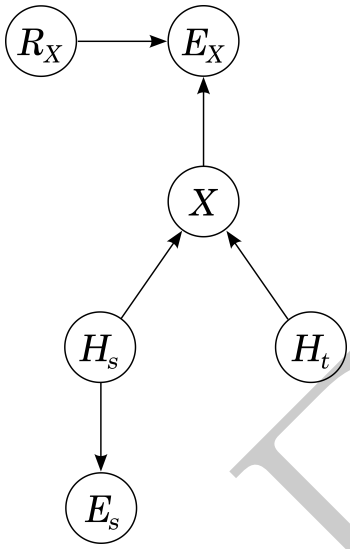

Figure 6: Alternative model for analogical inference

Thus, we assume that:

$$\forall i \in \{s,t\} : 0 < Pr(\mathrm{H}_i) < 1 \tag{5}$$

The motivation for Equation 5 is similar to the motivation for the original Equation 1: Since science is fallible, it should always be possible that new incoming evidence impacts our current beliefs.

Finally, we need to consider how exactly the probabilistic mechanism underlying the substructure $H_s \longrightarrow X \longleftarrow H_t$ works. We propose the following probabilistic constraint for characterizing this mechanism:

$$x_{\mathrm{H}} = x_{\bar{\mathrm{H}}} > x_{\bar{\mathrm{H}}} \tag{6}$$

The new expressions in Equation 6 are defined as follows:

$$
\begin{aligned}
x_{\mathrm{H}} &\coloneqq Pr(\mathrm{X}|\mathrm{H}_s, \mathrm{H}_t) \\
x_{\bar{\mathrm{H}}} &\coloneqq Pr(\mathrm{X}|\mathrm{H}_s, \bar{\mathrm{H}}_t) = Pr(\mathrm{X}|\bar{\mathrm{H}}_s, \mathrm{H}_t) \\
x_{\bar{\bar{\mathrm{H}}}} &\coloneqq Pr(\mathrm{X}|\bar{\mathrm{H}}_s, \bar{\mathrm{H}}_t)
\end{aligned}
\tag{7}
$$

The motivation for the probabilistic mechanism underlying $H_s \longrightarrow X \longleftarrow H_t$ is closely related to the one we gave for Equation 4 in section 2. A relevant structural similarity between the source and the target system (as postulated by X) would render the two hypotheses H_s and H_t more likely to be true or false together. This also goes the other way round: If the two hypotheses H_s and H_t are true or false together, this makes it for the very same reasons more likely that the two systems share relevant structural similarities, which would explain why both hypotheses are true or false together. This basic intuition is expressed by the inequality sign in Equation 6. This can be illustrated by means of the rat study example: If the immune system of rats indeed functions like the immune system of humans, then it is more likely that the antiviral compound to be tested leads to the intended effect for both or for none of these two kinds of organisms. And vice versa: The fact that it leads to the intended effect for both or for none of the two types of organisms increases the probability that their immune systems indeed share relevant features, which would explain that both types of organisms react to the antiviral compound in the same way.

We can say even more about the mechanism underlying $H_s \longrightarrow X \longleftarrow H_t$: If rats and humans indeed share relevant structural similarities, this will not discriminate between whether the antiviral compound will work on both or on none of the two kinds of organisms. The structural similarity tells us only that the two types of organisms will more likely react in the same way to the test, but nothing more. This is expressed by the part of Equation 6 stating that

$x_{\mathrm{H}} = x_{\bar{\mathrm{H}}}$. A similar line of thought can be used to motivate the part stating that $Pr(\mathrm{X}|\mathrm{H}_s, \bar{\mathrm{H}}_t) = Pr(\mathrm{X}|\bar{\mathrm{H}}_s, \mathrm{H}_t)$ in Equation 7: A relevant structural similarity between the immune systems of rats and humans would not discriminate between the case in which H_s is true and H_t is false and the case in which H_s is false and H_t is true. It would only tell us that it is more likely that both hypothesis are either true or false together than it is that one is true and the other is false.

With all the assumptions and constraints on our model's probability distribution introduced above, we can now explore our model's suitability as an alternative model for analogical inference. The first observation we can make is the following:

Theorem 3.1. *For every Bayesian network with the graph in Figure 6 whose probability distribution Pr satisfies Equations 5, 6, and 7 as well as the constraints in Table 1: If* $Pr(\mathrm{R}_X) > 0$*, then* $c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) > 0$.

Theorem 3.1 does the same job Theorem 2.1 did for Dardashti et al.'s (2019) original model: It shows that the model allows for analogical confirmation if certain assumptions are met. The model thus spells out conditions under which analogical inference qualitatively agrees with Bayesian updating. Note that positive confirmatory impact of E_s on H_t requires evidence E_X about the structural similarity to be taken into account. To this end, also the prior probability of R_X needs to be positive. This is not surprising, since $Pr(\mathrm{R}_X) = 0$ would mean that the source of the evidence E_X for the structural similarity hypothesis X is believed to be absolutely unreliable. Observing the evidence E_X would then tell us nothing more than if we had no access to this piece of evidence at all. (See Theorem 3.3 below for further discussion.)

Another interesting observation we can make is that all of the plausible probabilistic constraints from section 2 for Dardashti et al.'s (2019) original

model are also met by our alternative model:[16]

Theorem 3.2. *The probability distribution Pr of every Bayesian network with the graph in Figure 6 that satisfies Equations 5, 6, and 7 as well as the constraints in Table 1 also satisfies the assumptions expressed by Equations 1 and 4.*

Finally, and most importantly, our alternative model does not give rise to the problematic Theorems 2.3 and 2.4. Instead, it naturally captures the intuitions about analogical inference discussed in section 2 which stand in conflict with Theorems 2.3 and 2.4. These intuitions are captured by the following observation:

Theorem 3.3. *For every Bayesian network with the graph in Figure 6 whose probability distribution Pr satisfies Equations 5, 6, and 7 as well as the constraints in Table 1:*

(a) $c(\mathrm{H}_t;\mathrm{E}_s) = 0$.

(b) If $Pr(\mathrm{R}_X) = 1$, *then* $c(\mathrm{H}_t;\mathrm{E}_s|\mathrm{E}_X) = max$.

(c) For every probability distribution Pr^* *resulting from Pr by increasing* $Pr(\mathrm{R}_X)$ *while keeping all other parameters fixed:* $c^*(\mathrm{H}_t;\mathrm{E}_s|\mathrm{E}_X) > c(\mathrm{H}_t;\mathrm{E}_s|\mathrm{E}_X)$.

Theorem 3.3(a) together with Theorem 3.1 captures the basic intuition underlying the sort of analogical inference instantiated by our example on model organisms to the effect that analogical inference should be based on some evidence about the similarity of the source and the target system. In terms of the model, this means that E_s should confirm H_t to some extent only if one

[16] We avoided the assumption expressed by Equation 3 in Theorem 3.2 since it is one of the basic assumptions our model shares with Dardashti et al.'s (2019) and, thus, does not need to be proven.

has access to evidence E_X about the similarity and the source of the evidence is believed to be somewhat reliable (i.e., $Pr(R_X) > 0$), while there should be no confirmatory impact on H_t if E_X is not considered at all. This is also in line with Crowther et al.'s (2021) observation that analogical inference requires independent support for the analogy.

Theorem 3.3(b) captures another basic intuition discussed in section 2: Believing that the source and the target system are maximally similar w.r.t. some relevant feature—that is in terms of the model conditioning on the independent evidence E_X and assigning maximal reliability to the source providing us with this evidence—should give us as much confirmatory impact of E_s on H_t as possible. Decreasing the degree of reliability of the source to any value smaller than 1 would result in a smaller degree of confirmation.

Finally, Theorem 3.3(c) reflects the basic intuition that increasing the believed degree of similarity between the source and the target system—that is in terms of the model conditioning on E_X and increasing the reliability of the source—, increases the confirmatory impact E_s has on H_t as well. This also nicely matches extrapolation via analogy: The more similar we believe the immune system of rats and the human immune system are, the more impact do the findings in the rat study have on the expected effectiveness of the antiviral compound on humans.

