Model explanation versus model-induced explanation

By: Insa Lawler and Emily Sullivan[*]

Lawler, I. & Sullivan, E. (2020). Model explanation versus model-induced explanation. *Foundations of Science*. https://doi.org/10.1007/s10699-020-09649-1

Abstract:

Scientists appeal to models when explaining phenomena. Such explanations are often dubbed model explanations or model-based explanations (short: ME). But what are the precise conditions for ME? Are ME special explanations? In our paper, we first rebut two definitions of ME and specify a more promising one. Based on this analysis, we single out a related conception that is concerned with explanations that are *induced* from working with a model. We call them 'model-induced explanations' (MIE). Second, we study three paradigmatic cases of alleged ME. We argue that all of them are MIE, upon closer examination. Third, we argue that this undermines the building consensus that model explanations are special explanations that, e.g., challenge the factivity of explanation. Instead, it suggests that what is special about models in science is the epistemology behind how models induce explanations.

Keywords: Model | Explanation | Model-based Explanation | Idealization

Article:

1 Introduction

Models are frequently used in science. Some of them are used for merely *exploratory* purposes (cf., e.g., Kennedy 2012; Rohwer and Rice 2013; Gelfert (2016), ch. 4.). For instance, scientists construct models to calculate possible climate scenarios (see, e.g., Parker 2006; Werndl and Steele 2016), and quite a few models in economics are used to explore the behavior of ideal rational agents (see, e.g., Mäki 2005; Alexandrova 2008; Alexandrova and Northcott 2013; Grüne-Yanoff 2009; Marchionni 2017).[1] But scientists also appeal to models

[*] The order of authors is in alphabetical order as both authors have contributed equally to this work.
[1] Examples for explorative functions of models are the following (cf. Rohwer and Rice 2016, pp. 1141–1144): (i) Some models enable the modeler to view the phenomenon of interest from a novel perspective. (ii) Some models function as aids to discovering the right *kind* of explanations needed for the phenomenon at hand. (iii) Some models are used to justify important background beliefs for formulating an explanation.

when *explaining* phenomena. Philosophers typically describe this as explaining *with* the model itself. For instance, Bokulich claims: "[...] [O]n my view, Bohr's model [of atoms] does genuinely explain the Balmer series [...]" (Bokulich 2011, p. 44). Strevens asks "[...] how to interpret the ideal gas model, when it is proffered as an explanation of gases' Boylean behavior," (Strevens 2017, p. 38) and so forth. Such explanations are often dubbed model-based explanations or model explanations (short: ME).

Prima facie, ME are different from more familiar kinds of explanation and thus demand their own investigation. For example, ME play a crucial role for doubting the *factivity* of scientific explanation (see, e.g., Batterman 2009; Wayne 2011; Bokulich 2011, 2012; Kennedy 2012). Typically, models involve idealizations. Explanations with such models seem to involve idealizations, too. As Wayne writes (Wayne 2011, p. 831, our italics):

> Explanation in physics relies essentially on idealizations (idealized models) of physical systems, and the explanations themselves contain *false* statements about both the explanatorily relevant features of the physical system and the phenomenon to be explained.

This would violate factivity requirements on explanation, such as Hempel's requirement that "[t]he sentences constituting the explanans must be true" (Hempel 1965, p. 248).[2]

But what are the precise conditions for ME? And are ME special explanations? In what follows, we *first* clarify the notion of ME (Sect. 2) through a critical discussion of current accounts. Based on this analysis, we single out two different conceptions concerning the role of models in explanation and argue that only one of them is concerned with model explanation as a distinct kind of explanation. The other one is concerned with explanations that are *induced* from working with a model. We call them 'model-induced explanations' (MIE). *Second*, we study three paradigmatic cases of alleged ME (Sect. 3). What are their explananda? What are their explanantia? We argue that all of them are MIE—not ME—upon closer examination. *Third*, we argue that this undermines the building consensus that model explanations are special explanations that, e.g., challenge the factivity of explanation. Instead, it suggests that what is special about models in science is the epistemology behind how models induce explanations (Sect. 4).

2 Defining Model(-Based) Explanation

Models are devices scientists typically employ for examining objects or phenomena. We encounter them in many disciplines, including physics, chemistry, biology, psychology, linguistics, and the social sciences. Models are usually accounts of their so-called target objects or phenomena. But by their very nature, scientific models are not replicas or complete representations of them. As Hughes emphasizes, "[t]o have a model [...] is not to have a literally true account of the process or entity in question" (Hughes 1990, p. 71). Typically, one builds models to investigate particular features of the target phenomena. Take the double helix model of

[2] This challenge to a factive account of explanations has also been discussed by, e.g., Reiss (2012), Reiss (2013), Mäki (2013); for a critical account see, e.g., van Riel (2017), Sullivan and Khalifa (2019).

DNA as a paradigmatic example. Its target object, i.e. DNA, is modeled as having the form of a double helix. Using this model, one can explore this structural feature of DNA.

Models are construed based on *stipulations* about the target objects. For instance, the Ising model construes a macroscopic magnet as a collection of elementary magnets whose orientation determines the overall magnetization. Not uncommonly, these stipulations are *idealizations*. For instance, according to the optical Glauber model of atomic nuclei, these nuclei are perfect spheres of energy. The nature of these idealized stipulations is controversial (for an overview see, e.g., Weisberg 2007; Elliott-Graves and Weisberg 2014). For example, it is debated whether good models need to feature idealizations that can be de-idealized in the long run, whether some models involve *indispensable* idealizations, etc. Our treatment of model explanation is orthogonal to this debate.

Whereas some models are comprised of a set of mathematical equations, many models are not sentential entities. Models can also be materialized, e.g., the model of DNA can be a physical entity.[3] However, one can single out a model's *propositional content* by figuring out which propositions are true according to the model (see, e.g., Strevens 2013, p. 510; van Riel 2015, 2017; similarly Reiss 2012, pp. 49–50; Rohwer and Rice 2016, pp. 1129–1130). We can formulate the propositions that are true according to a model independent of whether the model itself contains these propositions.[4] For instance, one can state that, according to the double helix model, DNA has a helix structure, and that, according to the Ising model, macroscopic magnets consist of a collection of elementary magnets. We consider every proposition that can be singled out in that way as part of the model's propositional content.[5] Let us emphasize that this proposal does not involve the suggestion that models *represent* their target objects *by virtue of their propositional content*. The topic of whether or how models represent their target objects is a topic in its own right.[6]

As we saw before, the idea that models can explain phenomena is widespread. So, what are model explanations (ME)? Generally speaking, an *explanation* is an answer to a why-question or a how-question. For instance, one could cite a law together with other crucial conditions to answer why an event occurred. Answers to questions are standardly conceived as sets of propositions. The underlying assumption is that non-propositional methods of answering questions (e.g., nodding) could be described in terms of propositions. For instance, as Strevens emphasizes (Strevens 2013, p. 510), the content of explanations using visual information could be expressed in terms of propositions. Not everyone agrees here and perhaps

[3] We do not consider the particularity of *materialized* models here. For an overview of different kinds of models, see, e.g., Frigg and Hartmann (2012) or Gelfert (2016). One might also consider *model organisms*, such as the fruit fly *Drosophila melanogaster*, to be models (cf. Gelfert 2016, pp. 2–3). They can be considered a simplified form of the organisms in question.

[4] For an analysis of the nature of such according-to propositions, see, e.g., van Riel (2015).

[5] One issue to be discussed is whether every proposition that is entailed by a proposition that is true according to the model is also part of the model's propositional content. We remain neutral here.

[6] There is much discussion about whether or how models can be considered a *representation* of their target objects (for an overview see Frigg and Hartmann 2012; for particular accounts see, e.g., Hughes 1997; Bailer-Jones 2003; Giere 2004; Elgin 2007; Suárez 2010; Downes 2011; Frigg and Nguyen 2018), whether models are akin to fiction (e.g., Godfrey-Smith 2009; Frigg 2010; Toon 2012) or concerned with possibilities (e.g., Grüne-Yanoff 2013), etc. However, as we argue below, these issues can be separated from dealing with the nature of model(-based) explanations.

there are good reasons to allow for genuinely non-propositional explanations. But, in order to make progress on the question of model explanation, we follow Rohwer and Rice's (2016) lead and adopt the propositional account of explanation as a working hypothesis. However, we take it that the central conclusions drawn in this paper hold true even if we adopt a non-propositional account of explanation, as we indicate below.

Since not every answer to a why- or how-question counts as an explanation, one has to say more in order to define explanation. But this is not the agenda of our paper. Instead of adopting a specific account of explanation, we take a pluralistic stance. We do not presuppose that all explanations are causal. We include explanations that are typically considered to be non-causal, such as explanations of the fact that nobody can cross all of Königsberg's bridges exactly once (see, e.g., Pincock 2007; Lange 2013). In order to talk more precisely about explanation, we assume, for the sake of this paper, that explanations describe difference makers along the lines of Strevens' kairetic account (Strevens 2008). Causal explanations describe phenomena and facts that make a causal difference to the phenomenon to be explained. Other explanations might appeal to necessities to explain the phenomenon (e.g., Lange 2013). And so forth. Again, nothing hinges on our choice of this explanatory framework. As we hope to make clear throughout the paper, the issue of elucidating model explanation is not restricted to any specific conception of explanation.[7]

So, what makes an explanation model-based or a *model explanation* (ME) specifically? In the remainder of this section, we first consider two accounts of ME that are inspired by Bokulich's work (2011, 2012, 2017) and Rohwer and Rice's work (2016), respectively. We show that these accounts, as they stand, are too broad or too narrow. And we argue that one related conception of model explanation picks out what we call *model-induced* explanation rather than a distinct kind of explanation.

