PREPRINT
Forthcoming in
The Routledge Handbook of Philosophy of Scientific Modeling
Eds: Tarja Knuuttila, Natalia Carrillo, and Rami Koskinen

The puzzle of model-based explanation

N. Emrah Aydinonat

TINT - Centre for Philosophy of Social Science
Faculty of Social Sciences
University of Helsinki
emrah.aydinonat@helsinki.fi
Orcid: 0000-0001-6633-4044

Among the many functions of models, explanation is central to the functioning and aims of science. However, the discussions surrounding modeling and explanation in philosophy have largely remained separate from each other. This chapter seeks to bridge the gap by focusing on the puzzle of model-based explanation, asking how different philosophical accounts answer the following question: if idealizations and fictions introduce falsehoods into models, how can idealized and fictional models provide true explanations? The chapter provides a selective and critical overview of the available strategies for solving this puzzle, mainly focusing on idealized models and how they explain.

Introduction

Among the many functions of models, explanation is central to the functioning and aims of science, and models explain in various ways. However, the discussions surrounding modeling and explanation in philosophy have largely remained separate from each other. Accounts of models has mainly focused on questions of representation, idealization, and fiction, mostly paying attention to the relation between models and their targets (e.g., Weisberg 2013; Frigg and Nguyen 2020). Accounts of explanation, on the other hand, predominantly concentrated on the nature and types of explanation, developing alternative accounts of explanation (e.g., Woodward 2003; Strevens 2008). Because philosophers generally agree that idealizations play indispensable roles both in modelling and explanation, one possible way to bring together these two lines of inquiry is focus on the role of idealized models in explanation. In both literatures, idealizations are commonly conceived of as distortions (however, see Carrillo and Knuuttila 2022): like fictions, they introduce falsehoods into models. There is also a common presumption that explanations must be true. The question is, if idealizations and fictions are "false", how can idealized models provide true explanations? This is the puzzle of model-based explanation (henceforth, the puzzle). To solve it, one would need to resolve many debates in the philosophy of science, and ideally provide compatible accounts of models, truth, fiction, idealization, representation, understanding, and explanation. This chapter has the more modest aim of giving a selective and critical overview of the available strategies to solve the puzzle mainly focusing on idealized models—although the discussion naturally extends to the case of fictional models. The chapter does not explicitly address applied models (i.e., models fine-tuned to a specific particular real-world target) or statistical models (including econometric models, machine learning models and the like), although some of the strategies to solve the puzzle may apply to them as well.

What is the puzzle?

The puzzle has been discussed in a variety of ways. Let us look at *some* examples—reformulated here as dilemmas or trilemmas.

Strevens (2008, 297) discusses the puzzle in terms of the difficulty of explaining the widespread use of idealizations for causal accounts of explanation.

(S_i) Nonveridical models cannot explain.

(S_{ii}) Idealized causal models misrepresent their targets.

(S_{iii}) Idealized causal models are commonly used to provide explanations.

Bokulich (2008, 140, fn. 9) focuses on the tension between the requirement of truth for explanation, and the practice of giving model-based explanation that are "not entirely true" (Bokulich 2009, 105).

(B_i) "Widely received philosophical accounts of scientific explanation" have a "strict requirement of truth." (2009, 104)

(B_{ii}) Scientists nevertheless explain with idealized or fictional models and provide explanations that are "not entirely true."

In philosophy of economics the puzzle is dubbed as an explanation paradox:

(R_i) "Economic models are false.

(R_{ii}) Economic models are nevertheless explanatory.

(R_{iii}) Only true accounts can explain." (Reiss 2012, 49)

Love and Nathan (2015, 768) underscore the conflict between the goal of accurate representation in explanation and the "deliberate misrepresentation" of mechanisms in models:

(LN_i) Accurate representation is necessary for mechanistic explanations.

(LN_{ii}) Idealized models of mechanisms that are cited in mechanistic explanations misrepresent those mechanisms.

Potochnick (2017) highlights the contradiction between the beliefs that explanations must be true and that idealizations are untrue:

(P_i) Explanations must be true.

(P_{ii}) Idealizations are patently untrue.

(P_{iii}) Idealized models explain.

Examples can be multiplied. Formulations of the puzzle assume that (i) a good explanation is a true explanation, (ii) idealized models explain, and (iii) idealizations are falsehoods or distortions. Proposed solutions to the puzzle often involve the rebuttal of one or more of these assumptions.

To solve the puzzle, philosophers of science have employed multiple strategies (cf. Reiss 2012): (A) abandoning the requirement of truth for explanation (*Explanations need not be true*), (B) arguing that models cannot explain (*Models cannot explain … but they might help)*, (C) arguing that models can contain truths, enable correct inferences, or provide true explanations despite (or thanks to) idealizations (*Models explain)*, and (D) arguing that *Models are not explanations, but tools*. Without trying to be exhaustive, let us look at examples from each strategy.

Explanations need not be true

Catherine Elgin (2004; 2017) famously argued that "laws, models, idealizations, and approximations which are acknowledged not to be true […] figure *ineliminably* in the success of science" (2004, 113–14, emphasis added). Thus, she said, if we were to stick to the requirement of truth strictly, we would have to conclude that "much of our best science" is "epistemologically unacceptable" (2004, 114). Thinking of the puzzle, one way to follow Elgin is to argue that explanations need not be true. This would be a straightforward solution since there is nothing puzzling about "false" models providing false explanations. Even so, philosophers rarely follow this strategy explicitly, most likely because they commonly subscribe to the factivity of explanation.[1] One notable exception is Potochnik (2017), who argues that "because idealizations are patently untrue," (2017, 93) model-based explanations cannot be true either (2017, 134). Because Potochnik accepts that models are "false" and that models can explain, she sacrifices the factivity of explanation. However, on closer inspection, she does not give up on truth completely. She argues that "idealized representations can *truly depict* causal patterns", and that "scientific representations generate understanding of phenomena *in virtue of being true of* causal patterns. (2017, 119, emphasis added). She also substitutes the truth requirement with the following: explanations must depict real causal patterns. That is, according

[1] It is possible for a pragmatist to argue that an explanation need not be true, but as Achinstein (1984, 290) notes, "a pragmatic theory of explanation does not commit one to anti-realism" (or realism). Even versions of pragmatic theory of explanation employ some conditions concerning the truth or correctness of the explanation.

to Potochnik, a good explanation "must capture what is responsible for the explanandum", and "depict dependence relations" (2017, 135). Therefore, Potochnik transforms the puzzle into a new one: how can patently "untrue" models depict what is truly responsible for the explanandum? Consequently, we are no closer to the solution of the original puzzle we started with. Before moving on, note that if we were to brush aside Potochnik's points about explanation, her account would find a better home under *Model explain … thanks to representational failure*.