4 Comparing the two models

For easier reference, we will from now on refer to Dardashti et al.'s (2019) original model expanded by the reliability model as shown in Figure 4 as the *common origin model* and to our alternative model with the graph in Figure 6 as the *collider model*.[17] So far, we were mainly concerned with the technical

[17] A collider is a variable with two incoming arrows (cf. the Appendix).

results which showed that there are cases that are more adequately represented by the collider model than by the common origin model. Thus, the common origin model cannot serve as a general model for analogical inference. But what about the collider model? Can this model serve as a general model for analogical inference?

To answer this question, let us discuss whether there are any epistemic reasons for which we should prefer one way or the other to model analogical reasoning. Is there any diversity in the inferential paths underpinning different sorts of analogical reasoning that would justify the adoption of the common origin model vs. the collider model? Why should the variable X, standing for the similarity hypothesis, be treated as a common origin in one case and as a collider in another? We propose that both models have their proper place in scientific reasoning and may be taken to reflect specific epistemological dynamics. Probability (and therefore information) is propagated differently in the two models, and this may reflect the epistemic inputs and outputs in such diverse contexts.

From now on, we will look at two different interpretations of the similarity hypothesis X which we will label X_1 and X_2.[18] Let us start with having another look at the rat study case introduced earlier in order to highlight the inferential patterns associated with such distinct probability kinematics. So far, we were following the traditional literature on analogical inference and understood the similarity hypothesis X as follows: X_1 says that the immune systems of rats and humans are similar w.r.t. a certain feature. We assumed that this feature is relevant for the truth of the hypotheses H_s and H_t, but there might be other

[18]We would like to emphasize that this paper is not intended as a critique of Dardashti et al.'s (2019) application of their model to Hawking radiation. We thus leave it open under which type of interpretation their particular understanding of X as representing universality arguments falls.

features only realised by one of the two systems or not relevant for the truth of the hypotheses H_s and H_t. As we argued in sections 2 and 3, this understanding of the similarity hypothesis is best represented by the collider model.

However, there is another interpretation of the similarity hypothesis X that is, so we will argue, better represented by the common origin model. In this interpretation, X_2 states a theoretical model or general theory T about immune systems. Here we assume that rats and humans both fall in the domain of T. T might be a causal model or a set of equations or functions, but it might also be a full-fledged theory stating a set of strict or probabilistic laws.

The main difference between X_1 and X_2 is that X_2 requires much more than X_1 does. X_1 only requires the immune systems of rats and humans to be similar w.r.t. a feature relevant for the truth of H_s and H_t. It resembles a black box: Though we know some of the features of these systems, we have no information about how they are related (by laws, equations, causal relations, etc.). Thus, we cannot draw any conclusions about whether the hypotheses H_s and H_t are more likely to be true or false from the similarity of s and t w.r.t. the feature under consideration. We do only know that rats are more likely to react to the antiviral compound in the same way as humans do. Thus, the more similar we believe s and t to be w.r.t. the feature of interest, i.e., the more $Pr(X_1)$ approaches 1, the more impact observing evidence for the effectiveness in rats will have on the likelihood of the antiviral compound to be effective in humans, which is exactly what the collider model tells us.

X_2, on the other hand, gives us not only information about the presence of features relevant for the truth of the hypotheses about rats and humans, but also tells us how these factors are connected to each other (and to other factors). It resembles a white box: We have information about how the different features present in the immune systems of rats and humans are related to each

other (by laws, equations, causal relations, etc.). Thus, we can make inferences about how the antiviral compound will work in humans based on X_2. These inferences can be strict or probabilistic, depending on the specific model or theory (e.g., structural equation model without vs. structural equation model with error terms, classical physics vs. pharmacology, etc.).

Let us assume that we indeed have independent evidence E_{X_2} for X_2 which postulates a theory that covers different domains (such as s and t), and that we are not absolutely certain about X_2 (i.e., $Pr(\mathrm{X}_2)$ is not extreme). Thus, we can make predictions about whether the antiviral compound will work on humans based on such a theory. If we do the rat case study in addition and the result agrees with our prediction on the basis of the theory, this will give us even more confidence in our prediction. Thus, conditioning on E_s will always increase the confirmatory impact on H_t we already get from our confidence in the truth of the theory being shared by both domains (represented by conditioning on the independent evidence E_{X_2} together with a non-extreme probability $Pr(\mathrm{R}_{X_2})$). However, the more certain we become about the truth of X_2 (i.e., the more we increase $Pr(\mathrm{R}_{X_2})$), the stronger our confidence about H_t becomes. As a consequence, an independent study on a model organism will not give us that much of a boost in confirmation of H_t anymore. Thus, it becomes plausible that at some point an increase in certainty that the theory postulated by X_2 is correct goes hand in hand with a decrease of the confirmatory impact E_s has on H_t, which is exactly what Theorem 2.4 tells us about the common origin model.

Let us turn to the extreme case next: If we had certain knowledge that the theory is true, then we would have all the information we can ever hope to get for whether the antiviral compound works on humans. Carrying out the rat study would not give us any additional information about H_t in that case. The same would be true if we condition on $\bar{\mathrm{X}}_2$. Since postulating the theoretical

model or theory was all that linked the source and the target system together, these systems can be expected to be independent if the theory (or the fact that it covers both domains) is rejected with certainty. These two observations are exactly what Theorem 2.3(b) states for the common origin model: Fixing X_2 to any of its values will render H_t independent of E_s. Thus, the epistemic situation where we have some reason to believe that a specific theoretical model or full-fledged theory applies to diverse fields of reality is better captured by the common origin model than by the collider model.

Before we move on, we need to highlight one caveat: While confirmation via a theoretical model or theory agrees with Theorems 2.3(b) and 2.4 and, thus, nicely fits the epistemic situation represented by the common origin model, Theorem 2.3(a) is still problematic. The latter stands in contrast to Crowther et al.'s (2021) observation that evidence about the source system should be able to confirm the hypothesis about the target system only if the similarity hypothesis, that is X_2 in our example, has some independent support. In terms of the model this means that E_s should confirm H_t only conditional on E_{X_2}. We see no easy way to fix this problem in terms of the model and, thus, propose to follow the suggestion already discussed in section 2: Restrict the domain of application of the common origin model to cases where independent evidence E_{X_2} is available.

Let us now once again come back to the collider model. Can we say more about the epistemic situations adequately represented by it? The collider model seems to provide a valuable way to exemplify cases of extrapolation via analogy and extrapolation based on sampling assumptions such as the extrapolation of results from *in vivo* studies (experiments on animals) to humans, on the basis of some shared features, or the extrapolation of evidence from a study population to a target population based on statistical considerations (sampling procedures,

design protocols, etc.).[19]

Leaving aside cases of extrapolation exclusively relying on statistical considerations, and focusing on extrapolation via analogy, we observe that, in this case, the similarity between model and target is given by nature on the basis of a more or less loose set of shared features. Knowledge about similarity is typically incomplete and exact theoretical (or empirical) mechanisms may be inaccessible. What matters is that evidence about a specific process or an outcome in a given setting legitimates the idea that something similar may plausibly be considered to be at play elsewhere. In extrapolation via analogy, similarity is based on partial knowledge of both systems: The more we know about the features of the source and the target system, the more we learn about their similarity and, through such similarity, the more the two become epistemically dependent and relevant to each other.

We propose that our distinction between confirmation via theoretical model or theory vs. extrapolation via analogy, based on probability kynematics, can be considered as an ideal criterion for separating the two. Theoretical models are obviously affected by many uncertainties, equivocation, and noise (cf. Suárez & Bolinska, 2021). However, we think that the following qualitative features give further support to our distinction:

[19] Justification grounded on sampling procedures relies on the causal mechanisms underpinning the sampling process, i.e., on how the statistical units are selected from the target population and get included in the sample. Instead, in the phylogenetic case, extrapolation is justified because the model organism is taken to derive from a branch of the same tree to which the target organism belongs to (so the causal structure regards here the phylogenetic relations among species and taxa) (see also Levy & Currie, 2015). Although the former case can, strictly speaking, not be considered as an example of analogical reasoning, it shares with the latter the idea that sampling from a population/species (e.g., mammals) may give us plausible reasons for thinking that biological or other mechanisms observed in the sample may be at work also in the larger population, and therefore in other taxa of the same species.