2.1 Model Explanation

The first definition of model explanation (ME) we consider is based on a series of papers in which Bokulich prominently analyzes ME (Bokulich 2011, 2012, 2017). A basic idea is that ME are explanations where "[...] the explanans in question makes essential reference to a scientific model [...]." (Bokulich 2011, p. 38) Scientific models are "incomplete and idealized descriptions" of a target system (Bokulich 2017, p. 104; Bokulich 2011). Bokulich initially proposed that the essential reference to a model consists in the counterfactual structure of the model being isomorphic in the relevant respects to the counterfactual structure of the target phenomenon (Bokulich 2011, p. 39):

> More precisely, in order for a model M to explain a given phenomenon P, we require that the counterfactual structure of M be isomorphic in the relevant respects to the counterfactual structure of P.

[7] For instance, all our arguments are compatible with a Woodwardian concept of explanation that focuses on counterfactual dependence (e.g., Woodward 2003). This concept is used in the model explanation literature by, e.g., Bokulich (2011), Bokulich (2012), Rice (2018, 2019b).

So, Bokulich demands that the structural features of the model be (partially) isomorphic to the relevant structural features of the phenomenon to be explained.[8] But in her latest treatment of ME, Bokulich proposes a broader analysis (Bokulich 2017, p. 104):

> Model-based explanations (or model explanations, for short) are explanations in which the explanans appeal [sic] to certain properties or behaviors observed in an idealized model or computer simulation as part of an explanation for why the (typically real-world) explanandum phenomenon exhibits the features that it does.

The first definition of ME that we consider is based on this broader analysis. Bokulich adds two additional constraints on ME (Bokulich 2011, 2012): The model user is *justified* in using the model (justification) and the model explains by capturing patterns of counterfactual dependence that hold true for the phenomenon of interest (counterfactual dependence). The justificatory step is concerned with applying the model to the phenomenon to be explained. In addition, Bokulich argues that not just any model is explanatory. She discusses reductionist models in geomorphology that are constantly improved by *eliminating* idealizations (Bokulich 2017, p. 116):

> Here one tries to simulate the braided river in as much accurate detail and with as many different processes included as is computationally feasible, and then tries to solve the relevant Navier–Stokes equations in three dimensions. These reductionist models are the best available tools for predicting the features of braided rivers, but they are so complex that they yield very little insight into why the patterns emerge as they do.

These models are still idealized in some sense (e.g., they might involve abstractions and simplifications), but Bokulich denies that these models can *explain*. She argues that heavily de-idealized models, due to their complexity, are unable to provide explanations (or, for that matter, understanding). Heavily de-idealized models cannot give us the *why* of phenomena. Thus, we should make explicit that Bokulich restricts model explanation to what one could call *substantially-idealized-model* explanation.

Bokulich's last two constraints (counterfactual dependence and substantially idealized models) are concerned with what can explain phenomena. Since we want to stay as neutral as possible in this regard, we don't add them to the first definition of ME that we discuss. Moreover, we take it that Bokulich's justificatory step is part of what it means to appeal to a model. Ensuring that one is justified in using the model for the phenomenon in question strikes us as *justifiably appealing* to a model. We further assume that the analysis of ME is meant to include properties or behaviors that *define* the model, e.g., a modeling assumption that the population of interest is arbitrarily large. Such properties or behaviors are arguably not *observed*. So, we put the first definition of ME that we consider as follows:

---

[8] Fang develops a variant of Bokulich's account according to which it suffices that the *model user* hypothesizes that the counterfactual structures applies to the target phenomenon (Fang 2019). For a model user based account of ME see also, e.g., Jebeile and Kennedy (2015).

Model explanation (Appeal): An explanation is a model explanation iff the explanation justifiably appeals to properties or behaviors that define an idealized model or are observed in it.

To illustrate such an appeal analysis of ME, Bokulich uses the example of explaining why sparrows of a certain species vary in their feather coloration from pale to dark. With the aid of a game theory model, one can demonstrate that such a polymorphism can be used as a stable and successful strategy to mark the status of the sparrows (which avoids conflicts over resources). She then writes (Bokulich 2017, p. 104)

> The model demonstrates that such a strategy is stable and successful, and hence can be used as part of the explanation for why we find this polymorphism among sparrows [...].

We think that this a good illustration of ME_Appeal. We have a model, we are justified in applying it to the phenomenon to be explained, and we end up with a successful explanans *because* we used the model. However, this example also illustrates the main issue with ME_Appeal: It allows for the model's content to *not* be contained in the explanans. Take the example of the sparrows. The only reference to the game theory model is that the model demonstrated the success of the polymorphism strategy. The fact that the polymorphism strategy is successful is part of the explanation. But this polymorphism strategy (i.e., the variation in feather coloration) need not make any reference to the model. Only the outcome of applying the game theory model—the demonstration of the polymorphism strategy's success and stability—is part of the explanation. In other words, it is not the game theory model that explains but the polymorphism strategy. Because the model's content need not be part of the explanation, ME_Appeal provides us with *too weak* or *too loose* a connection between the model and the explanation to account for ME. We need a stronger link.

We ultimately think that the explanations picked out by ME_Appeal are part of what we call *model-induced* explanation. But before we go into detail, let us turn to an analysis that ensures a stronger connection between the model and the explanation.

Rohwer and Rice (2016), in asking how models and explanations are related, describe an alternative account of model explanation. This account follows the rough slogan 'The model is the explanans.' This slogan is understood as follows (Rohwer and Rice 2016, p. 1132, our italics):[9]

> …… the propositions that constitute the model are *identical* to the propositions that constitute the explanation the modeler is interested in.

We call the propositions that constitute a model the model's *content*. For the sake of being inclusive, we assume that the model's content is either its propositional content or its

[9] The account Rohwer and Rice describe seems to be in line with van Riel's definition of ME as explanations that are true according to a model (van Riel 2017).

representational[10] content—whatever the latter is precisely. This gives us a second pass at defining ME in terms of identity.

Model explanation (Identity): An explanation is a model explanation iff the model's content is identical to the explanation (or its explanans).[11]

The essence of ME_Identity is not restricted to *propositional* contents. If there were non-propositional explanations (and non-propositional model contents) and if one defined what an identity between non-propositional contents is, ME_Identity could have a non-propositional variant.

ME_Identity seems to be a plausible explication of the common claim that models can be explanations while being compatible with other accounts of ME (e.g., Craver 2006; Kaplan 2011). ME_Identity also clearly goes beyond appealing to a model and ensures a close relation between the model and the explanation. After all, the model's content is *identical* to the explanation. But straight away, this analysis raises at least two challenging questions: (i) Is it compatible with heavily idealized models? (ii) Does the identity thesis between model and explanation hold? What if the model contains more or fewer propositions than are necessary for explaining the target phenomenon? In what follows, we discuss each in turn.

One might worry that ME_Identity is not compatible with heavily idealized models. Can the content of heavily idealized models be identical to a *correct* explanation? This worry seems to presuppose that correct explanations cannot be idealized. But it is a substantial question whether they can. ME_Identity as such is perfectly compatible with heavily idealized models being ME. According to ME_Identity, it only follows that the explanations would be idealized, too.[12]

The main question concerning ME_Identity is, assuming that we have the right conception of a model's content: Does the identity thesis between model and explanation presumed by ME_Identity hold? To begin, Rohwer and Rice straightforwardly accept that a model may have *more* propositions than is necessary for explaining a target system. At least in some cases, the explanation still contains all those extra propositions. This makes the explanation a worse one, but the explanation is still a model explanation (cf. Rohwer and Rice 2016, p. 1133). For instance, a causal model explanation that cites more facts than necessary to explain the phenomenon of interest is not as good as it could be because it does not only focus on the factors that made a difference. The explainer would do better to choose a model without extraneous propositions needed to explain.

[10] It is controversial whether models represent their target phenomena (see fn. 6). But if so: The representational content could be an explanation if it can be expressed in terms of propositions.
[11] In some cases of explanations of singular occurrences of phenomena, the model's content might not contain descriptions of the phenomenon itself.
[12] For a similar reason, we think that it is misleading to call such an analysis of model explanation a 'representationalist account of model explanation' (cf. Kennedy 2012; Jebeile and Kennedy 2015; Fang 2019). The basic idea of a representationalist account is that the model accurately *represents* the phenomena of interest (or at least a substantial part thereof). However, ME_Identity is not concerned with accurate or complete representation. Moreover, what Kennedy proposes as a 'non-representationalist account of model explanation' picks out model-induced explanations, as we argue further below.