Models cannot explain … but they might help

The second strategy is to reject the premise that 'models explain', saying that most idealized models cannot provide true explanations *by themselves,* but they are nevertheless explanatorily useful. There are variations to this theme.

Consider McMullin's (1978) *hypothetico-structural* (HS) account of explanation. McMullin conceives of structural explanations as causal explanations that explain the "properties or behavior of a complex entity […] by alluding to the structure of that entity" (1978, 139). He argues that HS explanations, where a structure is *postulated with a theoretical model* (HS model) to explain a phenomenon, are common in science. They are *hypothetical* because "a different structure might also account for the features to be explained" (1978, 139). They are provisional and tentative because they do not satisfy the truth requirement and cannot be considered as complete definitive explanations. In Hempel's terms, HS explanations are *potential explanations*, i.e., explanations where the truth or falsity of the propositions constituting the explanans are not known yet (Hempel 1965, 338). They can be turned into true explanations, if their explanans can be justified by de-idealization.

Craver's (2006) account of mechanistic models also acknowledges the usefulness of models for explanation, while introducing strong requirements for explanations. According to Craver models have many explanatory functions: they are tools for demonstration, sketching explanations, conjecturing how-possibly explanations, and so on (2006, 355). However, to be an explanation or to explain, a model needs to "characterize the phenomenon", "describe the behavior of an underlying mechanism," and the components it describes "should correspond to components in the mechanism" in its real-world target (2006, 361). Accordingly, Craver sees models on a continuum based on how well they satisfy these requirements: (i) 'phenomenal' models, which are mere descriptions that do not explain (2006, 358), (ii) how-possibly models, which are "loosely constrained conjectures" (2006, 361), (iii) how-plausibly models, which are how-possibly models that fit better into what we already know, and (iv) how-actually models, which give complete descriptions of the actual mechanism "that in fact produces the phenomenon" and "show how a mechanism works, not merely how it might work" (2006, 361).

Craver's account does not accept anything less than a complete description of a mechanism for a true explanation. Note, however, that this statement concerns the descriptions of explanatory mechanism in an explanation, not models. It does not assume that more detailed models are better (Craver and Kaplan 2020). In this account, most idealized models cannot be considered *as* explanations, but they can be helpful for explaining by providing explanatorily relevant information that can be used in explanations. On the other hand, if a model explains, it must be because it captures the truths about actual mechanisms and idealizations must have been harmless in this sense. Either way the puzzle is resolved.

A related account is Kaplan's (2011) 3M account. It introduces "a model–mechanism–mapping (3M) constraint on explanatory mechanistic models" (2011, 347): components of the model should map onto and match with the actual mechanisms producing the phenomenon. Models that do not satisfy this requirement cannot provide true explanations, but only how-possibly explanations. 3M account does not necessarily ask for de-idealization for explanatory usefulness. If there is some "model–mechanism correspondence [...] the model will be endowed with explanatory force", Kaplan argues (2011, 348). Nevertheless, according to Kaplan, anything short of a complete description of the actual mechanism(s) will be an incomplete explanation (2011, 348).

McMullin, Craver, and Kaplan agree that even though most idealized models cannot be considered as explanations, they are still explanatorily useful. Many philosophers agree, and some openly propose a weaker reading of models. For example, Alexandrova (2008) suggests that we should conceive of models as *open-formulae* that help in formulating explanatory hypotheses. In this account, models are not explanations in themselves, but just recipes, schemata or templates for explanatory causal claims (2008, 397). Using models in explanations requires further steps like identifying the relevant causal hypothesis and ensuring that it holds for the case at hand.

As it should be clear by now, the philosophers who argue that most models cannot explain, do not deny that models can be useful in the process of producing true explanations. Models have many functions, most of which can help in producing explanations: they can generate explanatory hypotheses, help in exploring possible explanations, provide conceptual frameworks, assist in sketching explanations, aid in devising potential explanations, etc. (e.g., see Pielou 1981; Wimsatt 1987; Odenbaugh 2005). There is a considerable literature on the exploratory role of models (Aydinonat 2007; 2008; Gelfert 2015; Shech and Gelfert 2019; Massimi 2019), their modal functions (e.g., Rappaport 1989; Massimi 2019; Sjölin Wirling and Grüne-Yanoff 2021) and the relation between idealized models and how-possibly explanations (e.g., Craver 2006; Ylikoski and Aydinonat 2014; Bokulich 2014; Verreault-Julien 2019; Nguyen 2022). Most of this literature agrees with Craver, Kaplan, Alexandrova and others that idealized models can help us discover true explanations. Interestingly, as we will see shortly, philosophers who argue that models can and do explain are also happy to accept this claim, arguing that some models are useful in developing how-possibly explanations, potential explanations, sketches, or comparison cases. All this suggests that perhaps the solution of the puzzle is to be searched by analyzing how models are used as *tools* for explanatory purposes rather than conceiving of models *as* explanations (more on this below).

Models explain

Another way to solve the puzzle is to argue that models can provide true explanations thanks to their (i) representational adequacy, (ii) capacity to be used to make correct inferences, and (iii) falsities.

… *thanks to representational adequacy*

Showing that idealized models can be true or contain truths would make their ability to explain less puzzling. Many philosophers take this route. Consider Mäki's *functional decomposition account*. Mäki argues that idealized models represent selective aspects of their targets, isolate explanatorily relevant factors, and with respect to these aspects and factors they can be true (e.g., Mäki 1992; 2010). Strevens (2008) thinks that the function of idealizations is to remove explanatorily irrelevant aspects of the explanandum phenomenon from the model. He argues, *if "done right"* (2008, 300), an idealized model contains two parts: idealizations and "difference-makers for the explanatory target" (2008, 318). In both accounts, idealizations do not distort or misrepresent explanatory factors; they help in isolating them. If this were true, the puzzle would be resolved.