1. Whereas X_2 makes a theoretical postulate enjoying some independent evidential support, X_1 states more or less firmly established facts.

2. Consequently, the mapping between components and relations of the source and target system relevant for the analogy is explicit in X_2 (white box), whereas it tends to be ambiguous in X_1 (black box): It is impossible *in principle* to measure the noise and the equivocation of the model with respect to the target (Suárez & Bolinska, 2021; Bolinska, 2013).

3. To draw on Goodman's (1976) distinction between notational and dense systems of representation, one could see X_1 and X_2 as representing two qualitatively distinct ways to provide an inferential basis for transferring information associated with H_s to H_t. The common origin model based on a given law or set of laws resembles the way ideal notational systems ground the assignment of relata to notational signs (e.g., musical score). It can be thought as a white box in the terms mentioned above. Vice versa, in dense systems (e.g., painting and visual arts in general), neither signs nor their relata are discrete entities, let alone connected by functions or laws. Hence their interpretation is grounded on empirical facts or associations (black box).

In general, in the case of model organisms the similarity between the source and the target system is presumed on the basis of empirical knowledge and the joint implications of several overarching theories (biology, physiology, chemistry, etc.), which are not directly under test, and only indirectly contribute to generate the hypothesis of similarity. Also the scope of the inference differs in the two settings: Whereas in confirmation via theoretical model or theory the inference regards the behaviour of the system represented by the theoretical model or theory as a whole, in the case of extrapolation by analogy the inference refers only to specific (black box) mechanisms observed in the model organism and ex-

trapolated to another species. All the other mechanisms and inner workings of the two systems may not be directly accessible and transparent to the modeller. Hence, accruing evidence about the similarity of the two systems in general may provide further support to the hypothesis that the mechanism observed in the model also holds in the target.[20]

Summarizing, the two models formalize circumstances where analogical inference is grounded in different inputs, outputs, and types of (indirect) evidence. The common origin model represents cases where we have reasons to assume that the two systems share a richer common theoretical structure (in the form of laws, equations, etc.) which can be used to generate predictions about the truth of H_s and H_t. These reasons may, for example, stem from evidence about the similarity of the *phenomenological* behaviour of the two systems (see, e.g., the use of simulations in systems biology, Osimani & Poellinger, 2020). In the collider model, on the other hand, the hypothesis of similarity typically consolidates via the discovery of features shared between the two domains. This typically does not involve a deeper understanding of the underlying laws relevant for the truth of H_s and H_t. It can, however, in turn strengthen the hypothesis that the target system shares also other features, which have been empirically established only in the source system.

5 Corollary: Breaking the extrapolator's circle

An important consequence of the previous observations is that the collider model can help to break the so-called extrapolator's circle (Steel, 2007). This sort of epistemic paradox is pervasive in evidence-based policy and medicine and

[20]This feature is shared with extrapolation via sampling assumptions, where results in a study population are transferred to the general population, on the basis of methodological features such as research design, study protocol, and sampling procedures.

regards the fact that in order to apply evidence about causal effects observed in a source domain to another target domain, one needs to be confident enough about the fact that the latter is sufficiently similar in a causally relevant way to the former (cf. Cartwright, 2012; Cartwright & Hardie, 2012; Reiss, 2010). However, in order to know that this assumption is met, one would need to have sufficient causal knowledge about the target setting, which would make the extrapolation redundant. There has been several attempts to solve the circle, all relying on partial knowledge about the source and target structures complemented by indirect evidence about the similarity assumption itself (Steel, 2007). Khosrowi (2019) in particular advances the importance of integrating qualitative evidence about the similarity of the two systems at stake, i.e., our E_X, with quantitative methods developed in econometrics to account for the role of interactive covariates (Hotz, Imbens, & Mortimer, 2005; Muller, 2015).[21]

These proposals are based on pragmatic-methodological considerations:

> "If qualitative evidence increases our confidence that crucial features of causal mechanisms are qualitatively similar between populations [...] this can offer important support (although perhaps not full-fledged warrant) for the assumptions that are necessary for interactive covariate-based extrapolation to proceed. If this is successful, interactive covariate-based strategies may justifiably be used to obtain quantitative prediction of causal effects." (Khosrowi, 2019)

Our Bayesian formalization of extrapolation (via analogy) provides the required epistemic justificatory underpinning for this sort of inferential procedures, in that, by allowing probabilistic assessment (and updates) about the similarity of the two systems itself, it does not require full knowledge of the

[21] Causal factors that mediate the intensity of the causal effect of the treatment variable in different ways in different settings.

relevant causal mechanisms to get the extrapolation going, but at the same time it grounds the inferential leap on a consolidated approach to probabilistic inference. Furthermore, it shows how independent evidence E_X about X and its degree of reliability $Pr(\mathrm{R}_X)$ may contribute to updating H_t.

6 Conclusion

We started this paper with the question of whether Dardashti et al.'s (2019) model for establishing Hawking radiation by analogical inference can be understood as a general Bayesian model for analogical inference. In section 2 it turned out that there are situations in which the model fails to capture important basic intuitions about certain types of analogical inference. In section 3 we then proposed an alternative model that can take these basic intuitions into account. In section 4 it turned out that also our alternative model falls short of being a general Bayesian model for analogical inference. Actually, there are different types of analogical inference and different epistemic situations that require a different formal treatment in terms of these models.

The interpretation of X, that is, the epistemic basis of the similarity assumption, determines the justificatory structure of the inference, and, in turn, the topology of the Bayesian network. We formally vindicate Levy and Currie's (2015) distinction between theoretical models and model organisms (*pace* Weisberg, 2012), furthermore we provide a structure that allows to clearly explicate the role played by independent evidence about the similarity hypothesis X, and also by knowledge about the reliability of the source of such evidence (e.g., sampling procedures, randomization, and design protocols more generally).

The Bayesian formalization is not only interesting for illustrating the updating dynamics and implications at work in different evidential contexts. It is also interesting because the network topology sheds light on the kind of inference

at place and the related justificatory structure. Modeling diverse types of analogical inference through different Bayesian networks points to further avenues of research focusing on the interplay between semantic, ontological, and epistemic dimensions of scientific inferences (see also Godfrey-Smith, 2006; Bueno, 2021). Finally, Bayesian models of analogical inference (and especially the collider model) might also be useful for breaking the extrapolation circle (Steel, 2007). Further exploring these issues is a topic for future research.

Appendix

Bayesian networks

A Bayesian network is a structure $\langle \mathbf{V}, \mathbf{E}, Pr \rangle$. $\mathbf{V}$ is a set of random variables $X_1, ..., X_n$, $\mathbf{E}$ is a binary relation on $\mathbf{V}$, where $\langle X_i, X_j \rangle \in \mathbf{E}$ can be represented graphically by a directed edge $X_i \longrightarrow X_j$. Any concatenation of directed edges (regardless of their direction) connecting two variables X_i and X_j is called a path between X_i and X_j. $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ is assumed to be a directed acyclic graph (DAG), meaning that $\mathbf{G}$ does not feature paths of the form $X_i \longrightarrow ... \longrightarrow X_i$. Finally, Pr is a probability distribution over $\mathbf{V}$. For a structure $\langle \mathbf{V}, \mathbf{E}, Pr \rangle$ to count as a Bayesian network, it is required to satisfy the Markov factorization

$$Pr(x_1, ..., x_n) = \prod_{i=1}^{n} Pr(x_i | \mathbf{par}(X_i)),$$

where $x_1, ..., x_n$ are individual variables ranging over the possible values of $X_1, ..., X_n$ and $\mathbf{par}(X_i)$ stands for the instantiation of $\mathbf{Par}(X_i)$ to the values appearing on the left hand side of the equality sign. $\mathbf{Par}(X_i)$ is the set of X_i's parents which consists of all the variables $X_j \in \mathbf{V}$ for which $X_j \longrightarrow X_i$ holds.