An exception is made for idealized models. Idealized models, on their view, contain idealizations that are *not* necessary for *explaining* the phenomenon, but are nevertheless part of the model and the explanation. Examples are cases where idealizations are ineliminable, or idealizations that could be replaced with another idealization without explanatory loss (Rohwer and Rice 2016, pp. 1134–1137). This move concerning idealizations is made precisely because Rohwer and Rice want to hold on to (i) the claim that only true propositions can explain (which ME_Identity does *not* presuppose), (ii) the claim that idealized ME can be good explanations, and (iii) their identity analysis of ME. False idealizations are *extra* propositions of the model that are part of the explanation, but the success of the explanation only depends on the true propositions. Moreover, when the idealizations are ineliminable, there is no option to find an alternative model to explain with.

Take as an example their analysis of Chris Pincock's case of the 'deep water' idealization (Pincock 2012, 2014). This idealization postulates that the ocean is infinitely deep in order to model why regular wave patterns occur after irregular patterns of disturbance. The model contains the false proposition: (p1) the ocean is infinitely deep. Rohwer and Rice (2016, p. 1136) then argue that in order for the model's propositional content to be an explanation, the *model* must also contain a second, reinterpreted proposition, that is true—(p2) the depth of the ocean is above the threshold such that its particular value does not matter—in order to explain the wave patterns of interest. There is a problem here. Rohwer and Rice want to stay true to modeling practices, but the reinterpretation strategy *combined with* ME_Identity places a demand on modelers to (artificially) include propositions in their *model* that are reinterpretations of the idealizations they employ. According to ME_Identity, the model's content is identical to the explanation. So, to obtain the explanation using the reinterpretation strategy we need to change the model's content. But modelers typically don't change their models even when they know that the idealizations are not correct. They often knew this when they constructed the models. An analysis of ME should do well to capture actual modeling practices and not define ME in a way that excludes idealized models where the modeler does not do an interpretive step with their idealizations. The more inclusive reading of the deep water case (and what we suspect more closely resembles modeling practices) is that while the *explanation* might include (p2), the *model* need not.

What if there are cases where the model's propositional content has *fewer* propositions than the explananation? ME_Identity excludes cases where the explanans additionally involves propositions that are not true according to the model or just absent from the model's content from being full-fledged explanations. Such cases are only *partial* model explanations (Rohwer and Rice 2016, pp. 1138–1139). The model is necessary for the explanation, but it is not sufficient. This is a problem. According to our argument above, the case of the deep water model is arguably such a case where there are fewer propositions in the model than the explanation (if we reject the reinterpretive move), and we should want to include it as an instance of ME. Moreover, all cases where the explanation includes real-world features that are crucial for the explanation but not true according to the model are also plausible candidates for ME. Models are selective. Models do not always specify all aspects of target systems that explainers are interested in, but nevertheless the model aids in the explanation, and the explanation shares some of the model's content. On the one hand, it is *too narrow* a conception of ME to say that such explanations are not model explanations. After all, the explanation shares a crucial portion with the model's propositional content. On the other hand, in order for a partial model explanation to count as ME we need additional constraints. Otherwise, we are back to the worry that the analysis is *too broad*, rendering too many explanations ME.

Our proposal is not to abandon ME_Identity, but to revise it such that it avoids some of the aforementioned problems. We restrict the identity criterion to allow for cases where the model has more propositions than the explanation (or explanans) and cases where the explanation has more propositions than the model. Our main proposal is to focus on the *core* of the model's content, on the one hand, and on the *core* of the explanation, on the other hand. We have ME when the model's content or its core[13] is identical to the core part of the explanation:

Model explanation (Core): An explanation is a model explanation iff the model's core content is identical to the core of the explanation.[14]

According to ME_Core, the model's core content must be identical to the explanation's core in order for the explanation to be a model explanation. One might wonder about a case where the model's core content *contains* the core of the explanation but is not identical to it. Our definition excludes such cases and we think rightly so. Allowing for such cases would render the definition *too broad*. Then one could construct model explanations by creating rather arbitrary models with cores that involve the explanation's propositions but also many *irrelevant* other ones. A close connection between the model and the explanation is then lost.

A lot hinges in our definition on what constitutes an explanation's *core*. In this paper, we don't offer a full account of the core of an explanation or model. However, there are some general intuitive principles that are helpful here and can serve to motivate the remaining discussion in the paper.

First, consider the core of the explanation. What constitutes the core of the explanation depends on the kind of explanation. In the case of a law-based explanation, the citation of the law and the law's application conditions are arguably the core of the explanation. In the case of a mechanistic explanation, the description of crucial parts of the mechanism constitute the explanation's core. *Second*, we cannot simply define the core in terms of the sheer amount of propositions. A law-based explanation might only consists of a few propositions. *Third*, there needs to be a non-trivial relationship between the propositions in question and the explanatory power of the explanation. While certain boundary conditions might be necessary for entailing the explanandum, it is not central to the explanatory power of the explanation. For example, Sullivan (2019) argues that in order to delineate a causal explanation from a non-causal explanation, one must identify the 'primary reason' an explanation succeeds, and that boundary conditions are unlikely candidates. This sense of explanatory importance is what we mean by the core. For

[13] Recall that the propositional content of a model might include all the entailed propositions, as well (cf. footnote 5).
[14] ME_Core and ME_Identity are concerned with the case of a *single* model. In cases where one explains a phenomenon using multiple models at the same time, one would need to revise the definition such that a conjunction of the models' core contents is identical to the core of the explanation. (Note that multi-scalar models with inconsistent sub-model assumptions typically explain different *aspects* of a larger phenomenon and thus do not provide a joint explanation.)

example, in a causal-mechanistic explanation of an event, the causal mechanism is the core of the explanation with the peculiarities of the event in question being in the periphery.

Fourth, when we move to whether an explanation is a *model* explanation, we need to see whether the core propositions of the explanation play a non-trivial role in the model that the explanation refers to. Specifically, we want to exclude the possibility that an explanation is a model explanation simply in virtue of relying on a generic proposition that also just happens to be true of many other models. For example, many optimality models in biology rely on infinitely-sized populations. However, if a given explanation includes this idealization it is not thereby based on all possible optimality models or on all models that assume infinitely-sized populations. It is not just that the proposition in the explanation must play a crucial role in its explanatory power, but the same proposition must be central to the model in question. It needs to be a proposition that is *entrenched* with the other propositions of the model in such a way that it is recognizably doing real work in the model, e.g., the proposition uniquely discriminates the model in question, and cannot be easily separated. ME_Core captures the deep water case without arbitrarily stipulating whether interpretive idealizations are or are not part of a model.

Fifth, the notion of a model's core might be illuminated with so-called *robustness analysis*.[15] Roughly speaking, this is the study of similar, but distinct, models of the same target phenomenon. The basic idea is that if such models lead to similar results, we can "...separate the scientifically important parts and predictions of our models from the illusory ones that are accidents of representations," as Weisberg puts it (Weisberg 2006, p. 731). For instance, Woodward highlights that robustness analysis might lead to identifying casual relationships which are stable or invariant under changes (Woodward 2006, p. 235). The elements of the model that are robust or stable are arguably an element of the model's core. And the 'illusory ones' would be part of the model's periphery. Kuorikoski, Lehtinen, and Marchionni add further considerations about how to distinguish between the core of a model and its periphery (Kuorikoski et al. 2010). They separate what they call *substantial model assumptions* from assumptions that idealize "... away the influence of the confounding factors ..." (Kuorikoski et al. 2010, p. 547) and assumptions that need to be added to render the model mathematically tractable. The latter two kinds of assumptions could be described as the model's periphery. But whether models can be decomposed in that way is controversial (see, e.g., Rice 2019a). Either way, robustness analysis can contribute to sharpening the notion of a model's core.

Now that we have a more promising definition of ME, we can go back to the thesis that Bokulich's analysis might not be concerned with model explanation but with a related conception of models and explanations.

2.2 Model-Induced Explanation

When it comes to scientific models, an important relation is what Rohwer and Rice call an *epistemological* relation between a modeler, a model, and an explanation (Rohwer and Rice 2016, sect. 3). For instance, they describe cases where models are aids to discovering

[15] We thank an anonymous reviewer for this suggestion. For details on robustness analysis, see, e.g., Wimsatt 1981; Orzack and Sober 1993; Levins 1966; Weisberg 2006; Woodward 2006; Kuorikoski et al. 2010; for a critical view see, e.g., Odenbaugh and Alexandrova 2011).

explanations by helping to identify the kind of explanation needed for the phenomenon of interest (*ibid.*). In what follows, we argue that several conceptions of model explanation pick out a particular kind of epistemological relation rather than a distinct kind of explanation, namely what we call *inducing explanation*. We call the explanations that feature this relation *model-induced explanations* (MIE).