Both accounts presume that models have modular components and can be decomposed into idealized and difference-making parts. But can we decompose models this way? Rice (2019) argues that most models do not decompose this way for two main reasons. First, idealizations are indispensable for many mathematical techniques that are employed in model building and without them explanation would not be possible (2019, 193). Second, the assumption that idealizations will not distort a model's representation of explanatorily relevant (e.g., difference-making) relations is often not true. Hence, it is often not possible to "map the accurate parts of the model onto what is relevant and its inaccurate parts onto what is irrelevant" (2019, 194). This would at least require further steps, such as some *interpretation* of and *commentary* on the model, by the *model user*.

If Rice is right and if some idealizations are ineliminable (Batterman 2009; see also C. Z. Elgin 2004) then it becomes difficult to solve the puzzle with a naïve decompositional strategy. However, a closer look reveals that Mäki and Strevens' strategies are not so naïve after all. For example, Strevens agrees that some interpretation might be required to determine explanatory (ir)relevance and even gives role to *explanatory framework*, which could include "nature and goals of a particular conversation" (2008, 151) hence the explanatory practices, conventions and norms within a field. Similarly, Mäki (2010, 180) emphasizes the importance of the intention and *purpose* of the model user, and *model commentary* that connects a model's elements with the real world. Both Mäki and Strevens are aware that determining whether a model explains requires some interpretation and information about the context, but they do not provide enough guidance about concepts such as explanatory framework and model commentary. Moreover, both accounts allow for incomplete model-based explanations, model-based explanations with varying degrees of explanatory power and how-possible explanations.

To overcome difficulties that these accounts face with regards to ineliminable idealizations, Pincock (2020; 2021) recommends abandoning the commitment to the truth of models' parts that perform the explanatory task and accepting that generalizations generated by models are often only *partially true*. But how can partially false generalizations provide wholly true explanations? According to Pincock, presence of falsehoods in models is consistent with true model explanations if "there is an appropriate truth underlying each falsehood" (2021, 18). The problem with this is that we do not know how to determine the truths underlying falsehoods any better than we know the answer to the original puzzle. While Pincock talks about underlying truths, Niiniluoto (2018, 57) argues that although each idealization might not be partially or approximately true, "together with other claims, an idealized theory or model as a whole may be truthlike or sufficiently similar to the real system". Either way, it remains unclear on what basis the model user infers the true claims that will constitute the explanans.

An alternative route is to argue that model-based explanations are *partial* in the Hempelian sense. In a partial explanation, "the explanans does not account for the explanandum-phenomenon in the specificity with which it is characterized by the explanandum-sentence" (Hempel 1965, 416). Elgin and Sober (2002) think that models can provide partial explanations without necessarily being false. They argue that idealized models can explain, if their idealizations are *harmless* in the sense that removing these idealizations would not "make much difference in the predicted value of the effect variable", that is, the explanandum (2002, 448). In this account, the explanandum, *E*, need not be entailed by the explanans or be derivable from it: it is enough if explanans implies *E′*, provided that it is *close enough* to *E* (2002, 448). The difficulty is that this approach presumes not only that successful idealizations ("done right") will be harmless in the sense that they will distort the model results only slightly, but also that the idealizations do not influence the truth of the explanans. However, if idealizations are ineliminable, how can we know that they are harmless in both senses? The similarity between *E* and *E′* will not do. Robustness analysis might help (e.g., Levins 1966), but it has limited use without empirical evidence (Orzack and Sober 1993). So, after all, it appears that idealized models can explain only if we can make sure that their idealizations play no role whatsoever in explanations, other than removing disturbing factors. Hence, given the ineliminability of idealizations, the puzzle remains (see also Bokulich 2011, 36).

… thanks to correct inferences

The preceding accounts in this section agree that explanatory inferences are made possible if a model (M) successfully represents a real-world target (T). An alternative approach is to reconsider what "M represents T" means and to reverse the relation between explanatory inferences and representation. The inferential conception of representation does just this: it says, if one can draw inferences about T by using M, then M represents T (e.g., Suárez 2004). Can this approach solve the puzzle?

Recall that the puzzle is a puzzle because it starts with the premise that idealized models are "false" and explanations are true. The inferentialist approach does not impose truth conditions for inferences, it only requires that the model user can make inferences about T using M. That M represents T does not imply that M provides a true explanation. Hence, conceived this way the inferentialist approach does not even address the puzzle, let alone solving it. However, there is a version of inferentialism that explicitly addresses the puzzle.

Kuorikoski and Ylikoski (2015) amend the inferentialist approach to argue that "model-based (explanatory) reasoning" is "a matter of drawing conclusions from given assumptions using external inferential aids" (i.e., models) and this basically explains the "epistemic role of models" (2015, 3827). In this account, models help answering what-if questions and making what-if inferences. It is argued that if M can be used to make *correct* inferences about T, then M represents T (2015, 3827).

The puzzle is then transformed into a new one: how can "false" models help in making correct inferences about their targets, and what ensures the reliability of these inferences and the truth of their conclusions? It is in answering these questions that Kuorikoski and Ylikoski drift apart from the basic inferentialist view and draw close to Mäki and Strevens. First, they argue that some assumptions of a model help *isolate* real-world dependency relations and as such they are not the source of falsities in a model (2015, 3829). It is these substantial assumptions that allow model users to use what they learn about models as guides to inferences about real-world phenomena: an explanatory model, despite the falsities introduced by idealizations, "get[s] the target explanatory dependence right" (2015, 3831) thanks to its substantial assumptions. Second, they argue that derivational robustness analysis (Woodward 2006; Kuorikoski, Lehtinen, and Marchionni 2010) increases the reliability of model inferences.

In brief, in this account, substantial assumptions and robustness analysis are making the heavy lifting with respect to the solution of the puzzle. There is a concern, however. Ineliminability of idealizations also undermines robustness analysis: altering ineliminable idealizations will change the nature of the model, and this would make model comparisons, which are required for robustness analysis, problematic (Lisciandra 2017). Thus, the advertised epistemic benefits of robustness analysis might not realize, and the puzzle would remain (see also Verreault-Julien 2021).