Bayesian networks do not add anything that could not also be expressed in terms of probabilities alone, but they can often simplify probabilistic reasoning

and help in visualizing constraints on one's probability distribution. By looking at a Bayesian network's graph one can, for example, read off some of the probabilistic independence relations implied by the network's structure. In particular, a structure $\langle \mathbf{V}, \mathbf{E}, Pr \rangle$ satisfying the Markov factorization is equivalent with it satisfying the following condition as well:

Definition 6.1 (Markov condition)**.** *$\langle \mathbf{V}, \mathbf{E}, Pr \rangle$ satisfies the Markov condition if and only if every $X_i \in \mathbf{V}$ is probabilistically independent of its non-descendants conditional on its parents.*

A variable X_i's non-descendants are all the variables in $\mathbf{V}$ which are not among its descendants. A variable X_i's descendants are all those variables $X_j \in \mathbf{V}$ for which $X_i \longrightarrow ... \longrightarrow X_j$ holds. For technical reasons, X_i is considered to be its own descendant.

For DAGs, the Markov condition is equivalent with the *d*-separation criterion (Pearl, 2000, sec. 1.2.3):

Definition 6.2 (*d*-separation criterion)**.** *$\langle \mathbf{V}, \mathbf{E}, Pr \rangle$ satisfies the d-separation criterion if and only if for all $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ (with $\mathbf{X} \cap \mathbf{Y} = \varnothing$, $\mathbf{X} \cap \mathbf{Z} = \varnothing$, and $\mathbf{Y} \cap \mathbf{Z} = \varnothing$): If $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$, then $\mathbf{X}$ and $\mathbf{Y}$ are probabilistically independent conditional on $\mathbf{Z}$.*

Definition 6.3 (*d*-separation)**.** *Let $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ be a DAG and $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ (with $\mathbf{X} \cap \mathbf{Y} = \varnothing$, $\mathbf{X} \cap \mathbf{Z} = \varnothing$, and $\mathbf{Y} \cap \mathbf{Z} = \varnothing$). Then $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are not d-connected given $\mathbf{Z}$.*

Definition 6.4 (*d*-connection)**.** *Let $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ be a DAG and $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ (with $\mathbf{X} \cap \mathbf{Y} = \varnothing$, $\mathbf{X} \cap \mathbf{Z} = \varnothing$, and $\mathbf{Y} \cap \mathbf{Z} = \varnothing$). Then $\mathbf{X}$ and $\mathbf{Y}$ are d-connected given $\mathbf{Z}$ if and only if there is an $X \in \mathbf{X}$ and a $Y \in \mathbf{Y}$ connected by a path π such that:*

(a) All non-colliders on π are not in $\mathbf{Z}$, and

(b) all colliders on π are in $\mathbf{Z}$ or have a descendent in $\mathbf{Z}$.

A collider on a path π is a variable X such that $\longrightarrow X \longleftarrow$ is a part of π. The concepts of d-connection and d-separation are useful because they allow one to determine whether two sets of variables can be probabilistically dependent conditional on a third set of variables by looking at the graph's structure without any need to do the corresponding probabilistic calculations. Finally, two graphs that share the same d-connections and d-separations can form a Bayesian network together with the same probability distributions. We refer to such graphs as probabilistically indistinguishable:

Definition 6.5 (probabilistic indistinguishability)**.** *Two DAGs $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ and $\mathbf{G'} = \langle \mathbf{V'}, \mathbf{E'} \rangle$ are probabilistically indistinguishable if and only if $\mathbf{V} = \mathbf{V'}$ and all subsets of $\mathbf{V}$ and $\mathbf{V'}$ share the same d-connection and d-separation relations.*

Proofs

For easier reference, we use the following notation:

$x \coloneqq Pr(\mathrm{X})$	$\bar{x} \coloneqq Pr(\bar{\mathrm{X}})$
$x_{\mathrm{H}_s\mathrm{H}_s} \coloneqq Pr(\mathrm{X}\|\mathrm{H}_s, \mathrm{H}_t)$	$\bar{x}_{\mathrm{H}_s\mathrm{H}_s} \coloneqq Pr(\bar{\mathrm{X}}\|\mathrm{H}_s, \mathrm{H}_t)$
$x_{\mathrm{H}_s\bar{\mathrm{H}}_s} \coloneqq Pr(\mathrm{X}\|\mathrm{H}_s, \bar{\mathrm{H}}_t)$	$\bar{x}_{\mathrm{H}_s\bar{\mathrm{H}}_s} \coloneqq Pr(\bar{\mathrm{X}}\|\mathrm{H}_s, \bar{\mathrm{H}}_t)$
$x_{\bar{\mathrm{H}}_s\mathrm{H}_s} \coloneqq Pr(\mathrm{X}\|\bar{\mathrm{H}}_s, \mathrm{H}_t)$	$\bar{x}_{\bar{\mathrm{H}}_s\mathrm{H}_s} \coloneqq Pr(\bar{\mathrm{X}}\|\bar{\mathrm{H}}_s, \mathrm{H}_t)$
$x_{\bar{\mathrm{H}}_s\bar{\mathrm{H}}_s} \coloneqq Pr(\mathrm{X}\|\bar{\mathrm{H}}_s, \bar{\mathrm{H}}_t)$	$\bar{x}_{\bar{\mathrm{H}}_s\bar{\mathrm{H}}_s} \coloneqq Pr(\bar{\mathrm{X}}\|\bar{\mathrm{H}}_s, \bar{\mathrm{H}}_t)$
$h_s \coloneqq Pr(\mathrm{H}_s)$	$\bar{h}_s \coloneqq Pr(\bar{\mathrm{H}}_s)$
$h_{s\mathrm{X}} \coloneqq Pr(\mathrm{H}_s\|\mathrm{X})$	$\bar{h}_{s\mathrm{X}} \coloneqq Pr(\bar{\mathrm{H}}_s\|\mathrm{X})$
$h_{s\bar{\mathrm{X}}} \coloneqq Pr(\mathrm{H}_s\|\bar{\mathrm{X}})$	$\bar{h}_{s\bar{\mathrm{X}}} \coloneqq Pr(\bar{\mathrm{H}}_s\|\bar{\mathrm{X}})$
$h_t \coloneqq Pr(\mathrm{H}_t)$	$\bar{h}_t \coloneqq Pr(\bar{\mathrm{H}}_t)$
$h_{t\mathrm{X}} \coloneqq Pr(\mathrm{H}_t\|\mathrm{X})$	$\bar{h}_{t\mathrm{X}} \coloneqq Pr(\bar{\mathrm{H}}_t\|\mathrm{X})$
$h_{t\bar{\mathrm{X}}} \coloneqq Pr(\mathrm{H}_t\|\bar{\mathrm{X}})$	$\bar{h}_{t\bar{\mathrm{X}}} \coloneqq Pr(\bar{\mathrm{H}}_t\|\bar{\mathrm{X}})$
$e_{s\mathrm{H}_s} \coloneqq Pr(\mathrm{E}_s\|\mathrm{H}_s)$	$\bar{e}_{s\mathrm{H}_s} \coloneqq Pr(\bar{\mathrm{E}}_s\|\mathrm{H}_s)$
$e_{s\bar{\mathrm{H}}_s} \coloneqq Pr(\mathrm{E}_s\|\bar{\mathrm{H}}_s)$	$\bar{e}_{s\bar{\mathrm{H}}_s} \coloneqq Pr(\bar{\mathrm{E}}_s\|\bar{\mathrm{H}}_s)$
$r_X \coloneqq Pr(\mathrm{R}_X)$	$\bar{r}_X \coloneqq Pr(\bar{\mathrm{R}}_X)$
$e_{X\mathrm{X}\mathrm{R}_X} \coloneqq Pr(\mathrm{E}_X\|\mathrm{X}, \mathrm{R}_X)$	$\bar{e}_{X\mathrm{X}\mathrm{R}_X} \coloneqq Pr(\bar{\mathrm{E}}_X\|\mathrm{X}, \mathrm{R}_X)$
$e_{X\mathrm{X}\bar{\mathrm{R}}_X} \coloneqq Pr(\mathrm{E}_X\|\mathrm{X}, \bar{\mathrm{R}}_X)$	$\bar{e}_{X\mathrm{X}\bar{\mathrm{R}}_X} \coloneqq Pr(\bar{\mathrm{E}}_X\|\mathrm{X}, \bar{\mathrm{R}}_X)$
$e_{X\bar{\mathrm{X}}\mathrm{R}_X} \coloneqq Pr(\mathrm{E}_X\|\bar{\mathrm{X}}, \mathrm{R}_X)$	$\bar{e}_{X\bar{\mathrm{X}}\mathrm{R}_X} \coloneqq Pr(\bar{\mathrm{E}}_X\|\bar{\mathrm{X}}, \mathrm{R}_X)$
$e_{X\bar{\mathrm{X}}\bar{\mathrm{R}}_X} \coloneqq Pr(\mathrm{E}_X\|\bar{\mathrm{X}}, \bar{\mathrm{R}}_X)$	$\bar{e}_{X\bar{\mathrm{X}}\bar{\mathrm{R}}_X} \coloneqq Pr(\bar{\mathrm{E}}_X\|\bar{\mathrm{X}}, \bar{\mathrm{R}}_X)$

Proof Theorem 2.1.