As we argued above, ME_Appeal is too weak or too loose a definition of model explanation. However, ME_Appeal captures an important aspect of many explanations discussed in the literature: The cases that drive the debate are concerned with models that seem to be *epistemically crucial* to the explanation. In contrast to merely using a model as a tool to look for the right (kind of) explanation of the phenomenon of interest (a case discussed by Rohwer and Rice 2016, p. 1141), the models of interest play an *enabling* role. Without using the model, the explanation would have not been discovered.[16] So, the models play a crucial role in the process of obtaining the explanations (see also Lawler 2019). We focused above on Bokulich's proposal, but this conception of model explanation is visible throughout the literature.[17] Take, for example, Graham Kennedy's 'non-representationalist' account of model explanation (Kennedy 2012, pp. 331–332, see also Jebeile and Kennedy 2015):

> Comparison cases explain by allowing model users to identify those factors which make a difference to the behavior of the modeled target system. [...] This type of explanation occurs with many scientific models. [...] In cases where the actual value of a variable is known, a non actual or false value can be used to generate a comparison with the more realistic case. In cases where the actual value of a component is not known, two non actual limiting comparison cases can be used to encompass the (unknown) actual value. These comparison cases allow the user to learn about the behavior and/or evolution of the phenomenon in question, and thereby to explain components of the target system being modeled.

What Graham Kennedy's and Bokulich's analyses have in common is that the relation between the explanation and the model is located in the *process of obtaining the explanation*. The model is claimed to play an important (if not even an essential) role in that process. The game theory model establishes the success of the polymorphism, which can then be used to explain *the why* of the variation in feather coloration. When Kennedy argues for her 'non-representationalist' account of model explanation, she also emphasizes this role of the model. She describes two astrophysics models which help the model user to arrive at the explanation of interest by functioning as a comparison case for the phenomenon at hand. Graham Kennedy writes (Kennedy 2012, p. 331):

[16] As one reviewer remarked, another interesting epistemic role might be the role of models in *justifying* the explanations of interest. Discussing the relation between justification and models is a topic in its own right. We don't discuss it here.

[17] Marchionni, for example, describes conceptions of ME as being between two opposite sides of a continuum (cf. Marchionni 2017, p. 609). We think that they are better described as two different conceptions for the reasons given in what follows.

> The simplified two-dimensional models are themselves required for explanation because they enable the scientists to identify which factors make a causal difference to the evolution of the disks.

Let us suppose that she is right about her case studies. Let us suppose that the models are *required* for the desired explanation. Even if that were true, this does not mean that the model itself explains the phenomenon of interest. The claim that the model is required for obtaining the explanation is merely a claim about how one arrives at the explanation. This claim is compatible with the explanans not making *any* reference to the model. For instance, in Kennedy's example the models help to identify the causal difference makers for the evolution of the disks. But only the latter need to be cited in the resulting explanations. So, Graham Kennedy's 'non-representationalist' account of model explanation turns out to be not about model explanation. Instead, this account is better described to be about what we call *model-induced explanation*. The resulting explanation is closely related to the model because *working* with the model opens up new *epistemic* perspectives. The explanation is induced by working with the model. Graham Kennedy hints at this epistemic function when she writes (Kennedy 2012, p. 327, italics omitted):

> I propose that, in many cases, the idealizations within scientific models play a more active explanatory role, by allowing scientists to determine what is causally relevant.

And in a later paper with Jebeile she claims (Jebeile and Kennedy 2015, p. 384):

> [...] idealizations [in models] should be seen as having an active role in making possible the identification of explanatory components in models.

This enabling function is crucial and should not be neglected. But it is important not to conflate it with the results of utilizing it (i.e., the obtained explanations) (see also Lawler 2019). Elgin describes this function of models (and of idealized scientific devices more generally) as providing *epistemic access* to the phenomenon of interest: "Each model exemplifies different features and affords epistemic access to different aspects of the target." (Elgin 2017, p. 267) So, different models open up different epistemic perspectives on the phenomenon. Constructing and working with the model highlights aspects of the phenomenon that are otherwise difficult to examine or describe. One plausible reading of what Graham Kennedy describes is that she illustrates an instance of this general function highlighted by Elgin. Also, other claims about model explanation can plausibly be read as claims about model-induced explanation. Take as an example Rice's analysis of how idealized models can explain. He writes (Rice 2018, p. 2803, our italics):

> Only by pervasively distorting the features of real-world systems can physicists apply the mathematical modeling techniques required to provide *epistemic access to the explanations we seek*.

This case sounds like a paradigmatic case of a model-induced explanation.

It goes without saying that this relation between models and explanations is an interesting one. However, it is important to flag that it as a largely *epistemic* one. The model plays no more or less than an important role in what one might want to call the *discovery* of the explanation in question. One should not conflate this enabling role of models with model explanation. The mere epistemic role of models falls short of philosophers' ambitions when they discuss model explanation. In order to keep track of the difference between these two conceptions, we define 'model-induced explanation' (MIE) as follows:[18]

Model-induced explanation: An explanation is model-induced iff constructing or using the model constitutes a decisive part of arriving at the explanation.

Model explanation (Core): An explanation is a model explanation iff the model's core content is identical to the core of the explanation.

To illustrate the importance of this distinction, take Morrison's claim about what makes models explanatory (Morrison 1999, p. 63):

> The reason models are explanatory is that in representing these systems, they exhibit certain kinds of structural dependencies.

Her claim can mean substantially different things. Exhibiting structural dependencies does not mean that these are identical to the dependencies cited in the respective explananatia; they could merely point to them. If Morrison only requires that the dependencies exhibited by the model have this pointing function, she is concerned with model-induced explanation. If, instead, she requested the isomorphism relation, she would be concerned with model explanation.

In what follows, we analyze paradigmatic cases of (alleged) model explanation in the literature to examine whether they are ME or merely MIE.

3 Paradigmatic Cases: Bees, Fluids, and Rainbows

There are at least three kinds of models that are frequently discussed in the literature on scientific models and explanations: optimality models, phase-transition models involving the thermodynamic limit, and models that are analyzed as fictions rather than idealizations. For each kind, we analyze one paradigmatic case that is claimed to provide explanations: models for explaining the honeybee foraging behavior (used in Rice 2016), lattice gas models for explaining patterns of fluid flow (used in Batterman and Rice 2014; Rice 2018), and models for explaining supernumerary arcs of rainbows (used in Batterman 2005; Pincock 2011; Saatsi (forthcoming)).

[18] This distinction is roughly related to Rohwer and Rice's proposal to draw "[...] a distinction between a model being a stand-alone explanation [model explanation] versus merely being explanatory [model-induced explanation]" (Rohwer and Rice 2013, p. 335). But their notion of an 'explanatory model' is much weaker than our notion of a model-induced explanation. According to them, "[a]n explanatory model is one that produces scientific understanding relevant to answering a why question [...]" (Rohwer and Rice 2013, p. 335). By contrast, we demand that the results of working with the model are parts of the answers to the why-question and that using the model is decisive for obtaining the answers.

In this section, we ask whether these paradigmatic cases are genuine instances of ME. Our guiding questions in what follows are (a) What is explained, i.e., what is the explanandum-phenomenon? (b) What is the explanation? (c) How does the model figure into the latter?[19] The upshot of our analyses is that the alleged ME turn out to be MIE. As we argue in the next section, this fact undermines the building consensus that model explanations are special explanations that, e.g., challenge our standard concepts of explanation.

3.1 Bees

In the realm of biology, so-called *optimality or optimization models* are frequently used (cf. Rice 2012, 2018, sect. 3.2; Elgin and Sober 2002, pp. 446–448; Potochnik 2007, 2009, 2010; Bokulich 2017, pp. 104–105).[20] These are models that highly idealize their target objects or phenomena. As Rice puts it (Rice 2018, p. 2808),

> [such models] [...] pervasively misrepresent the features and processes of their target system(s), including those that are assumed to be the difference makers for the target explanandum [phenomenon].

The basic goal of optimality models is to analyze why particular phenotypic traits occur. They do so by determining optimal strategies for obtaining particular features, such as the net energy intake given a set of limiting factors and trade-offs such as the costs of finding or consuming food. In order to determine the optimal strategy, such models don't simply involve some distorting idealizations; they involve distorting idealizations for the most part. An example Rice gives is a model of the foraging behavior of honey bees (as presented in Schmid-Hempel et al. 1985). It is assumed, among other things, that the honey bee population is arbitrarily large, that there is no intergenerational overlap, that the selection pressure in the honey bee population remains constant, etc. (cf. Rice 2016, p. 89). In short, there are barely any propositions that are true according to the model that are actually true. Moreover, as in the case of any optimality model, it is assumed that natural selection is the only evolutionary factor that matters for the phenotypic trait's evolution.

How can biologist explain with such a model? In the honey bee example, the explanandum-phenomenon is the fact that honey bees tend to leave food sources when their crops (i.e., their honey sacks/stomachs) are only *partially* filled. This is a puzzling fact because one would expect them to fill it completely (or at least as much as possible). The explanation for this behavior is that the honey bees maximize their energy efficiency rather than the rate of energy intake (cf. Rice 2016, p. 89). The foraging pattern seems to be an adaptive response to a trade-off between energy efficiency maximization and energy intake rate maximization. Visiting more food sources would reduce their energy efficiency. That is why honey bees leave them when their energy intake is high enough.