On the positive side, Kuorikoski and Ylikoski avoid overemphasizing representation and settle for a modest claim concerning model explanation: models "capture a small set of explanatory dependencies that are assumed to be central" (2015, 3830) and when they are used to explain particular empirical phenomena, they do not necessarily provide complete or actual explanations: a model can sometimes be merely "a part of a how-possibly explanation" (2015, 3831). Both by emphasizing the role of robustness in enabling model-based inferences and acknowledging the selectiveness and partiality of representation, Kuorikoski and Ylikoski establish that model-based explanations cannot be fully understood by examining an isolated model: often a family-of-models perspective is needed (Ylikoski and Aydinonat 2014; see also Love and Nathan 2015).

… thanks to representational failure

We have seen that accounts that focus on representational adequacy encounter difficulties with the ineliminability of idealizations. Batterman (2009, 45) argues that some idealizations are necessary for explanation, and de-idealization might even reduce the explanatory power of some models. Batterman and Rice (2014) take this argument one step further, arguing that "highly idealized models can play explanatory roles despite *near complete representational failure*" (2014, 355, emphasis added). They argue that accounts that focus merely on representational adequacy *fail to explain* why idealizations are explanatory (2014, 365). To make their point, Batterman and Rice focus on a class of explanations of macrolevel patterns across systems using highly idealized models. They show that as a representation of any *particular* system these models are inadequate because they leave out the important particular details of individual systems. Nevertheless, they argue, these models are explanatory exactly because they leave out these details. If one asks why a set of different systems are strikingly similar in a certain aspect (e.g., a macrolevel pattern or feature), this might make the details of individual systems unnecessary from an explanatory point of view: the reason why these systems are similar might have nothing to do with their particular details but with some general features that are shared by all of them. If this is the case, adding detail – to increase the representational adequacy of the model from the perspective of one given individual system – would hinder the explanatory focus and power of the model. Thus, in such a case, idealization would be in fact necessary for explanation.

This point is well taken, but does it really go against the representational adequacy point of view? Representational adequacy depends on the explanatory task at hand. If the task is to explain common macro features of heterogenous systems, a model that focuses only on a small number of common features among these systems would be representationally adequate, even according to a hardheaded representationalist. When Batterman and Rice talk about "complete representational failure", they are obviously talking about the representational adequacy of the model with respect to a particular system, which is not relevant given the explanatory task. Thus, contrary to the appearances, the disagreement is not that severe (see also Lange 2015; Reutlinger 2017). Whereas representationalists argue that falsities introduced by idealizations are irrelevant, Batterman and Rice ask for *an explanation* of why the left-out details are irrelevant. They argue that, at least for the class of models they discuss, "the real explanatory work is done by showing why the various heterogeneous details of these systems are irrelevant and, along the way, by demonstrating the relevance of the common features" (2014, 365). Using examples from fluid dynamics and biology, they argue that these models are explanatory because they have a *backstory* showing that the model and the heterogenous systems it is supposed to explain belong to the same universality class. Note that merely providing a model that is in the same universality class with the phenomena it is supposed to explain does not provide much information. Batterman and Rice are asking for more: for a demonstration, for a story that explains the explanatoriness of the model. "The models are explanatory in virtue of a *there being a story* about why large classes of features are irrelevant," they say (2014, 356, emphasis added).[2] For the class of models that Batterman and Rice analyzing, this appears to solve the

---

[2] In later work Rice (2019, 201) loosens this requirement: "scientists can justifiably use idealized models within a universality class to explain the behaviours of real-world systems in that class

puzzle, in principle. In practice, however, explaining explanatory irrelevance involves considering the context of modeling and explanation. This is perhaps the larger lesson to extract from Batterman and Rice: answering why the relevant isolations are in place, why they were introduced, what modelers discovered by employing certain idealizations, etc. is crucial to an understating of explanatory value. In this regard, studying the broader the context of modelling is often superior to just studying an isolated model-target pair (Aydinonat and Köksal 2019). As we will see, philosophers who see models as tools take this suggestion one step further.

Although many philosophers offer potential solutions to the puzzle, only a very few addresses it directly. Bokulich is one of these exceptions and sets her task to show that "idealizations themselves are capable of doing some real explanatory work" (2011, 36). She first defines model-based explanation or *model explanation* as an explanation whose explanans "makes essential reference to" (2011, 38) an idealized or fictional model. Next she defines what it means for a model to explain: a *model explains* when it shows how its elements "correctly capture the patterns of counterfactual dependence in the target system" (2017, 106) or can "'reproduce' the relevant features of the explanandum phenomenon" (2011, 39), enabling model users to answer a wide range of what-if questions. How does this solve the puzzle? How can a "false" model get the counterfactual structure right (i.e., provide a true explanation)? To answer this, Bokulich introduces another step, a *justificatory step* that specifies the model's domain of applicability, shows that the explanandum "falls within that domain", and ensures that it "adequately capture[s] the relevant features of the world" (2011, 39).

According to Bokulich, justification might come from theory, showing that "model can be trusted as *an adequate representation* of the world" or "through various empirical investigations" (2011, 39, emphasis added). Moreover, justificatory step is "to be understood as playing a role analogous to Hempel's condition of truth […]." It is "intended to rule out as explanatory those models that we know to be merely phenomenological." (2011, 39, fn. 11). So, in this account, the justificatory step does "the heavy lifting" (2012, 736).

Where are we at concerning the puzzle? Bokulich's account is not too different from representationalist accounts insofar as the justificatory step is intended to ensure that falsities or fictionalizations in the model are *harmless* with respect to the model's ability to capture the truths about the counterfactual structure of the explanandum phenomenon given the explanatory task. A model might be idealized or refer to fictional entities, but what matters for explanation is whether it gets the explanatory relations, connections, structures, etc. right. The important point is, without the justificatory step, which is often contextual and dependent on the current state of knowledge (Bokulich 2012), we cannot know whether the explanatory hypotheses

even when they fail to have a complete explanation of why that universality class occurs." Also see Woodward (2018) on the sufficiency of information about irrelevance for explanation.

generated using the model are true or not. Without it, we only have sketches, templates, potential explanations.

