A defense of truth as a necessary condition on scientific explanation[*]

Christopher Pincock – December 26, 2020 – 10373 words

Abstract: How can a reflective scientist put forward an explanation using a model when they are aware that many of the assumptions used to specify that model are false? This paper addresses this challenge by making two substantial assumptions about explanatory practice. First, many of the propositions deployed in the course of explaining have a non-representational function. In particular, a proposition that a scientist uses and also believes to be false, i.e. an "idealization", typically has some non-representational function in the practice, such as the interpretation of some model or the specification of the target of the explanation. Second, when an agent puts forward an explanation using a model, they usually aim to remain agnostic about various features of the phenomenon being explained. In this sense, explanations are intended to be autonomous from many of the more fundamental features of such systems. I support these two assumptions by showing how they allow one to address a number of recent concerns raised by Bokulich, Potochnik and Rice. In addition, these assumptions lead to a defense of the view that explanations are wholly true that improves on the accounts developed by Craver, Mäki and Strevens.

I. Introduction

Nearly all scientific explanations are presented using scientific models. These models involve claims that the scientist believes are false of their target phenomenon. I call these claims "idealizations". When a reflective scientist puts forward an explanation, they confront a challenge. Suppose they aspire to present an explanation that is wholly true. This may be

[*] Forthcoming, *Erkenntnis*.

because they assume that truth is a necessary condition on explanation. I call this view "veritism". The challenge is that much of what the scientist says in giving that explanation is believed to be false. The main conclusion of this paper is that the best way for the reflective scientist to preserve an initial commitment to veritism is to qualify the role of the false claims in their explanation and to detach their explanation from the fundamental features of the physical world. The first element of this defense of veritism can be called "anti-representationalism", while the second element of this defense maintains the metaphysical autonomy of scientific explanation. An anti-representationalist rejects the representationalist dogma that the only function of a proposition put forward in a scientific context is to represent some fact (Price 2013). A reflective scientist that puts forward propositions that they believe to be false cannot coherently take the function of these speech acts to be to represent some fact. Some other genuine function must be identified. In sections II and III I consider some of the functions that false propositions can perform in the presentation of model-based scientific explanations that have been recently emphasized by Bokulich and Potochnik. Many of these roles are perfectly consistent with veritism, and so a good way to defend veritism is to acknowledge just these non-representational functions.

Many philosophers of science have offered defenses of veritism that emphasize the non-representational functions of propositions that are believed to be false of the target of explanation. In section IV I consider the proposals of Craver, Mäki, Rice and Strevens. These defenses of veritism differ on what non-representational functions they allow for, and also on how genuine explanations are tied to the fundamental features of the physical world. I argue in section IV that the best way for the reflective scientist to maintain veritism is to follow

Woodward and endorse the metaphysical autonomy of scientific explanation. At the same time, much of what I say builds on Strevens' account of the value of idealization in presenting a scientific explanation. So one novel contribution in section IV is to show how Strevens' basic approach to idealization can be detached from the links to fundamental reality that Strevens imposes. The challenge is to provide a coherent account of what makes an explanation genuine without saddling the scientist with excessive metaphysical commitments. Here I draw on Yablo's notion of partial truth.

II. Models and Propositions

Why do the tides occur twice a day at various locations on the earth, such as San Francisco, Fiji and Cape Town? The target of this explanatory question is a true proposition: that the tides occur twice a day at these locations. The veritist supposes that an explanation of such a target just is one or more true propositions. For the explanation to be a genuine explanation, these propositions must be explanatorily relevant to the target. One species of explanatory relevance is causal relevance, and in this paper I will usually take for granted that Woodward's interventionist account of causal explanation is the best way to approach causal, explanatory relevance (Woodward 2003a). For Woodward, there are two kinds of propositions that are causally relevant to some target. First, there are causal generalizations that correctly tie some causal variables to the variable features of the explanatory target. Second, there are propositions that truly characterize the actual values of some of these causal variables. In the tides case the formulation of these propositions is somewhat complicated, but in more ordinary cases we have no difficulty arriving at these two sorts of propositions. For example, why do certain flamingoes have pink feathers? First, consuming canthaxanthin (a natural pink dye) causes pink feathers (causal generalization). Second, these flamingos consume canthaxanthin (actual values of some causal variables).