---

[19] A brief methodological remark: Philosophers when discussing idealized models often assume that scientists actually succeed in doing what they claim they do. In particular, they take for granted that scientists correctly explain with at least some models (cf., e.g., Wayne 2011, pp. 831–832; Rice 2018, p. 2799). In this paper, we do not discuss whether this assumption is apt. Instead, we evaluate a conditional question: If scientists provide us with correct explanations: How do the models figure into such explanations?

[20] Optimization models are also used in other disciplines, as Rice points out (Rice 2018, p. 2803).

How does the model figure into the latter? The core of the explanation is constituted by the trade-off claim. This claim, in turn, is the result of the above described optimality model, according to which the honey bees maximize their energy efficiency (cf. Rice 2016; Schmid-Hempel et al. 1985). The patterns that this model predicts are very similar to the patterns observed for the real-world honey bees' foraging behavior. Alternative models did not reproduce these patterns. So, it seems that the model's stipulation that honey bees maximize their energy efficiency is correct. The model seems to capture correctly this fact of the foraging behavior.

The model seems to be essential for the explanation (at least at that time). It is the one that produces the observed patterns. It is also true that the trade-off claim is part of the model's content. After all, this claim is true according to the model. So, it looks as if the explanation is a ME. However, let us not jump to a conclusion here. Recall our analysis of ME:

Model explanation (Core): An explanation is a model explanation iff the model's core content is identical to the core of the explanation.

The condition for ME is not fulfilled, upon closer examination. As Rice argues in detail (Rice 2016, 2018), the model's trade-off claim *cannot be quarantined* from the idealized stipulations mentioned above, such as the claim that the honey bee population is arbitrarily large. Only the stipulations taken together have that claim as a result. So, arguably the core of the *model*'s content contains at least a substantial amount of these stipulations; the core is not just constituted by the trade-off claim. Importantly, none of these idealized stipulations are part of the explanation; only the trade-off claim is. The explanation does not involve the assumptions that the honey bee population is arbitrarily large or that there is no intergenerational overlap. At best, the explanation and the model both involve the claim that evolution leads to traits that maximize energy efficiency.[21] But that would not render it a ME. The model's core content is not identical to the core of the explanation. The core of the model, but not the explanation, involves the crucial idealizations. If so, the honey bee case is not a case of a ME. The model itself does not explain the honey bees' foraging behavior.

The honey bee case can, however, be analyzed as a MIE, i.e., a model-induced explanation. Constructing the energy maximization model constituted a decisive part of arriving at the trade-off claim. The model induced the explanation, so to speak. It played a substantial role in arriving at the explanation.

3.2 Fluids

In physics, we also encounter idealized models, such as models involving the thermodynamic limit. "This widely used modeling assumption is the limit in which (roughly speaking) the number of particles of the system approaches infinity," as Rice puts it (Rice 2018, pp. 2800–2801). The volume of the system is assumed to go to infinity, as well. Models that employ the $n\to\to\infty\infty$ assumption and the $V\to\to\infty\infty$ assumption are so-called phase-transition models. Phase transitions are abrupt changes of the qualitative macroscopic properties of a system or substance, such as water's freezing into ice, the transition from liquid to gas, or the

---

[21] We thank an anonymous reviewer for emphasizing this point.

magnetization of iron. The thermodynamic limit is claimed to be essential for such models because the phenomenon of a phase transition cannot be *produced* with a model that assumes finite particles. We cannot model phase transitions by employing finite systems, say, systems based on statistical mechanics.[22] So, it seems that we cannot explore phase transitions without taking for granted the thermodynamic limit. Such models are hence a promising candidate for ME.

An example of models that employ the thermodynamic limit are particular lattice gas automaton models that model fluid flow. These models not only employ the thermodynamic limit but also involve several other distorting idealizations. According to such models, fluids consist of point particles that could move in just six directions and only on a hexagonal lattice. Despite these utterly false assumptions, an application of a lattice gas automaton model *reproduces* macroscopic behaviors of real-world fluids to a relevant degree of similarity (for details see, e.g., Batterman and Rice 2014; Rice 2018).

So, how can we explain with such a model? In the fluid flow example, the explanandum-phenomenon is the fact that the momentum density profile in a pipe is parabolic. Batterman and Rice propose that this fact can be explained by the patterns resulting from the lattice gas automaton model (Batterman and Rice 2014). Their basic idea is that such a 'model explanation' is possible when the use of the model's idealizations can be *justified* (Batterman and Rice 2014; Rice 2018). As Rice puts it (Rice 2018, p. 2796, our italics):

> [...] [H]ow can models that provide holistically distorted representations explain? In order to answer this question, I will propose an alternative method for *justifying* scientists' use of idealized models to explain [...].

The idea that justification plays an important role for ME is also advocated by Bokulich (2011), as we mentioned before. According to her, the justification consists of specifying the domain of applicability of the model and to show that the phenomenon to be explained falls within that domain (*ibid.*). Rice and Batterman's proposal differs from that. The basic idea is as follows (Rice 2018, p. 2796):

> [...] I will propose an alternative method for justifying scientists' use of idealized models to explain that appeals to *universality*: the fact that systems with (perhaps very) different physical features will display similar patterns of macroscale behavior.

So, their proposal is that we are justified in using the lattice gas automaton model because it and the fluid are in the same *universality* class (Batterman and Rice 2014; Rice 2018; Batterman 2009, pp. 437–438). Universality is the fact that very different systems display highly similar macrobehaviors despite their differences. One example for this are phase transitions.

[22] There is a debate about the thermodynamic limit in philosophy of physics. Some argue that it is dispensable (e.g., Butterfield 2011; Norton 2012; for an overview see, e.g., Shech 2017). For some useful discussion see also, e.g., Shech (2013), Feintzeig (2017). For the sake of argument, we take for granted here that the thermodynamic limit is necessary.

Very different entities, such as fluids and ferromagnets, can undergo phase transitions that are remarkably similar in their features.[23]

So, why is universality considered to show us how models can explain something? The idea is as follows. *First*, we have to establish an appropriate *link* between the results of the idealized model and the real-world phenomenon of interest, e.g., a link between phase transitions that result from the model application and the real-world phase transitions. The link is that the idealized model and the real-world physical system are in the same universality class (Rice 2018, p. 2812).

Second, that the real-world system and the idealized system are in the same universality class suggests that "[...] the stability of such macobehaviors [sic] is due to the fact that the features [...] are largely independent of the details of the components or dynamical processes that operate in the system" (Rice 2018, p. 2813, italics omitted). In other words, we can conclude that "[...] many of the details that distinguish the physical systems from one another are irrelevant for their universal behavior [...]" (Batterman 2002, p. 42). This shall give us a good enough reason to believe that genuine idealized models are explanatory. As Batterman and Rice put it (Batterman and Rice 2014, p. 356):

> The models are explanatory in virtue of there being a story about why large classes of features are irrelevant to the explanandum phenomenon.

Or as Rice puts it (Rice 2018, p. 2816):

> [...] [T]he reason these idealized models are able to explain is that, as long as the system is within the relevant universality class, most of the physical details of the system are irrelevant for the occurrence of certain universal macrobehaviors.

Yet, this cannot be the whole story. That some features are irrelevant does not explain the phenomenon of interest. In his 2018 paper, Rice indeed limits his claim about the role of universality to the claim that appealing to universality can *justify* the use of idealized models to explain phenomena (Rice 2018). However, Batterman and Rice make further statements about in virtue of what facts they consider genuine idealized models to be explanatory. They put it as follows (Batterman and Rice 2014, p. 363):

> A derivative, or by-product, of this [universality] analysis is the identification of the shared features of the class of systems. In this case, the by-product is a realization that all the systems within the universality class share the common features locality, conservation, and symmetry. [...] This answers [the] question ['Why do very different fluids have features, such as symmetry, in common?'] and provides, given the answer to [the question 'Why are the heterogeneous details irrelevant for the occurrence of the phenomenon?'], an answer to [the question 'Why are the common features necessary for the phenomenon to occur?'].

Let us suppose they are right. The argument from universality then gives us the following:

[23] The same holds true for certain models in biology (cf. Batterman and Rice 2014, Sect. 4; Rice 2018, p. 2802).

> (i) The fact that the model and its target system are in the same universality class justifies using the former to explore the latter.
>
> (ii) The fact that different kinds of real-world systems are in the same universality class explains why their "[....] patterns occur across such varied physical system [sic]" (Rice 2018, p. 2802).
>
> (iii) The fact that different kinds of real-world systems are in the same universality class explains why they share some features.
>
> (iv) The fact that different kinds of real-world systems are in the same universality class explains why the common features of systems in a universality class are *necessary* for the pattern of interest to occur.

It goes without saying that these are important results. However, none of them in isolation or taken together gives us the desired explanation or justifies treating the models as providing us with ME. Recall that the explananda of interest are facts like the fact that the momentum density profile in a pipe is parabolic or other facts about features of fluids, liquids, etc. Neither (i), (ii), (iii), nor (iv), nor a combination of (i)–(iv) explains these facts.