Nguyen (2021) argues that to get the counterfactual dependence right, a model must represent the dependence relation in its target, say, between A and B, correctly. However, in contrast to Bokulich, he contends that since the explanation concerns the relation between A and B, it cannot be said that the falsities in the model plays any role in the explanation even though they "play an essential role in *generating* the explanation" (2021, 3232, emphasis added). More generally, according to Frigg and Nguyen's (2020) DEKI (Denotation, Exemplification, Keying-up, and Imputation) account of representation, idealized and fictional models *can* explain provided that they appropriately represent the target. This, however, requires (i) an appropriate *interpretation* of the model given the goals of modeling and explanation, and (ii) a *key*, that translates model's properties to the properties that will be imputed onto the target. Although, Frigg and Nguyen's solution to the puzzle is like Bokulich's solution in that it argues that models can explain thanks to representational failure, it does not assume that models explain by themselves: without interpretation and keying-up there would be no model explanation according to the DEKI account. Frigg and Nguyen argue that idealizations and fictions could play an essential role in *producing* the explanation; they do not argue that they are necessarily a part of the explanation. In this sense, their account would perhaps be more at home next to those who argue that models explain thanks to their representational adequacy.

The importance of context and goals of modeling and explanation appears to be a point agreed by most philosophers, despite their differences. Another point of agreement, without explicit acknowledgement, seems to be that merely focusing on the model-target relation is not entirely helpful in understanding or solving the puzzle: such things as interpretation, model commentary, model use, explanatory goals, model justification, and exploration have been repeatedly invoked in dealing with the puzzle.

Models are not explanations, but tools

Models are not explanations

If one assumes that *explanations must be true* and *idealized models are false*, then considering false models as explanatory seems paradoxical. However, the paradox arises if we also assume either that (i) models are explanations or (ii) that models are featured in the set of explanans directly, without any interpretation. If models are not explanations and are not commonly used in the explanans without modification, the puzzle would dissolve because that models contain idealizations would not necessarily mean that the explanantia of model-based explanations are false.

Consider the first assumption. Can an idealized or fictional model be an explanation? One difficulty with equating a model to an explanation is that models and explanations might be different sort of things. If this is true, conceiving of models as explanations would be misguided. However, even if we assume that models and explanations are the same sort of things, it is hard to conceive of idealized or fictional models as explanations. For the sake of the argument, Rohwer and Rice (2016) assume that both models and explanations can be "characterized or reinterpreted as sets of propositions" (2016, 1130) and explore where this assumption leads us to. They show that if this assumption were true, a model and an explanation would be identical only for some simple cases that do not involve idealizations or fictions. For a model to be identical to an explanation, the assumptions of the model (or a subset of these assumptions) must constitute the explanans, and the model result they imply must be identical to the explanandum. If a model were to employ idealizing assumptions, this would mean that the explanans of the model explanation cannot be true—unless the model's idealizing assumptions are reinterpreted in some way. In short, for the case of idealized and fictional models it is hard to say that there would be an identity preserving matching between the elements of a model and an explanation if we hold on to the truth requirement for explanation. In fact, Rohwer and Rice (2016) show that in most cases some interpretation of a model is required for explanation. Relatedly, Marchionni (2017) argues that seeing models as explanations is too limiting and leaves out many explanatory models, particularly explanatory idealized models. In most cases, models *help* explain rather than being explanations in themselves.

If most idealized models are not explanations, perhaps the second assumption is true, and models are featured in the set of explanans directly, without any interpretation. Recall that Bokulich argues that the explanans of a model explanation "makes an essential reference to" (2011, 38) a model. Thus, Bokulich does not equate model with explanations but argues that models are featured in explanations. In her other work, she uses alternative formulations: "makes central use of" (2018, 144) and "appeal[s] to certain properties or behaviors observed in" (2017, 104) a model. But what do these mean? Essential in what sense? What kind of reference, use or appeal? Bokulich does not answer these questions. Moreover, her justificatory step requirement, which is external to the model, implies that there must be some interpretation of the model involved in a model explanation. In conclusion, there does not appear to be good reasons believe in either of the two presumptions of the puzzle. This constitutes yet another solution: it is perhaps a pseudo puzzle after all.

Even though clarifying the relation between a model and an explanation is a promising strategy to resolve the puzzle, there are only a few explicit attempts to do this. We have seen that Bokulich tells us that model explanations makes an essential reference to models. In contrast, Marchionni (2017) argues that we should not consider any explanation that cites a model as explanatory. She recommends asking whether the model provides *explanatorily relevant information* independently of whether the model or some of its parts are cited in the explanans. Lawler and Sullivan (2020), on the other hand, advise us against seeing model-based explanations as a special kind of explanation. The sheer diversity of models and their explanatory uses suggest that they might have a point. They argue that in most cases 'model explanations' are just *model-induced explanations*, rather than models being explanations.

The statements of the puzzle appear to make the implicit assumption that idealized models, their premises, or results are or could be somehow added to the set of explanantia without modification and that the falsity of idealizations are preserved in the explanatory context. However, throughout the chapter we have seen that when challenged, philosophers repeatedly invoked concepts such as justification, interpretation, commentary, and context to defend their versions of how models explain. In most cases, they argued that models contribute to explanations in several ways.

Models are tools

Taking the arguments concerning various explanatory functions of models, importance of context, exploration and justification seriously suggests that we should not ignore what scientists do with their models and how they use them to explain. Looking at how models are used and manipulated for explanatory purposes can provide a key to the puzzle. There are several arguments to this effect. For example, Kennedy (2012), and Jebeile and Kennedy (2015) argue that false idealizations enable model-based explanation by allowing scientists to produce *comparison cases*. Idealizations then allow "scientists to determine what *is* causally relevant" (Kennedy 2012, 327) by comparing the model to the real-world case at hand. Jebeile and Kennedy suggest that merely focusing on representational adequacy is a mistake: explanatory functions of models can be better understood if we consider *models as* "*epistemic tools* that are designed by and for scientists to make inferences, and explanations (2015, 384, emphasis added), and *explanation as* "a process or *an activity*, rather than simply a product" (2015, 384, emphases added). In other words, model-based explanation cannot be fully understood without studying how model users use models to explain.