An explanation of the twice-daily ("diurnal") character of some tides seems to work quite differently.[1] Figure 1 presents the critical step in Newton's celebrated explanation of the tides (Newton 1999, Bk. 3, Prop. 24, 837).

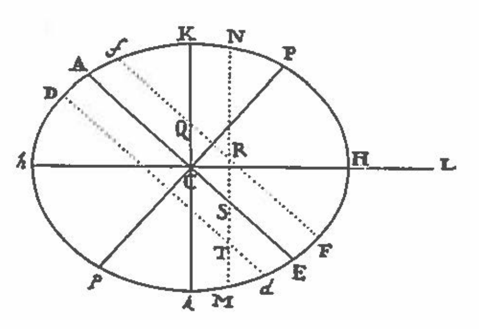

Figure 1: Newton's model of the tides (Newton 1999, 837)

Consider a spherical earth with equator ACE that is completely covered with water, with the moon at some distance along the line CHL. The gravitational forces exerted by the moon on the surface of the earth give rise to a distribution of so-called "tidal forces" that will raise the waters of the Earth on the side closest to the moon (at H) and on the side farthest from the moon (at h). Newton supposes that the water will immediately change its height once the tidal forces are applied or changed. The upshot of his discussion is the "spheroid" corresponding to these altered heights of the waters of the earth in response to the tidal forces: "this spheroid [HKhA] will represent the figure of the sea very nearly" (Newton 1999, 837). The rotation of the earth on its own axis PCp then leads to a given location experiencing two high tides per day. With this argument, it looks like Newton is explaining why the tides occur twice a day (at certain locations) using two claims. First, the moon's gravitational forces bring about this distribution of tidal forces, which in turn immediately alters the height of the water. Second, the earth completes one rotation around its own axis each day.

Newton's explanation of the tides will serve as our primary case study of an explanation that raises challenges for veritism. In this section I consider the problem posed just by the use of a scientific model. The explanation does not seem to be made up of propositions whose subject matter is the target of the explanation. Instead, it looks as if Newton has interposed a distinct model earth whose features are very different from the actual earth: the model earth is spherical and completely covered with water, while the actual earth is not a sphere, and has land in addition to oceans and seas. If the explanation deploys this model, then the explanation is not even truth-apt as a model is not the right sort of thing to be true or false. And if we look around for truth-apt propositions, the only propositions involved here seem to be about the model, and not about the target.

In a number of papers Bokulich has argued for a distinct kind of scientific explanation that she calls model explanation: "what makes something a *model* explanation is that the explanans in question makes essential reference to a scientific model, and that scientific model (as I believe is the case with all models) involves a certain degree of idealization and/or fictionalization" (2011, 38).[2] Whatever a model is, it is not a proposition or a collection of propositions. So, model explanations are a direct challenge to veritism. Here we have explanations that have models as their parts, and so there is no way to maintain that every genuine explanation is wholly true. As Bokulich summarizes the concern, "This use of models to explain is at odds with traditional philosophical accounts of explanation insofar as models are not literally true descriptions of their target systems; rather, they involve all sorts of falsehoods, including idealizations, abstractions and outright fictions" (2012, 726).

Bokulich has also offered an account of what makes a model explanation a genuine explanation: "the model explains the explanandum by showing how elements of the model correctly capture the pattern of counterfactual dependence of the target system" (2011, 39). A proposed explanation is genuine, then, when it correctly encapsulates a salient pattern of counterfactual dependence. And such an explanation is a model explanation when it captures this pattern in a way that requires the use of a model. Bokulich often presents her account as a generalization of what Woodward has offered. While Woodward was right to emphasize counterfactual dependence, he erred in supposing that the only way to present such dependencies was through propositions that are true of the target system. In addition, it is often possible, and perhaps even required, for scientists to resort to models to capture such dependencies.[3]

To illustrate Bokulich's position, I will summarize what her account entails for the tides case. The model here is some kind of imagined, fictional scenario where the earth is covered with water and the height of the water is immediately shifted by the impressed tidal forces. A counterfactual dependence is captured by this fictional scenario because it is fully realized there. For example, in fact and in the fictional model scenario, Cape Town's location experiences two high tides every twenty-four hours as the earth rotates on its own axis once every twenty-four hours. In the fictional model scenario, were the earth to rotate much more slowly, then that location would not experience two high tides every twenty-four hours, but only one high tide. So, in the fictional model scenario, the number of high tides in a twenty-four hour period counterfactually depends on the rate of rotation of the earth on its own axis. And this dependence, which is realized by the fictional scenario, is also found in the actual earth-moon system. The model correctly captures a genuine counterfactual dependence by realizing it, and the model thereby explains this aspect of the tides on the real earth.