Take (i): We gladly accept that the idealized model being in the same universality class as the system to be explained justifies the exploration of the latter with the former. However, this only justifies the *use* of the idealized model. It does not give us any explanation yet or show that the model's core is identical to the explanation's core.

(ii) might give us an explanation. But (ii) is at best an explanation for the question 'Why do similar patterns occur across different fluids?'[24] On the one hand, this question is substantially different from the question of interest, namely, say, 'Why is the momentum density profile in a pipe parabolic?' On the other hand, the resulting explanation is clearly *not* a ME. The explanation is that all the different fluids are in the same universality class. Such an explanation contains no reference to a model in any interesting sense. The same holds true for (iii). Indeed, being in the same universality class might be relevant for explaining the commonalities of features. But this answers the question 'Why do very different fluids share features X, Y, Z?' and not the questions of interest and it doesn't seem to involve the *model*'s content.[25]

We are somewhat skeptical that the argument from universality gives us (iv). But even if it does, no ME is obtained. (iv) addresses the question 'Why are the common features necessary for the phenomenon to occur?' This is an interesting question and an answer to it might constitute *part* of an explanation for why the pattern of interest occurs. But, on the one hand, this does not suffice for ME. It does not give us the *core* of the desired explanation. The latter arguably consists of more than a list of some necessary features. On the other hand, these

[24] (ii) might also give us a good reason to believe that we only need *one* explanation for the variety of the systems which exhibit the pattern.

[25] Batterman and Rice also suggest that we can explain particular behaviors of fluids by pointing out that the fluids are in a particular universality class where all members exhibit these behaviors (Batterman and Rice 2014, p. 364). But, again, the explanatory information is the membership in the universality class and not some model information.

necessary features themselves do not make a reference to the model. They seem to be independent of the model. So, (iv) does not give us ME, either.

What about a combination of (i)–(iv)? (i) is merely concerned with the justificatory step. (ii) and (iii) have closely related *explananda*. (ii) might answer 'Why do similar patterns occur across different fluids?' and (iii) might answer 'Why do very different fluids share features X, Y, Z?' The combined answer can be further connected with the result of (iv), namely that the common features of systems in a universality class are *necessary* for the pattern of interest to occur. Recall that our question of interest is a question like 'Why is the momentum density profile in a pipe parabolic?' (ii)–(iv) taken together also don't provide us with a ME. Only knowing of *necessary* conditions of the momentum density profile and knowing that these conditions are shared with systems in the same universality class doesn't give us a full explanation. But let us suppose that they do or that they can be combined with other information to arrive at a full explanation. Even if so, it has only been established that the explanatory decisive information is the membership in the universality class. We don't have evidence that the model's core content is identical to the explanation's core.

So, none of the explanatory virtues of universality seems to lead to ME. However, models featuring the thermodynamic limit can clearly provide us with model-induced explanations. In fact, we think that Batterman and Rice's analyses are best construed as analyses of MIE. Consider how Rice substantiates the universality claim. He illustrates it by means of the example of the discovery that melt ponds are in these same universality class as other systems that undergo phase transitions. He concludes (Rice 2018, p. 2816, our italics):

> [...] by discovering that these melt ponds are in the same universality class as other physical (and model) systems, these modelers were able to apply various mathematical modeling tools (e.g. homogenization) to *extract explanatory information* about real-world systems without having to accurately represent the entities, processes, or ontology of those systems. In this way, *these mathematical modeling techniques enabled access to explanations* and understanding that would otherwise have been inaccessible.

These extracting and epistemic access functions are at the heart of MIE. One extracts explanatory information by working with the model and one gains epistemic access to explanations. Moreover, the claim that—by means of universality—one can identify which features are necessary for the phenomenon to occur also fits the conception of MIE better than the conception of ME. Hence, we think it is safe to conclude that one can obtain MIE with the aid of universality but not ME.

3.3 Rainbows

Lastly, consider a candidate of ME that involves, in Bokulich's terminology (Bokulich 2012), an *explanatory fiction*: An explanation of the supernumerary arcs of rainbows. In cases where light waves moving through the raindrop exhibit constructive and destructive interference, extra bands of color can form inside the primary rainbow, with some space between the primary bow and the extra bands. These extra bands are supernumeraries.

The best explanation of why supernumeraries form requires reference to features of the wave theory of light and the fictitious ray theory (Batterman 2002; Pincock 2011; Saatsi (forthcoming)). Saying the *best* explanation here is not accidental. There are more complex computational models that also capture supernumeraries without appealing to light rays. In particular, the Lorenz-Mie model is able to capture the phenomenon utilizing electromagnetic theory. However, just as Bokulich is skeptical that the hyper-realistic models of braided rivers are explanatory, since they fail to provide *the why*, so too Batterman (2002), Pincock (2011), and Saatsi (forthcoming) argue that the Lorenz-Mie model fails to provide us with *the why* of supernumeraries. Instead, it is argued, the complex angular momentum approach (CAM), which utilizes the fiction of light rays, provides us with the best explanation.

Following Pincock (2011), the size of β—a dimensionless parameter that is the product of the wavelength number, k (2π/wavelength), and the radius of the raindrop, a—determines the rainbow patterns that emerge. For example, if β is too small, then a rainbow is not observed, or certain colors may be distorted. The ray representation results from the wave representation when $\beta \rightarrow \infty$ In this case, the wave-theoretic aspects of the light are not relevant to trace the path of the light through the raindrop. In other words, when the wavelength of light is much smaller than the radius of the raindrop, the dominant contributions to the light begin to approach the behavior of rays instead of waves (Pincock 2011, p. 19). Importantly Pincock (Pincock 2011, p. 16) notes that:

> we do not represent the wave crests as forming a continuous straight line, but only claim that the distance between crests is so small with respect to the radius of the drops that it is not relevant to the path of the wave.

This means the ray theory helps to give a useful frame for understanding light's behavior, even though the ray theory ignores key aspects of the characteristics of light, such as the way that light is diffracted by a sphere. Interestingly, in explaining supernumeraries, one needs to incorporate the interference and diffraction effects provided by the wave-theory, while also incoporating the ray-theoretic representation. In particular, the CAM method provides a "rapidly converging expression in terms of 'poles' and 'saddle points' in a complex-valued angular momentum space, representing the main contributions to the scattering amplitude at the primary rainbow angle" (Saatsi (forthcoming), p. 12). Saddle points occur where the first derivative of the scattering amplitudes S with respect to λ (the angular momenta of the components of the light that hits the drop) is 0. Poles, on the other hand, occur where S lacks a derivative of some order (Pincock 2011, p. 18). These saddle points and poles play a different interpretive function in the mathematical theory. Saddle points correspond to rays which appear more sharply as the ratio between the raindrop radius and the wavelength increases. On the other hand, poles correspond to waves, pointing to the importance of diffraction (Pincock 2011, p. 19). This, what Pincock refers to as an 'interpretive conjecture,' is what allows us to plainly see what the overall scattering process corresponding to supernumeraries depends on.

This brings us to the question: What role does the ray model play in explaining the supernumeraries? No doubt, the ray fiction plays an important role in isolating which explanation variables crucially explain the explanandum (Batterman 2005; Pincock 2011; Saatsi (forthcoming)). Without the light ray model, we would not be able to see the fundamental difference makers in the sea of Mie computations. However, does this role go beyond model-induced explanation (MIE)?

Pincock describes the ray model as playing a largely *interpretive* function (Pincock 2011, p. 19):

> A scientist must ascend from the wave theory to the ray representation before she is able to get the 'physical insight' into the supernumeraries which CAM provides. This does not mean that she must believe the ray theory is correct. Instead she must use the results of one idealization to inform the proper interpretation of another idealization.

The ray model allows us to gain an understandable interpretation of the physical behavior of light in the context of other idealizing assumptions (e.g., the introduction to the limit). Saatsi shares Pincock's interpretation that the light ray fiction is an interpretive exercise. He says of CAM's improvement that (Saatsi (forthcoming), p. 12, original italics):

> This improvement is not a matter of introducing new variables that ontologically transcend the Lorenz-Mie theory (cf. Pincock 2011). Nor is it a matter of providing more fine-grained information about the explanatory dependence. Rather the improvement has to do with the way in which the CAM approach defines critical explanation variables upon which the explanandum depends *in a simple way*.

On this interpretation, the propositions of the fictitious ray model are not part of the explanation that explain supernumeraries. Instead, it is an interpretive frame for understanding certain behavioral and mathematical realities of light explained in terms of other concepts (e.g., saddle points). Thus, the ray model plays an interpretive epistemic role in understanding how key concepts relate in an explanation, but it is not a core aspect of the explanation itself. The variables in the explanation without this interpretive gloss would still do the same explanatory work. The light ray model extracts how we should think about the explanation variables, but it is not identical to any part of the explanation. The "association between the saddle points and rays lets us appreciate how the light behaves in some respects as the ray theory would predict" (Pincock 2011, p. 19), in a way that furthers an epistemic aim, such as understanding, but is not part of the explanation in any proper sense. Thus, the ray model plays a genesis function characteristic of MIE, not of ME.