Another example is an argument by Boesch (2021) who says that dissimilarities found in models enable "novel forms of manipulation" (2021, 504) and thereby facilitate the attainment of epistemic aims, such as explanation. Many representationalists would agree on the point about dissimilarly or function of false idealizations: 'it is thanks to the dissimilarities we are able to focus on what matters', they would say (e.g., see Mäki 2011). However, Boesch, Kennedy and Jebeile are right in arguing that representationalists put too little emphasis on how model use and manipulation make explanatory inferences possible, crippling their ability to solve the puzzle.

This point is closely related to and follows from the view that sees models as tools that scientists built and manipulate to learn about the world (Morgan and Morrison 1999; Morgan 2012). In this view, models have been characterized in a variety of related ways: as mediators (Morgan and Morrison 1999), epistemic artefacts (Knuuttila 2005), and erotetic devices (Carrillo and Knuuttila 2022; Knuuttila 2021). In contrast to the representationalist accounts of models, which start from questions concerning representation and model-target relations, this view focuses on how models are built, used, and manipulated to allow epistemic access to the world. It is argued that the widely held view that idealizations are distortions is misleading since it moves the focus away from the process and context of modelling to mere comparisons between models and their targets (e.g., Carrillo and Knuuttila 2022). This approach emphasizes that understanding models as tools that can perform useful epistemic functions such as explanation requires moving beyond the model-target dyad and taking into account the purposes of model building and manipulations, as well as the context of modeling and its place in scientific practice (Knuuttila 2010; 2011; see also Morgan 2012).

How does this so-called artefactual approach view the puzzle? First, it sees the puzzle as pointless, since its proponents assume that there is no independent way of accessing the world without representation. Nevertheless, one lesson we can extract is the following: faced with the ineliminability of idealizations, solving the puzzle appears to require more than a focus on the model-target dyad (Knuuttila 2010; Carrillo and Knuuttila 2022). Following up on this point requires getting rid of the straightjackets of representationalist and inferentialist accounts, and more detailed case studies on actual model-based explanations. Second, more recent work that characterizes models as erotetic devices provide a more explicit link between models and explanations. Recall that several philosophers argued that models provide how-possibly explanations. Knuuttila (2021) argues that by seeing models as erotetic devices that are constructed to answer theoretical and explanatory questions, we can understand the modal functions of models and hence how they can provide how-possibly explanations better. This appears to be a fruitful line of research that could help in resolving the puzzle conceived as an *inference gap*; i.e., gap between what we know about the model and our model-based inferences concerning the real world.

Concluding remarks

The chapter started by saying that to solve the puzzle, one needs resolve many debates in the philosophy of science, and ideally provide compatible accounts of models, truth, fiction, idealization, representation, understanding, and explanation. This is because the puzzle is about all of these things. Philosophical accounts of models and explanation, on the other hand, are like scientific models in that they employ many abstractions and idealizations. They were set out to answer very specific questions concerning a limited set of philosophical problems, but not about the full set of questions relating to how models help us explain. For this reason, although each account provided insights into how model-based explanations work and what they might be, they were also vulnerable to criticism, being limited by their assumptions. This short discussion suggests that we still have a long way to go in explicating how model-based explanations explain.

What should be the next steps?

Firstly, it should be obvious that preconceptions concerning what model explanations are can only take us so far. Given that there are several ways in which models can contribute to explanations, more detailed studies of how explanations are produced using models are needed (Rice, Rohwer, and Ariew 2019). Moreover, the roles of interpretation, model commentary and explanatory context (and all other escape routes we encountered) in model-based explanation need to be studied further, and with more case studies. Doing this might require a more historical approach (Aydinonat and Köksal 2019). It will also be useful if such case studies explicitly and clearly state the the explanandum and explanans of the model-based explanations that they discuss.

Secondly, and relatedly, we should pay more attention to the diversity of types of models and model-based explanation. Both Aydinonat (2008) and Marchionni (2017) suggest that when discussing model-based explanations one needs to make further elementary distinctions. Model-based explanations have different types of explananda. Some explain singular events, some explain generic events, and some explain laws and law-like generalizations. Accordingly, we have singular and generic model-based explanations, as well as model-based explanations of laws. Some model-based explanations are complete, others are incomplete, and incomplete ones are incomplete in different ways. Then we have potential explanations, possible explanations, actual explanations, causal explanations, structural explanations, non-causal explanations, equilibrium explanations, etc. Moreover, in practice, explanations are never perfect: they are far from the ideals set by philosophers. Consequently, as Marchionni (2017) suggests, if we would like to study model-based explanations, we should be also willing to incorporate varying degrees of explanatory power to our frameworks.

Thirdly, it appears that seeing models as tools or epistemic artifacts will serve the useful purpose of settling many debates, if proponents of this view can show how model use and manipulation contribute to explanation, understanding or learning—i.e., providing an account of how the inference gap is closed.

Fourthly, recognizing that in practice many explanations make use of multiple models will help seeing the actual explanatory contribution of individual models.

And finally, more attention needs to be paid to models that fail to explain—to avoid the positive results bias in the philosophy of science.

References

Achinstein, Pete. 1984. "The Pragmatic Character of Explanation." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1984 (2): 274–92. https://doi.org/10.1086/psaprocbienmeetp.1984.2.192509.

Alexandrova, Anna. 2008. "Making Models Count." *Philosophy of Science* 75 (3): 383–404. https://doi.org/10.1086/592952.

Aydinonat, N Emrah. 2007. "Models, Conjectures and Exploration: An Analysis of Schelling's Checkerboard Model of Residential Segregation." *Journal of Economic Methodology*. Routledge. https://doi.org/10.1080/13501780701718680.

Aydinonat, N. Emrah. 2008. *The Invisible Hand in Economics: How Economists Explain Unintended Social Consequences*. London: Routledge.

Aydinonat, N. Emrah, and Emin Köksal. 2019. "Explanatory Value in Context: The Curious Case of Hotelling's Location Model." *The European Journal of the History of Economic Thought* 26 (5): 879–910. https://doi.org/10.1080/09672567.2019.1626460.

Batterman, Robert W. 2009. "Idealization and Modeling." *Synthese* 169 (3): 427–46. https://doi.org/10.1007/s11229-008-9436-1.