Theorem 2.1 was proven by Dardashti et al. (2019, Theorem 1). Since we will rely on some steps of Dardashti et al.'s proof later on, we briefly reproduce

it here. We need to show that $c(\mathrm{H}_t;\mathrm{E}s) > 0$ follows from the assumption in Theorem 2.1. To this end, we show that

$$\Delta := \frac{Pr(\mathrm{H}_t, \mathrm{E}_s)}{Pr(\mathrm{E}_s)} - Pr(\mathrm{H}_t) > 0.$$

The probabilities figuring in this inequality can be computed as follows, where α, β are defined as:

$$\alpha := h_{s\mathrm{X}} e_{s\mathrm{H}_s} + \bar{h}_{s\mathrm{X}} e_{s\bar{\mathrm{H}}_s}$$
$$\beta := h_{s\bar{\mathrm{X}}} e_{s\mathrm{H}_s} + \bar{h}_{s\bar{\mathrm{X}}} e_{s\bar{\mathrm{H}}_s}$$

$$\begin{aligned} Pr(\mathrm{H}_t, \mathrm{E}_s) &= \sum_{X, H_s} Pr(X, \mathrm{H}_t, H_s, \mathrm{E}_s) \\ &= \sum_{X, H_s} Pr(X) Pr(\mathrm{H}_t|X) Pr(H_s|X) Pr(\mathrm{E}_s|H_s) \\ &= x h_{t\mathrm{X}} (h_{s\mathrm{X}} e_{s\mathrm{H}_s} + \bar{h}_{s\mathrm{X}} e_{s\bar{\mathrm{H}}_s}) + \bar{x} h_{t\bar{\mathrm{X}}} (h_{s\bar{\mathrm{X}}} e_{s\mathrm{H}_s} + \bar{h}_{s\bar{\mathrm{X}}} e_{s\bar{\mathrm{H}}_s}) \\ &= x h_{t\mathrm{X}} \alpha + \bar{x} h_{t\bar{\mathrm{X}}} \beta \end{aligned} \tag{8}$$

$$\begin{aligned} Pr(\mathrm{E}_s) &= \sum_{X, H_s} Pr(X, H_s, \mathrm{E}_s) \\ &= \sum_{X, H_s} Pr(X) Pr(H_s|X) Pr(\mathrm{E}_s|H_s) \\ &= x\alpha + \bar{x}\beta \end{aligned} \tag{9}$$

$$\begin{aligned} Pr(\mathrm{H}_t) &= \sum_{X} Pr(X) Pr(\mathrm{H}_t|X) \\ &= x h_{t\mathrm{X}} + \bar{x} h_{t\bar{\mathrm{X}}} \end{aligned} \tag{10}$$

With Equations 8-10 we obtain

$$\begin{aligned}\Delta &= \frac{x h_{t\mathrm{X}}\alpha + \bar{x} h_{t\bar{\mathrm{X}}}\beta}{x\alpha + \bar{x}\beta} - x h_{t\mathrm{X}} + \bar{x} h_{t\bar{\mathrm{X}}} \\ &= \frac{x\bar{x}(h_{t\mathrm{X}} - h_{t\bar{\mathrm{X}}})(h_{s\mathrm{X}} - h_{s\bar{\mathrm{X}}})(e_{s\mathrm{H}} - e_{s\bar{\mathrm{H}}})}{x\alpha + \bar{x}\beta}. \end{aligned} \tag{11}$$

By Equations 1-3, all simple terms in the fraction are greater than 1 and also the results of the subtractions in the numerator are greater than 1. Thus, it follows that $\Delta > 0$. □

Proof Theorem 2.2.

We need to show that $c(\mathrm{H}_t; \mathrm{E}s) > 0$ also follows from the assumptions made in Theorem 2.1 if $Pr(\mathrm{H}_i|\mathrm{X}) > Pr(\mathrm{H}_i|\bar{\mathrm{X}})$ is replaced by $Pr(\mathrm{H}_i|\mathrm{X}) < Pr(\mathrm{H}_i|\bar{\mathrm{X}})$ (for $i \in \{s, t\}$). This follows directly from Equation 11: Again, all simple terms in the fraction are greater than 1. But this time, $(h_{t\mathrm{X}} - h_{t\bar{\mathrm{X}}})$ and $(h_{s\mathrm{X}} - h_{s\bar{\mathrm{X}}})$ are negative, while $(e_{s\mathrm{H}} - e_{s\bar{\mathrm{H}}})$ is still positive. Thus, it follows that $\Delta > 0$. □

Proof Theorem 2.3.

(a) Since marginalizing out $\{R_X, E_X\}$ from a Bayesian network with the graph in Figure 4 results in a Bayesian network with the graph in Figure 2, the probability distribution of the latter will be the probability distribution of the former restricted to $\{X, H_t, H_s, E_s\}$. Thus, $c(\mathrm{H}_t; \mathrm{E}_s) > 0$ is a direct consequence of Theorem 2.2.

(b) We need to show that $c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) = 0$ if the assumptions in Theorem 2.3 hold and $Pr(\mathrm{R}_X) = 1$. To this end, it suffices to show that

$$\Delta^* := \frac{Pr^*(\mathrm{H}_t, \mathrm{E}_s)}{Pr^*(\mathrm{E}_s)} - Pr^*(\mathrm{H}_t) = 0,$$

where $Pr^*(\cdot) = Pr(\cdot|\mathrm{E}_X)$. As a first step, we show that $Pr^*(\mathrm{X}) = 1$ follows from

our assumptions. To this end, we compute $Pr(\mathrm{X}|\mathrm{E}_X)$ as follows:

$$\begin{aligned} Pr(\mathrm{X}|\mathrm{E}_X) &= \frac{Pr(\mathrm{E}_X|\mathrm{X})Pr(\mathrm{X})}{Pr(\mathrm{E}_X)} \\ &= \frac{\sum_{R_X} Pr(\mathrm{E}_X|\mathrm{X}, R_X)Pr(R_X)Pr(\mathrm{X})}{\sum_{X,R_X} Pr(\mathrm{E}_X|X, R_X)Pr(R_X)Pr(X)} \\ &= \frac{e_{X\mathrm{X}\mathrm{R}_X} r_X x + e_{X\mathrm{X}\bar{\mathrm{R}}_X} \bar{r}_X x}{e_{X\mathrm{X}\mathrm{R}_X} r_X x + e_{X\mathrm{X}\bar{\mathrm{R}}_X} \bar{r}_X x + e_{X\bar{\mathrm{X}}\mathrm{R}_X} r_X \bar{x} + e_{X\bar{\mathrm{X}}\bar{\mathrm{R}}_X} \bar{r}_X \bar{x}} \end{aligned} \tag{12}$$

By assumption, $r_X = 1$. Thus, Equation 12 transforms to

$$Pr(\mathrm{X}|\mathrm{E}_X) = \frac{e_{X\mathrm{X}\mathrm{R}_X} x + e_{X\mathrm{X}\bar{\mathrm{R}}_X} 0x}{e_{X\mathrm{X}\mathrm{R}_X} x + e_{X\mathrm{X}\bar{\mathrm{R}}_X} 0x + e_{X\bar{\mathrm{X}}\mathrm{R}_X} \bar{x} + e_{X\bar{\mathrm{X}}\bar{\mathrm{R}}_X} 0\bar{x}}. \tag{13}$$