Batterman (2005), on the other hand, takes the ray model to be setting the boundary conditions of the explanation. This is more promising for ME. If the ray model is part of the boundary conditions, then perhaps this is enough to be an instance of ME. Batterman says (Batterman 2005, p. 159):

> In order to see what boundary conditions to impose on the partial differential equation in the first place, we must conceptualize the problem as one in which (to a first approximation) we are considering specular reflection off the back of the raindrop. It involves, that is, thinking about light behaving as rays on the physical boundaries. Without the physical interpretation to begin with, we would not know what boundary conditions to join to the differential equation. Neither, would we know how to join those boundary conditions to the equation. Put another way, we must examine the physical

> details of the boundaries (the shape, reflective and refractive details of the drops, etc.) in order to set up the boundary conditions required for the mathematical solution to the equation.

Notice though that Batterman does not go so far to say that the ray model is part of the explanation. On the contrary, the ray model is 'setting up' what is needed to solve the equation and generate the explanation. It is not the propositions of the fictitious ray model that are part of the explanation. The model gives us a physical interpretation to extract the necessary boundary conditions that later figure into the explanation (i.e., the shape and size of the raindrops and their reflective details). So again, we fall short of ME. The fiction is the genesis of explanatory information, namely picking out what information is explanatorily relevant, but the fictitious model is no way identical to the explanation, even in our restricted sense.

Batterman's interpretation is different from Pincock's and Saatsi's. For Pincock and Saatsi, the interpretive role the ray model plays seems largely secondary to the the mathematical model, as a step to improve understanding. Whereas on Batterman's view, the ray model uncovers the boundary conditions for the start of a possible explanation. However, in both cases, the resulting explanation utilizing the ray model is an instance of MIE, not an instance of ME.

3.4 Upshot of the Survey

The result of our survey of paradigmatic alleged ME is that all of them turn out to be MIE. In each case, the relevant explanatory information is *independent* of the model but only closely intertwined with the model due to the history of obtaining the information. In the honey bee case, it is the information that the honey bees seem to maximize their energy efficiency rather than the rate of energy intake. In the fluid flow case, we learn which physical details are irrelevant and which ones are necessary for the phenomenon of interest. In the rainbow case, working with the fiction of light rays provides us with necessary boundary conditions (Batterman 2005) and an interpretive framework for a mathematical theory (Pincock 2011).

The observation that the explanatory information is information independent of the model is not limited to the examined cases. We chose them because they are paradigmatic cases of alleged ME. What seems to be a model explanation turns out to be a model-induced explanation. For instance, the optimality model of the eider duck's foraging behavior (Rice 2018) also at best provides us with MIE, but not with ME. We expect to get similar results for other alleged cases of ME. If so, we don't have a case of ME yet. As we argue in the next section, this suggests that what is special about models in science is how they induce explanations.

4 Are Model Explanations Special?

So far, we have argued that there are two different conceptions of explanations using models. Model-induced explanations are explanations where constructing or using the model constitutes a decisive part of arriving at the explanation. Model explanations are explanations where the model's core content is identical to the core of the explanation. In this section, we argue that both are not special from an explanatory point of view.

Although it is a special characteristic of model(-induced) explanations that they are closely related to a model, this does not necessarily render such explanations special *qua* explanation. ME can simply be instances of more general kinds of explanation. For instance, Bokulich prominently introduces a taxonomy of model explanations that reflect familiar kinds of explanation (Bokulich 2011, sec. 2 & 3): According to her, *mechanistic model explanations* are particular mechanistic explanations, namely descriptions of mechanisms based on a model. *Covering-law model explanations* are particular covering-law explanations, namely explanations which, *inter alia*, cite model-based laws in their explanantia. *Causal model explanations* are particular causal explanations, namely explanations where one explains observed features by postulating underlying structures whose features are causally responsible for the properties. *Structural model explanations* are particular non-causal explanations, namely explanations where the explanandum-phenomenon is shown to be a consequence of particular structural features of the theories employed in the model.[26] Importantly, in none of these cases is the *model* aspect of the explanation doing much work. On the contrary, it is the causal, law-covering or structural aspect that makes the explanation a 'special' kind of explanation demanding its own treatment.

Moreover, without loss, this taxonomy could be also used as a taxonomy of model-induced explanation. All the categories apply equally well to MIE. For instance, the model-based law, or non-accidental regularity, could be the one that is discovered by working a model, such as the regularity that honey bees tend to maximize their energy efficiency rather their energy intake. So again, there is nothing about the role of models that makes the explanation *qua* explanation special or different.

At the very least, it seems that in order to show that an explanation constitutes a new kind of explanation one needs to show that either a unique kind of why- or how-question is being asked, or that there is a fundamentally different way to answer why- or how-questions. However, our interlocutors have not given us that much. Instead, we suggest that what philosophers of science take to be interesting about the models that we discussed is actually the *epistemology* behind how models help produce explanations. MIE explanations are obtained in close relationship with working with a model. Models figure in the *process* of obtaining the explanation and they might even be *required* in that process. Without the model we might not arrive at the explanations, because we lack the crucial epistemic access to the desired information. For instance, Marchionni emphasizes that the models she examines crucially depend (Marchionni 2017, p. 606, our italics):

> [...] on assumptions known to be false [...] such assumptions are *indispensable* for the derivation of the results.

We think that such features make explanations with the aid of models unique. But it is important to not conflate the special features of the discovery of an explanation with the features of the explanation itself. For instance, special features of the discovery of laws are not special features of explanations using these laws.

---

[26] The more precise definitions of these kinds of model-based explanations are not important here and we also do not discuss the taxonomy's adequacy. For criticisms of Bokulich's account of structural model explanations, see, e.g., King (2016).

The distinction between the explanation and how we arrive at it has implications for not only our taxonomies of explanations, but also for larger debates about the nature of explanation and scientific practice, especially the factivity debate as we briefly discuss below. Thus, it is important that we no longer conflate model explanation with model-induced explanation.

5 Concluding Remarks

In this paper, we argued that there are two substantially different conceptions of model explanation, which should not be conflated, but often are: model explanation and model-induced explanation. We argued that model explanations are best understood in terms of an identity relation between the explanation's core and the core of the respective model's content (ME_Core). By contrast, model-induced explanations (MIE) only feature explanantia that have been obtained by working with a model. We further argued that paradigmatic cases of alleged ME do not fulfill the criteria for ME; instead they turn out to be MIE. It seems that philosophers of science have taken up an interest, not with model explanation, as they claim, but with model-induced explanation. The special or interesting aspects of these explanations with models is due to the epistemic discovery behind the explanations—that is how models induce, enable, or generate explanations—not the properties of the explanations themselves. Thus, philosophers need to reconsider the unique way that models explain. Our notion of ME_Core suggests thinking through what the central notions of a model are and how they provide explanatory power as a way forward.

Lastly, we also expect that, in light of our results, the argument for the anti-factivity of scientific explanation loses much of its force. The anti-factivity debate is driven by two fundamental assumptions: (i) that explanations with models are ME, and (ii) that at least some explanations with models involve the idealizations stipulated by the model. In this paper, we substantially undermined (i). The survey of the paradigmatic cases of alleged ME shows that they are really cases of MIE. We have also called into question (ii). In none of the cases of idealizations that we discussed are the idealizations themselves part of the respective explanantia. However, we stopped short of offering a decisive argument against the possibility of (ii). Whether there are explanations that include idealizations stipulated by a model demands a closer look (for arguments against (ii), see, e.g., Lawler (2019); Rice (2019b)).

Acknowledgements. Discussions with colleagues and advisors contributed to shaping the view that we defend in this article. We're grateful to (in alphabetical order) Mark Alfano, Finnur Dellsén, Anna-Maria Asunta Eder, Catherine Elgin, Benjamin Feintzeig, Roman Frigg, Kareem Khalifa, Christian Nimtz, Juha Saatsi, Henrik Sova, Thomas Spitzley, Michael Strevens, Raphael van Riel, Kate Vredenburgh, and the participants of Thomas Spitzley's and Christian Nimtz's research groups. We also thank the audiences in Aarhus, Atlanta, Barcelona, Bochum, Bordeaux, Exeter, Ghent, Lund, Pärnu, and Seattle, as well as three anonymous reviewers for their constructive criticisms and suggestions. Insa Lawler gratefully acknowledges that part of her research for this article was funded by the Volkswagen Foundation for the project 'A Study in Explanatory Power', by the German Academic Exchange Service (DAAD) for a research stay at New York University (2015–2016), and by an Emmy Noether Grant from the German Research Council (DFG), Reference No. BR 5210/1-1.

Conflict of interest. On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alexandrova, A. (2008). Making models count. *Philosophy of Science*, *75*(3), 383–404.

Alexandrova, A., & Northcott, R. (2013). It's just a feeling: Why economic models do not explain. *Journal of Economic Methodology*, *20*(3), 262–267.

Bailer-Jones, D. (2003). When scientific models represent. *International Studies in the Philosophy of Science*, *17*(1), 59–74.

Batterman, R. (2002). *The devil in the details: Asymptotic reasoning in explanation, reduction and emergence*. Oxford: Oxford University Press.