Batterman, Robert W., and Collin C. Rice. 2014. "Minimal Model Explanations." *Philosophy of Science* 81 (3): 349–76. https://doi.org/10.1086/676677.

Boesch, Brandon. 2021. "Scientific Representation and Dissimilarity." *Synthese* 198 (6): 5495–5513. https://doi.org/10.1007/s11229-019-02417-0.

Bokulich, Alisa. 2008. *Reexamining the Quantum-Classical Relation: Beyond Reductionism and Pluralism*. Cambridge, UK ; New York: Cambridge University Press.

———. 2009. "Explanatory Fictions." In *Fictions in Science: Philosophical Essays on Modeling and Idealization*, edited by Mauricio Suárez, 91–109. London: Routledge.

———. 2011. "How Scientific Models Can Explain." *Synthese* 180 (1): 33–45. https://doi.org/10.1007/s11229-009-9565-1.

———. 2012. "Distinguishing Explanatory from Nonexplanatory Fictions." *Philosophy of Science* 79 (5): 725–37. https://doi.org/10.1086/667991.

———. 2014. "How the Tiger Bush Got Its Stripes: 'how Possibly' vs. 'How Actualy' Model Explanations." *The Monist* 97 (3): 321–38.

———. 2017. “Models and Explanation.” In *Springer Handbook of Model-Based Science*, edited by Lorenzo Magnani and Tommaso Bertolotti, 103–18. Dordrecht: Springer.

———. 2018. “Searching for Non-Causal Explanations in a Sea of Causes.” In *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, edited by Alexander Reutlinger and Juha Saatsi, 141–63. Oxford University Press. https://doi.org/10.1093/oso/9780198777946.003.0008.

Carrillo, Natalia, and Tarja Knuuttila. 2022. “Holistic Idealization: An Artifactual Standpoint.” *Studies in History and Philosophy of Science* 91 (February): 49–59. https://doi.org/10.1016/j.shpsa.2021.10.009.

Craver, Carl F. 2006. “When Mechanistic Models Explain.” *Synthese* 153 (3): 355–76. https://doi.org/10.1007/s11229-006-9097-x.

Craver, Carl F., and David M. Kaplan. 2020. “Are More Details Better? On the Norms of Completeness for Mechanistic Explanations.” *The British Journal for the Philosophy of Science* 71 (1): 287–319. https://doi.org/10.1093/bjps/axy015.

Elgin, Catherine Z. 2004. “True Enough.” *Philosophical Issues* 14: 113–31. https://doi.org/10.1111/j.1533-6077.2004.00023.x.

———. 2017. *True Enough*. Cambridge, MA: MIT Press.

Elgin, Mehmet, and Elliott Sober. 2002. “Cartwright on Explanation and Idealization.” *Erkenntnis* 57 (3): 441–50. https://doi.org/10.1023/A:1021502932490.

Frigg, Roman, and James Nguyen. 2020. *Modelling Nature: An Opinionated Introduction to Scientific Representation*. Cham, Switzerland: Springer.

Gelfert, Axel. 2015. *How to Do Science with Models*. New York, NY: Springer Berlin Heidelberg.

Hempel, Carl G. 1965. “Aspects of Scientific Explanation.” In *Scientific Explanation: Essays in Philosophy of Science*, 331–496. New York: Free Press.

Jebeile, Julie, and Ashley Graham Kennedy. 2015. “Explaining with Models: The Role of Idealizations.” *International Studies in the Philosophy of Science* 29 (4): 383–92. https://doi.org/10.1080/02698595.2015.1195143.

Kaplan, David Michael. 2011. “Explanation and Description in Computational Neuroscience.” *Synthese* 183 (3): 339. https://doi.org/10.1007/s11229-011-9970-0.

Kennedy, Ashley Graham. 2012. “A Non Representationalist View of Model Explanation.” *Studies in History and Philosophy of Science Part A*, Structures and Strategies in Ancient Greek and Roman Technical Writing, 43 (2): 326–32. https://doi.org/10.1016/j.shpsa.2011.12.029.

Knuuttila, Tarja. 2005. *Models as Epistemic Artefacts: Toward a Non-Representationalist Account of Scientific Representation*. Philosophical Studies from the University of Helsinki 8. Vantaa: Department of Philosophy, University of Helsinki.

———. 2010. “Some Consequences of the Pragmatist Approach to Representation: Decoupling the Model-Target Dyad and Indirect Reasoning.” In *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association*, edited by Mauricio Suárez, Mauro Dorato, and Miklós Rédei, 139–48. Dordrecht: Springer. https://doi.org/10.1007/978-90-481-3263-8.

———. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42 (2): 262–71. https://doi.org/10.1016/j.shpsa.2010.11.034.

———. 2021. “Epistemic Artifacts and the Modal Dimension of Modeling.” *European Journal for Philosophy of Science* 11 (3): 65. https://doi.org/10.1007/s13194-021-00374-5.

Kuorikoski, Jaakko, Aki Lehtinen, and Caterina Marchionni. 2010. “Economic Modelling as Robustness Analysis.” *The British Journal for the Philosophy of Science* 61 (3): 541–67. https://doi.org/10.1093/bjps/axp049.

Kuorikoski, Jaakko, and Petri Ylikoski. 2015. “External Representations and Scientific Understanding.” *Synthese* 192 (12): 3817–37. https://doi.org/10.1007/s11229-014-0591-2.

Lange, Marc. 2015. “On ‘ Minimal Model Explanations ’ : A Reply to Batterman and Rice.” *Philosophy of Science* 82 (2): 292–305.

Lawler, Insa, and Emily Sullivan. 2020. “Model Explanation Versus Model-Induced Explanation.” *Foundations of Science*, February. https://doi.org/10.1007/s10699-020-09649-1.

Levins, Richard. 1966. “The Strategy of Model Building in Population Biology.” In *Conceptual Issues in Evolutionary Biology*, edited by Eliot Sober, 18–27. Cambridge, MA: MIT Press.

Lisciandra, Chiara. 2017. “Robustness Analysis and Tractability in Modeling.” *European Journal for Philosophy of Science* 7 (1): 79–95. https://doi.org/10.1007/s13194-016-0146-0.