Since the conditional probabilities of a reliability model's parameters are specified as in Table 1, Equation 13 transforms to

$$Pr(\mathrm{X}|\mathrm{E}_X) = \frac{1x + 0x}{1x + 0x + 0\bar{x} + 0\bar{x}} = 1.$$

From Equation 11 we know that

$$\Delta^* = \frac{x^*\bar{x}^*(h_{t\mathrm{X}} - h_{t\bar{\mathrm{X}}})(h_{s\mathrm{X}} - h_{s\bar{\mathrm{X}}})(e_{s\mathrm{H}} - e_{s\bar{\mathrm{H}}})}{x^*\alpha + \bar{x}^*\beta}, \tag{14}$$

where α, β are defined as in the proof of Theorem 2.1. Since $x^* = Pr^*(\mathrm{X}) = Pr(\mathrm{X}|\mathrm{E}_X) = 1$, Equation 14 reduces to

$$\Delta^* = \frac{0(h_{t\mathrm{X}} - h_{t\bar{\mathrm{X}}})(h_{s\mathrm{X}} - h_{s\bar{\mathrm{X}}})(e_{s\mathrm{H}} - e_{s\bar{\mathrm{H}}})}{x^*\alpha + \bar{x}^*\beta}. \tag{15}$$

Since the numerator in Equation 15 is 0, it follows that also $\Delta^* = 0$.[22] □

Proof Theorem 2.4.

The following table fully specifies a probability distribution Pr for the Bayesian

[22] We assume that $\Delta^* = 0$ if the denominator in Equation 15 is 0.

network in Figure 4 that satisfies all of the assumptions in Theorem 2.4. Again, we assume that $i \in \{s, t\}$:

$Pr(\mathrm{E}_s\|\mathrm{H}_s)$	$Pr(\mathrm{E}_s\|\bar{\mathrm{H}}_s)$	$Pr(\mathrm{H}_i\|\mathrm{X})$	$Pr(\mathrm{H}_i\|\bar{\mathrm{X}})$	a	$Pr(\mathrm{X})$
0.9	0.1	0.9	0.1	0.5	0.1

Varying $Pr(\mathrm{R}_X)$ while keeping all the other probabilities in this table fixed produces the curve shown in Figure 5, where $c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) = Pr(\mathrm{H}_t|\mathrm{E}_s, \mathrm{E}_X) - Pr(\mathrm{H}_t|\mathrm{E}_X)$ can be computed as follows:[23]

$$
\begin{aligned}
&Pr(\mathrm{H}_t|\mathrm{E}_s, \mathrm{E}_X) = \frac{Pr(\mathrm{H}_t, \mathrm{E}_s, \mathrm{E}_X)}{Pr(\mathrm{E}_s, \mathrm{E}_X)} \\
&= \frac{\sum_{X,R_X,H_s} Pr(\mathrm{H}_t|X)Pr(\mathrm{E}_s|H_s)Pr(H_s|X)Pr(\mathrm{E}_X|X,R_X)Pr(X)Pr(R_X)}{\sum_{X,R_X,H_s} Pr(\mathrm{E}_s|H_s)Pr(H_s|X)Pr(\mathrm{E}_X|X,R_X)Pr(X)Pr(R_X)} \\
&= \frac{h_{\mathrm{X}} e_{s\mathrm{H}_s} h_{\mathrm{X}} x r_X + h_{\mathrm{X}} e_{s\bar{\mathrm{H}}_s} \bar{h}_{\mathrm{X}} x r_X + h_{\mathrm{X}} e_{s\mathrm{H}_s} h_{\mathrm{X}} a x \bar{r}_X + h_{\mathrm{X}} e_{s\bar{\mathrm{H}}_s} \bar{h}_{\mathrm{X}} a x \bar{r}_X +}{e_{s\mathrm{H}_s} h_{\mathrm{X}} x r_X + e_{s\bar{\mathrm{H}}_s} \bar{h}_{\mathrm{X}} x r_X + e_{s\mathrm{H}_s} h_{\mathrm{X}} a x \bar{r}_X + e_{s\bar{\mathrm{H}}_s} \bar{h}_{\mathrm{X}} a x \bar{r}_X + } \\
&\frac{h_{\bar{\mathrm{X}}} e_{s\mathrm{H}_s} h_{\bar{\mathrm{X}}} a \bar{x} \bar{r}_X + h_{\bar{\mathrm{X}}} e_{s\bar{\mathrm{H}}_s} \bar{h}_{\bar{\mathrm{X}}} a \bar{x} \bar{r}_X}{e_{s\mathrm{H}_s} h_{\bar{\mathrm{X}}} a \bar{x} \bar{r}_X + e_{s\bar{\mathrm{H}}_s} \bar{h}_{\bar{\mathrm{X}}} a \bar{x} \bar{r}_X}
\end{aligned}
$$

$$
\begin{aligned}
Pr(\mathrm{H}_t|\mathrm{E}_X) &= \frac{Pr(\mathrm{H}_t, \mathrm{E}_X)}{Pr(\mathrm{E}_X)} \\
&= \frac{\sum_{X,R_X} Pr(\mathrm{H}_t|X)Pr(\mathrm{E}_X|X,R_X)Pr(\mathrm{X})Pr(R_X)}{\sum_{X,R_X} Pr(\mathrm{E}_X|X,R_X)Pr(X)Pr(R_X)} \\
&= \frac{h_{\mathrm{X}} x r_X + h_{\mathrm{X}} a x \bar{r}_X + h_{\bar{\mathrm{X}}} a \bar{x} \bar{r}_X}{x r_X + a x \bar{r}_X + a \bar{x} \bar{r}_X}
\end{aligned}
$$

This shows that there are distributions Pr^* resulting from Pr by increasing $Pr(\mathrm{R}_X)$ while keeping all other parameters of the original Bayesian network fixed such that: $Pr(\mathrm{R}_X) < 1$ and $c^*(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) < c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X) > 0$. □

Proof Theorem 3.1.

[23] Since we assume that $Pr(\mathrm{H}_s) = Pr(\mathrm{H}_t)$, we define $h_{\mathrm{X}} := h_{s\mathrm{X}} = h_{t\mathrm{X}}$.

We need to show that $c(\mathrm{H}_t;\mathrm{E}_s|\mathrm{E}_X) > 0$ if $Pr(\mathrm{R}_X) > 0$ holds under the assumptions made in Theorem 3.1. To this end, it suffices to show that the likelihood ratio

$$l^* := \frac{Pr^*(\mathrm{E}_s|\bar{\mathrm{H}}_t)}{Pr^*(\mathrm{E}_s|\mathrm{H}_t)}$$

is smaller than 1, where

$$Pr^*(\cdot) = Pr(\cdot|\mathrm{E}_X).$$

We can compute the relevant probabilities as follows:[24]