Batterman, R. (2009). Idealization and modeling. *Synthese*, *169*, 427–446.

Batterman, R., & Rice, C. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349–376.

Batterman, R. W. (2005). Response to Belot's "Whose Devil? Which Details?". *Philosophy of Science*, *72*(1), 154–163.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, *180*, 33–45.

Bokulich, A. (2012). Distinguishing explanatory from nonexplanatory fictions. *Philosophy of Science*, *79*(5), 33–45.

Bokulich, A. (2017). Models and explanation. In L. Magnani & T. Bertolotti (Eds.), *Handbook of model-based science* (pp. 103–118). Berlin: Springer.

Butterfield, J. (2011). Less is different: Emergence and reduction reconciled. *Foundations of Physics*, *41*(6), 1065–1135.

Craver, C. (2006). When mechanistic models explain. *Synthese*, *153*, 355–376.

Downes, S. (2011). Scientific models. *Philosophy Compass*, *6*(11), 757–764.

Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, *132*(1), 33–42.

Elgin, C. (2017). *True enough*. Cambridge: MIT Press.

Elgin, M., & Sober, E. (2002). Cartwright on explanation and idealization. *Erkenntnis*, *57*(3), 441–450.

Elliott-Graves, A., & Weisberg, M. (2014). Idealization. *Philosophy Compass*, *9*(3), 176–185.

Fang, W. (2019). An inferential account of model explanation. *Philosophia*, *47*(1), 99–116.

Feintzeig, B. (2017). Deduction and definability in infinite statistical systems. *Synthese* (online first).

Frigg, R. (2010). Models and fictions. *Synthese*, *172*(1), 251–268.

Frigg, R. & Hartmann, S. (2012). Models in science. In E. N. Zalta EN (Ed.) *The Stanford encyclopedia of philosophy*, fall 2012 edn.

Frigg, R., & Nguyen, J. (2018). The turn of the valve: Representing with material models. *European Journal for Philosophy of Science*, *8*(2), 205–224.

Gelfert, A. (2016). *How to do science with models*., A philosophical primer Berlin: Springer.

Giere, R. (2004). How models are used to represent reality. *Philosophy of Science*, *71*, 742–752.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies*, *143*, 101–116.

Grüne-Yanoff, T. (2009). Learning from minimal economic models. *Erkenntnis*, *70*(1), 81–99.

Grüne-Yanoff, T. (2013). Genuineness resolved: A reply to Reiss' purported paradox. *Journal of Economic Methodology*, *20*(3), 255–261.

Hempel, C. (1965). *Aspects of scientific explanation, and other essays in the philosophy of science*. New York: Free Press.

Hughes, R. (1990). The Bohr atom, models, and realism. *Philosophical Topics*, *18*, 71–84.

Hughes, R. (1997). Models and representation. *Philosophy of Science*, *64*, 325–336.

Jebeile, J., & Kennedy, A. (2015). Explaining with models: The role of idealizations. *International Studies in the Philosophy of Science*, *29*(4), 383–392.

Kaplan, D. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*, 339–373.

Kennedy, A. (2012). A non representationalist view of model explanation. *Studies in History and Philosophy of Science, Part A*, *43*(2), 326–332.

King, M. (2016). On structural accounts of model-explanations. *Synthese*, *193*, 2761–2778.

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *British Journal for the Philosophy of Science*, *61*(3), 541–567.

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science*, *64*, 485–511.

Lawler, I. (2019). Scientific understanding and felicitous legitimate falsehoods. *Synthese*,. https://doi.org/10.1007/s11229-019-02495-0.

Levins, R. (1966). The strategy of model building in population biology. In E. Sober (Ed.), *Conceptual issues in evolutionary biology* (pp. 18–27). Cambridge: MIT Press.

Mäki, U. (2005). Models are experiments, experiments are models. *Journal of Economic Methodology*, *12*(2), 303–315.

Mäki, U. (2013). On a paradox of truth, or how not to obscure the issue of whether explanatory models can be true. *Journal of Economic Methodology*, *20*(3), 268–279.

Marchionni, C. (2017). What is the problem with model-based explanation in economics? *Disputatio*, *9*(47), 603–630.

Morrison, M. (1999). Models as autonomous agents. In M. Morgan & M. Morrison (Eds.), *Models as mediators: Perspectives on natural and social science* (pp. 38–65). Cambridge: Cambridge University Press.

Norton, J. (2012). Approximations and idealizations: Why the difference matters. *Philosophy of Science*, *79*, 207–232.

Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology and Philosophy*, *26*(5), 757–71.

Orzack, S., & Sober, E. (1993). A critical assessment of Levins's the strategy of model building in population biology' (1966). *Quarterly Review of Biology*, *68*(4), 533–546.

Parker, W. (2006). Understanding pluralism in climate modeling. *Foundations of Science*, *11*(4), 349–368.

Pincock, C. (2007). A role for mathematics in the physical sciences. *Noûs*, *41*, 253–275.

Pincock, C. (2011). Mathematical explanations of the rainbow. *Studies in History and Philosophy of Science Part B*, *42*(1), 13–22.

Pincock, C. (2012). *Mathematics and scientific representation*. Oxford: Oxford University Press.

Pincock, C. (2014). How to avoid inconsistent idealizations. *Synthese*, *191*, 2957–2972.

Potochnik, A. (2007). Optimality modeling and explanatory generality. *Philosophy of Science*, *74*(5), 680–6915.

Potochnik, A. (2009). Optimality modeling in a suboptimal world. *Biology and Philosophy*, *24*(2), 183–197.

Potochnik, A. (2010). Explanatory independence and epistemic interdependence: A case study of the optimality approach. *The British Journal for the Philosophy of Science*, *61*(1), 213–233.

Reiss, J. (2012). The explanation paradox. *Journal of Economic Methodology*, *19*(1), 43–62.

Reiss, J. (2013). The explanation paradox redux. *Journal of Economic Methodology*, *20*(3), 280–292.

Rice, C. (2012). Optimality explanations: A plea for an alternative approach. *Biology & Philosophy*, *27*, 685–703.

Rice, C. (2016). Factive scientific understanding without accurate representation. *Biology & Philosophy*, *31*(1), 81–102.

Rice, C. (2018). Idealized models, holistic distortions, and universality. *Synthese*, *195*(6), 2795–2819.

Rice, C. (2019a). Models don't decompose that way: A holistic view of idealized models. *The British Journal for the Philosophy of Science*, *70*(1), 179–208.

Rice, C. (2019b). Understanding realism. *Synthese*. https://doi.org/10.1007/s11229-019-02331-5.

Rohwer, Y., & Rice, C. (2013). Hypothetical pattern idealization and explanatory models. *Philosophy of Science*, *80*(3), 334–355.

Rohwer, Y., & Rice, C. (2016). How are models and explanations related? *Erkenntnis*, *81*(5), 1127–1148.

Saatsi, J. (forthcoming). Realism and explanatory perspectivism. In M. Massimi, & C. McCoy C (Eds.) *Understanding perspectivism: Scientific challenges and methodological prospects*. Routledge.

Schmid-Hempel, P., Kacelnik, A., & Houston, A. (1985). Honeybees maximize efficiency by not filling their crop. *Beahavioral Ecology and Sociobiology*, *17*, 61–66.

Shech, E. (2013). What Is the paradox of phase transitions? *Philosophy of Science*, *80*(5), 1170–1181.

Shech, E. (2017). Idealizations, essential self-adjointness, and minimal model explanation in the Aharonov–Bohm effect. *Synthese* (online first).

Strevens, M. (2008). *Depth. An account of scientific explanation*. Cambridge: Harvard University Press.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science*, *44*(3), 510–515.

Strevens, M., et al. (2017). How idealizations provide understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 37–48). Abingdon: Routledge.

Suárez, M. (2010). Scientific representation. *Philosophy Compass*, *5*(1), 91–101.

Sullivan, E. (2019). Universality caused: The case of renormalization group explanation. *European Journal for Philosophy of Science*, *9*(3), 36. https://doi.org/10.1007/s13194-019-0260-x.

Sullivan, E., Khalifa, K. (2019). Idealizations and understanding: Much ado about nothing? *Australasian Journal of Philosophy*, *97*(4), 673–689.

Toon, A. (2012). *Models as make-believe: Imagination, fiction, and scientific representation*. London: Palgrave-Macmillan.

van Riel, R. (2015). The content of model-based information. *Synthese*, *192*(12), 3839–3858.

van Riel, R. (2017). What is the problem of explanation and modeling? *Acta Analytica*, *32*(3), 263–275.

Wayne, A. (2011). Expanding the scope of explanatory idealization. *Philosophy of Science*, *78*(5), 830–841.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730–742.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, *104*(12), 639–659.

Werndl, C., & Steele, K. (2016). The diversity of model tuning practices in climate science. *Philosophy of Science*, *83*(5), 1133–1144.

Wimsatt, W. (1981). Robustness, reliability, and overdeterminationy. In M. Brewer & Collins (Eds.), *Scientific inquiry and the social science* (pp. 124–163). San Francisco: Jossey-Bass.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*(2), 219–240.