Love, Alan C., and Marco J. Nathan. 2015. “The Idealization of Causation in Mechanistic Explanation.” *Philosophy of Science* 82 (5): 761–74. https://doi.org/10.1086/683263.

Mäki, Uskali. 1992. “On the Method of Isolation in Economics.” In *Idealization IV: Intelligibility in Science (Poznan Studies in the Philosophy of the Sciences and the Humanities 26)*, edited by C. Dilworth, 317–51. Amsterdam: Rodopi.

———. 2010. “Models and Truth: The Functional Decomposition Approach.” In *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association*, edited by Mauricio Suárez, Mauro Dorato, and Miklós Rédei, 177–87. Springer.

———. 2011. “Models and the Locus of Their Truth.” *Synthese* 180 (June 2009): 47–63. https://doi.org/10.1007/s11229-009-9566-0.

Marchionni, Caterina. 2017. “What Is the Problem with Model-Based Explanation in Economics?” *Disputatio* 9 (47): 603–30. https://doi.org/10.1515/disp-2017-0020.

Massimi, Michela. 2019. “Two Kinds of Exploratory Models.” *Philosophy of Science* 86 (December): 869–81. https://doi.org/10.1086/705494.

McMullin, Ernan. 1978. “Structural Explanation.” *American Philosophical Quarterly* 15 (2): 139–47.

Morgan, Mary S. 2012. *The World in the Model: How Economists Work and Think*. Cambridge ; New York: Cambridge University Press.

Morgan, Mary S., and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science (Ideas in Context)*. Vol. 1. Cambridge University Press.

Nguyen, James. 2021. “Do Fictions Explain?” *Synthese* 199 (1–2): 3219–44. https://doi.org/10.1007/s11229-020-02931-6.

———. 2022. “Scientific Modeling.” In *The Palgrave Encyclopedia of the Possible*, 1–10. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-98390-5_183-1.

Niiniluoto, Ilkka. 2018. “Explanation by Idealized Theories.” *Kairos. Journal of Philosophy & Science* 20 (1): 43–63. https://doi.org/10.2478/kjps-2018-0003.

Odenbaugh, Jay. 2005. “Idealized, Inaccurate but Successful: A Pragmatic Approach to Evaluating Models in Theoretical Ecology.” *Biology and Philosophy* 20 (2–3): 231–55. https://doi.org/10.1007/s10539-004-0478-6.

Orzack, Steven Hecht, and Elliott Sober. 1993. “A Critical Assessment of Levins’s The Strategy of Model Building in Population Biology (1966).” *The Quarterly Review of Biology* 68 (4): 533–46.

Pielou, E. C. 1981. “The Usefulness of Ecological Models: A Stock-Taking.” *The Quarterly Review of Biology* 56 (1): 17–31. https://doi.org/10.1086/412081.

Pincock, Christopher. 2020. "Concrete Scale Models, Essential Idealization, and Causal Explanation." *The British Journal for the Philosophy of Science*, December, 000–000. https://doi.org/10.1093/bjps/axz019.

———. 2021. "A Defense of Truth as a Necessary Condition on Scientific Explanation." *Erkenntnis*, January. https://doi.org/10.1007/s10670-020-00371-9.

Potochnik, Angela. 2017. *Idealization and the Aims of Science*. Chicago: The University of Chicago Press.

Rappaport, Steven. 1989. "The Modal View of Economic Models." *Philosophica* 44: 61–80.

Reiss, Julian. 2012. "The Explanation Paradox." *Journal of Economic Methodology* 19 (1): 43–62. https://doi.org/10.1080/1350178X.2012.661069.

Reutlinger, Alexander. 2017. "Do Renormalization Group Explanations Conform to the Commonality Strategy?" *Journal for General Philosophy of Science* 48 (1): 143–50. https://doi.org/10.1007/s10838-016-9339-7.

Rice, Collin. 2019. "Models Don't Decompose That Way: A Holistic View of Idealized Models." *The British Journal for the Philosophy of Science* 70 (1): 179–208. https://doi.org/10.1093/bjps/axx045.

Rice, Collin, Yasha Rohwer, and André Ariew. 2019. "Explanatory Schema and the Process of Model Building." *Synthese* 196 (11): 4735–57. https://doi.org/10.1007/s11229-018-1686-y.

Rohwer, Yasha, and Collin C. Rice. 2016. "How Are Models and Explanations Related?" *Erkenntnis* 81: 1127–48. https://doi.org/10.1007/s10670-015-9788-0.

Shech, Elay, and Axel Gelfert. 2019. "The Exploratory Role of Idealizations and Limiting Cases in Models." *Studia Metodologiczne* 39 (195–232): 38. https://doi.org/10.14746/sm.2019.39.8.

Sjölin Wirling, Ylwa, and Till Grüne-Yanoff. 2021. "The Epistemology of Modal Modeling." *Philosophy Compass* n/a (n/a): e12775. https://doi.org/10.1111/phc3.12775.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, Mass: Harvard University Press.

Suárez, Mauricio. 2004. "An Inferential Conception of Scientific Representation." *Philosophy of Science* 71 (5): 767–79. https://doi.org/10.1086/421415.

Verreault-Julien, Philippe. 2019. "How Could Models Possibly Provide How-Possibly Explanations?" *Studies in History and Philosophy of Science Part A* 73 (February): 22–33. https://doi.org/10.1016/j.shpsa.2018.06.008.

———. 2021. "Factive Inferentialism and the Puzzle of Model-Based Explanation." *Synthese* 199 (3–4): 10039–57. https://doi.org/10.1007/s11229-021-03235-z.

Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Wimsatt, William C. 1987. "False Models as Means to Truer Theories." In *Neutral Models in Biology*, edited by Matthew H. Nitecki and Antoni Hoffman, 23–55. Oxford: Oxford University Press.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

———. 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13 (2): 219–40. https://doi.org/10.1080/13501780600733376.

———. 2018. *Some Varieties of Non-Causal Explanation*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780198777946.003.0007.

Ylikoski, Petri, and N. Emrah Aydinonat. 2014. "Understanding with Theoretical Models." *Journal of Economic Methodology* 21 (1): 19–36. https://doi.org/10.1080/1350178X.2014.886470.