$$\begin{aligned}
Pr^*(\mathrm{E}_s|\bar{\mathrm{H}}_t) &= Pr(\mathrm{E}_s|\bar{\mathrm{H}}_t,\mathrm{E}_X) = \frac{Pr(\bar{\mathrm{H}}_t,\mathrm{E}_s,\mathrm{E}_X)}{Pr(\bar{\mathrm{H}}_t,\mathrm{E}_X)} \\
&= \frac{\sum_{X,R_X,H_s} Pr(\mathrm{E}_s|H_s)Pr(\mathrm{E}_X|X,R_X)Pr(R_X)Pr(X|H_s,\bar{\mathrm{H}}_t)Pr(H_s)Pr(\bar{\mathrm{H}}_t)}{\sum_{X,R_X,H_s} Pr(\mathrm{E}_X|X,R_X)Pr(R_X)Pr(X|H_s,\bar{\mathrm{H}}_t)Pr(H_s)Pr(\bar{\mathrm{H}}_t)} \\
&= \frac{e_{s\mathrm{H}_s} r_X x_{\bar{\mathrm{H}}} h_s \bar{h}_t + e_{s\bar{\mathrm{H}}_s} r_X x_{\mathrm{H}} \bar{h}_s \bar{h}_t + e_{s\mathrm{H}_s} a\bar{r}_X x_{\bar{\mathrm{H}}} h_s \bar{h}_t +}{r_X x_{\bar{\mathrm{H}}} h_s \bar{h}_t + r_X x_{\mathrm{H}} \bar{h}_s \bar{h}_t + a\bar{r}_X x_{\bar{\mathrm{H}}} h_s \bar{h}_t +} \\
&\quad \frac{e_{s\bar{\mathrm{H}}_s} a\bar{r}_X x_{\mathrm{H}} \bar{h}_s \bar{h}_t + e_{s\mathrm{H}_s} a\bar{r}_X \bar{x}_{\bar{\mathrm{H}}} h_s \bar{h}_t + e_{s\bar{\mathrm{H}}_s} a\bar{r}_X \bar{x}_{\mathrm{H}} \bar{h}_s \bar{h}_t}{a\bar{r}_X x_{\mathrm{H}} \bar{h}_s \bar{h}_t + a\bar{r}_X \bar{x}_{\bar{\mathrm{H}}} h_s \bar{h}_t + a\bar{r}_X \bar{x}_{\mathrm{H}} \bar{h}_s \bar{h}_t} \\
&= \frac{\overbrace{r_X(e_{s\mathrm{H}_s} x_{\bar{\mathrm{H}}} h_s + e_{s\bar{\mathrm{H}}_s} x_{\mathrm{H}} \bar{h}_s) + \bar{r}_X a(e_{s\mathrm{H}_s} h_s + e_{s\bar{\mathrm{H}}_s} \bar{h}_s)}^{\alpha :=}}{\underbrace{r_X(x_{\bar{\mathrm{H}}} h_s + x_{\mathrm{H}} \bar{h}_s) + \bar{r}_X a}_{\beta :=}}
\end{aligned} \tag{16}$$

[24] Recall that $x_{\mathrm{H}} = x_{\mathrm{H}_s\mathrm{H}_t} = x_{\bar{\mathrm{H}}_s\bar{\mathrm{H}}_t}$ and $x_{\bar{\mathrm{H}}} = x_{\mathrm{H}_s\bar{\mathrm{H}}_t} = x_{\bar{\mathrm{H}}_s\mathrm{H}_t}$ hold due to Equations 6 and 7.

$$
\begin{aligned}
Pr^*(\mathrm{E}_s|\mathrm{H}_t) &= Pr(\mathrm{E}_s|\mathrm{H}_t,\mathrm{E}_X) = \frac{Pr(\mathrm{H}_t,\mathrm{E}_s,\mathrm{E}_X)}{Pr(\mathrm{H}_t,\mathrm{E}_X)} \\
&= \frac{\sum_{X,R_X,H_s} Pr(\mathrm{E}_s|H_s)Pr(\mathrm{E}_X|X,R_X)Pr(R_X)Pr(X|H_s,\mathrm{H_t})Pr(H_s)Pr(\mathrm{H}_t)}{\sum_{X,R_X,H_s} Pr(\mathrm{E}_X|X,R_X)Pr(R_X)Pr(X|H_s,\mathrm{H}_t)Pr(H_s)Pr(\mathrm{H}_t)} \\
&= \frac{e_{s\mathrm{H}_s} r_X x_\mathrm{H} h_s h_t + e_{s\bar{\mathrm{H}}_s} r_X x_{\bar{\mathrm{H}}} \bar{h}_s h_t + e_{s\mathrm{H}_s} a\bar{r}_X x_\mathrm{H} h_s h_t +}{r_X x_\mathrm{H} h_s h_t + r_X x_{\bar{\mathrm{H}}} \bar{h}_s h_t + a\bar{r}_X x_\mathrm{H} h_s h_t +} \\
&\quad \frac{e_{s\bar{\mathrm{H}}_s} a\bar{r}_X x_{\bar{\mathrm{H}}} \bar{h}_s h_t + e_{s\mathrm{H}_s} a\bar{r}_X \bar{x}_\mathrm{H} h_s h_t + e_{s\bar{\mathrm{H}}_s} a\bar{r}_X \bar{x}_{\bar{\mathrm{H}}} \bar{h}_s h_t}{a\bar{r}_X x_{\bar{\mathrm{H}}} \bar{h}_s h_t + a\bar{r}_X \bar{x}_\mathrm{H} h_s h_t + a\bar{r}_X \bar{x}_{\bar{\mathrm{H}}} \bar{h}_s h_t} \\
&= \frac{\overbrace{r_X(e_{s\mathrm{H}_s} x_\mathrm{H} h_s + e_{s\bar{\mathrm{H}}_s} x_{\bar{\mathrm{H}}} \bar{h}_s) + \bar{r}_X a(e_{s\mathrm{H}_s} h_s + e_{s\bar{\mathrm{H}}_s} \bar{h}_s)}^{\gamma :=}}{\underbrace{r_X(x_\mathrm{H} h_s + x_{\bar{\mathrm{H}}} \bar{h}_s) + \bar{r}_X a}_{\delta :=}}
\end{aligned} \tag{17}
$$

With Equations 16 and 17, l^* transforms to

$$
l^* = \frac{\frac{\alpha}{\beta}}{\frac{\gamma}{\delta}}. \tag{18}
$$

From this we can see that $\alpha < \beta$ and $\gamma < \delta$. Now setting $Pr(\mathrm{R}_X) = 0$ results in

$$
l^* = \frac{\frac{\bar{r}_X a(e_{s\mathrm{H}_s} h_s + e_{s\bar{\mathrm{H}}_s} \bar{h}_s)}{\bar{r}_X a}}{\frac{\bar{r}_X a(e_{s\mathrm{H}_s} h_s + e_{s\bar{\mathrm{H}}_s} \bar{h}_s)}{\bar{r}_X a}} = 1.
$$

Setting $Pr(\mathrm{R}_X) > 0$, on the other hand, results in

$$
\frac{\alpha}{\beta} < \frac{\gamma}{\delta}.
$$

Thus, $l^* < 1$ if $Pr(\mathrm{R}_X) > 0$. □

Proof Theorem 3.2.

First, we show that our alternative model satisfies Equation 1. From the

Bayesian network's structure it follows that

$$\begin{aligned} Pr(\mathrm{X}) &= \sum_{H_s,H_t} Pr(\mathrm{X}|H_s,H_t)Pr(H_s)Pr(H_t) \\ &= x_{\mathrm{H}_s\mathrm{H}_t}h_sh_t + x_{\mathrm{H}_s\bar{\mathrm{H}}_t}h_s\bar{h}_t + x_{\bar{\mathrm{H}}_s\mathrm{H}_t}\bar{h}_sh_t + x_{\bar{\mathrm{H}}_s\bar{\mathrm{H}}_t}\bar{h}_s\bar{h}_t \\ &= h_s(x_{\mathrm{H}_s\mathrm{H}_t}h_t + x_{\mathrm{H}_s\bar{\mathrm{H}}_t}\bar{h}_t) + \bar{h}_s(x_{\bar{\mathrm{H}}_s\mathrm{H}_t}h_t + x_{\bar{\mathrm{H}}_s\bar{\mathrm{H}}_t}\bar{h}_t). \end{aligned} \tag{19}$$

Let α, β be defined as

$$\begin{aligned} \alpha &:= (x_{\mathrm{H}_s\mathrm{H}_t}h_t + x_{\mathrm{H}_s\bar{\mathrm{H}}_t}\bar{h}_t) \\ \beta &:= (x_{\bar{\mathrm{H}}_s\mathrm{H}_t}h_t + x_{\bar{\mathrm{H}}_s\bar{\mathrm{H}}_t}\bar{h}_t). \end{aligned}$$

Then Equation 19 transforms to

$$Pr(\mathrm{X}) = h_s\alpha + \bar{h}_s\beta. \tag{20}$$

Since $0 \leq Pr(\mathrm{X}|H_s,H_t) \leq 1$ and α, β are weighted averages, it follows from Equations 5 and 6 that $0 < \alpha, \beta < 1$. From this and the fact that the expression on the right hand side of the equality sign in Equation 20 is a weighted average, it follows with Equation 5 that $0 < Pr(\mathrm{X}) < 1$.

Next, we show that our alternative model satisfies Equation 4. To this end, we show that $Pr(\mathrm{X}|\mathrm{H}_i) - Pr(\mathrm{X}) \neq 0$ (where $i \in \{s,t\}$). We begin by showing that $Pr(\mathrm{X}|\mathrm{H}_s) - Pr(\mathrm{X}) \neq 0$. From the Bayesian network's structure it follows that

$$\begin{aligned} Pr(\mathrm{X}|\mathrm{H}_s) &= \sum_{H_t} Pr(\mathrm{X}|\mathrm{H}_s,H_t)Pr(H_t) \\ &= x_{\mathrm{H}_s\mathrm{H}_t}h_t + x_{\mathrm{H}_s\bar{\mathrm{H}}_t}\bar{h}_t. \end{aligned} \tag{21}$$

If we define

$$\Delta := Pr(\mathrm{X}|\mathrm{H}_s) - Pr(\mathrm{X}),$$

it follows with Equations 19 and 21 that

$$\Delta = \alpha - (h_s\alpha + \bar{h}_s\beta).$$

$h_s\alpha + \bar{h}_s\beta$ is a weighted average. We already know that $0 < \alpha, \beta < 1$. Thus, the only possibility for Δ to equal 0 consists in α's weight h_s being 1. This, however, is excluded by Equation 5. Therefore, Δ cannot be 0.

A proof for $Pr(\mathrm{X}|\mathrm{H}_t) - Pr(\mathrm{X}) \neq 0$ can be constructed similarly to the proof for $Pr(\mathrm{X}|\mathrm{H}_s) - Pr(\mathrm{X}) \neq 0$ above. We only need to replace H_s by H_t and vice versa in Equation 21. □

Proof Theorem 3.3.

(a) follows directly from applying the d-separation criterion (Definition 6.2) to the Bayesian network depicted in Figure 6, which tells us that H_t is probabilistically independent of E_s unconditionally.

(b) follows trivially from Theorem 3.3(c).

(c) Let Pr^* result from Pr by increasing $Pr(\mathrm{R}_X)$ while keeping all parameters of the original Bayesian network fixed. Let

$$l := \frac{Pr(\mathrm{E}_s|\bar{\mathrm{H}}_t, \mathrm{E}_X)}{Pr(\mathrm{E}_s|\mathrm{H}_t, \mathrm{E}_X)}$$
$$l^* := \frac{Pr^*(\mathrm{E}_s|\bar{\mathrm{H}}_t, \mathrm{E}_X)}{Pr^*(\mathrm{E}_s|\mathrm{H}_t, \mathrm{E}_X)}.$$

With Equation 18 we now get

$$l^* = \frac{\frac{\alpha^*}{\beta^*}}{\frac{\gamma^*}{\delta^*}} < l = \frac{\frac{\alpha}{\beta}}{\frac{\gamma}{\delta}},$$

where $\alpha, \beta, \gamma, \delta$ are defined as in Equations 16 and 17 and $\alpha^*, \beta^*, \gamma^*, \delta^*$ result from $\alpha, \beta, \gamma, \delta$ by replacing r_X by r_X^* in Equations 16 and 17. Thus, the higher $Pr^*(\mathrm{R}_X)$ is compared to $Pr(\mathrm{R}_X)$, the smaller l^* is compared to l and, because of that, the higher $c^*(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X)$ is compared to $c(\mathrm{H}_t; \mathrm{E}_s|\mathrm{E}_X)$. □

Acknowledgements: Funded by the European Union - Next Generation EU, Mission 4, Component 1, CUP I53D23006890001. We would like to thank two anonymous reviewers for helpful comments on an earlier version of this paper.

Compliance with Ethical Standards

The authors declare that there are no conflicts of interest.

References

Boge, F. J. (2020). How to infer explanations from computer simulations. *Studies in History and Philosophy of Science Part A*, *82*, 25–33.

Bolinska, A. (2013). Epistemic representation, informativeness and the aim of faithful representation. *Synthese*, *190*(2), 219–234.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Bueno, O. (2021). Structural representation and the ontology of models. In *Models and idealizations in science* (pp. 199–216). Springer.

Cartwright, N. (2012). Presidential address: Will this policy work for you? predicting effectiveness better: How philosophy helps. *Philosophy of Science*, *79*(5), 973–989.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Crowther, K., Linnemann, N. S., & Wüthrich, C. (2021). What we cannot learn from analogue experiments. *Synthese*, *198*, S3701–S3726.

Dardashti, R., Hartmann, S., Thébault, K., & Winsberg, E. (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*, 1–11.

Dardashti, R., Thébault, K., & Winsberg, E. (2015). Confirmation via analogue simulation: What dumb holes could tell us about gravity. *British Journal for the Philosophy of Science*, *68*(1), 55–89.

Dawid, R., Hartmann, S., & Sprenger, J. (2014). The no alternatives argument. *British Journal for the Philosophy of Science*, *66*(1), 213–234.

de Pretis Francesco, Landes, J., & Osimani, B. (2019). E-synthesis: A Bayesian framework for causal assessment in pharmacosurveillance. *Frontiers in Pharmacology*, *10*.

Feldbacher-Escamilla, C. J., & Gebharter, A. (2019). Modeling creative abduction Bayesian style. *European Journal for Philosophy of Science*, *9*(1), 9.

Feldbacher-Escamilla, C. J., & Gebharter, A. (2020). Confirmation based on analogical inference: Bayes meets Jeffrey. *Canadian Journal of Philosophy*, *50*(2), 174–194.

Field, G. E. (2021). Putting theory in its place: The relationship between universality arguments and empirical constraints. *British Journal for the Philosophy of Science*.

Glymour, C. (2019). Creative abduction, factor analysis, and the causes of liberal democracy. *Kriterion – Journal of Philosophy*, *33*(1), 1–22.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and philosophy*, *21*(5), 725–740.

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett publishing.

Henderson, L., & Gebharter, A. (2021). The role of source reliability in belief polarisation. *Synthese*.

Hesse, M. B. (1966). *Models and analogies in science*. University of Notre Dame Press.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*(1-2), 241–270.

Khosrowi, D. (2019). Extrapolation of causal effects–hopes, assumptions, and the extrapolator,Äôs circle. *Journal of economic methodology*, *26*(1), 45–58.

Landes, J., Osimani, B., & Poellinger, R. (2017). Epistemology of causal inference in pharmacology. *European Journal for Philosophy of Science*, *8*(1), 3–49.

Levy, A., & Currie, A. (2015). Model organisms are not (theoretical) models. *The British Journal for the Philosophy of Science*, *66*(2), 327–348.

Muller, S. M. (2015). Causal interaction and external validity: Obstacles to the policy relevance of randomized evaluations. *The World Bank Economic Review*, *29*(suppl. 1), S217–S225.

Osimani, B., & Landes, J. (2020). Varieties of error and varieties of evidence in scientific inference. *The British Journal for the Philosophy of Science*.

Osimani, B., & Poellinger, R. (2020). A protocol for model validation and causal inference from computer simulation. *A Critical Reflection on Automated Science: Will Science Remain Human?*, 173–215.

Pearl, J. (2000). *Causality* (1st ed.). Cambridge: Cambridge University Press.

Reiss, J. (2010). Across the boundaries: Extrapolation in biology and social science, daniel p. steel. oxford university press, 2007. xi+ 241 pages. *Economics & Philosophy*, *26*(3), 382–390.

Sprenger, J. (2015). The probabilistic no miracles argument. *European Journal for Philosophy of Science*, *6*(2), 173–189.

Sprenger, J., & Hartmann, S. (2016). *Bayesian philosophy of science*.

Steel, D. (2007). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

Suárez, M., & Bolinska, A. (2021). Informative models: idealization and abstraction. In *Models and idealizations in science* (pp. 71–85). Springer.

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Winsberg, E. (2009). A tale of two methods. *Synthese*, *169*, 575–592.