What is the best veritist response to the challenge from model explanations? In line with the general strategy announced in section I, I maintain that the best option is to identify the non-representational functions of the propositions used in this explanatory context. The veritist should agree with Bokulich that many of the propositions used here function to specify and interpret a scientific model. These propositions are not intended to represent features of the actual earth. If this is right, then the falsity of these propositions, when taken to represent the actual earth, is not a problem for the veritist. For the veritist maintains only that a genuine explanation is made up of true propositions that are explanatorily relevant to the target. And propositions that function to specify and interpret a model are not designated as part of the explanation, as the veritist conceives it.

To flesh out this response, it is important to spell out the two-stage picture of model explanation that is crucial to this veritist line of defense. In the first stage, an agent will use propositions to specify and interpret a scientific model. In the tides case, the specification of the model involves indicating which mathematical structure is intended.[4] For Newton, we have a sphere with a spheroid superimposed and various lines added, as shown in figure 1. In addition, this mathematical structure must be interpreted. The interpretation will involve additional propositions that, among other things, must indicate what elements of the model stand for what real-world entities. For example, one line of the model stands for the earth's axis, while another line is made to stand for the earth's equator. Different accounts of how models represent target systems will flesh out how this interpretation works in somewhat different ways. For example, some accounts will emphasize a proposed structural relation that an agent puts forward as obtaining between a model and target, while others will emphasize a translation key that takes features of the model and outputs purported features of the target. The veritist need not endorse any specific account of how models represent. The important point for the veritist is just that many propositions presented by agents serve to interpret the model. Their function is not to represent the target.

Once the first stage of modeling is complete, one has a model that is specified and interpreted. This means that the model will *generate* propositions that are about the target. For example, Newton's model of the earth generates the proposition that the earth rotates on its own axis once every twenty-four hours. The veritist should identify a model explanation with some of the propositions generated by a model about the target. As long as all of these propositions are true, then veritism can be maintained. On this view, the point of the model is to generate propositions, some of which are true and explanatorily relevant to the target. An explanatory model, then, functions to provide a traditional, propositional explanation.

One can defend veritism by arguing that an explanatory model enables a traditional, propositional explanation that is wholly true. But there is more than one way to implement this defense. An instrumentalist defense of veritism maintains that any model that generates a propositional explanation counts as explanatory, no matter how this is accomplished. This proposal seems to ignore what is special about how some models do this. In section IV I will argue that we should consider how the model generates these propositions. Only some models should count as explanatory because only some models will generate these propositions in the right way. However, before describing what I take this right way to be, I want to get clearer on the notion of explanatory relevance.

III. Idealizations and Explanatory Relevance

We have regimented the term "idealization" by saying that an idealization is a proposition that an agent deploys when considering some target, but that the agent believes to be false of the target. In the last section I argued that the veritist should not be troubled by idealizations that function to specify and interpret a scientific model, even when that model is used for explanatory purposes. The obvious anti-veritist response is that many idealizations will be generated by the interpreted model. If it turns out that any of the idealizations generated by the model are part of a genuine explanation of the target, then veritism is again in trouble. Many anti-veritists appeal to scientific practice to support this objection. Potochnik is the person who has put the point most forcefully for she has identified several positive explanatory contributions for model-generated idealizations.[5] If these contributions are genuine, then veritism is refuted.

I claim that the best veritist response to this kind of objection is to carefully consider what these contributions are. Once they are considered, it becomes clear that the function of these idealizations in explanatory practice is non-representational. Here is how Potochnik summarizes her position: "Idealizations, including even the introduction of fictional entities, can contribute to causal pattern explanation when they help represent real causal patterns by representing phenomena as if they had features they do not" (Potochnik 2017, 144). The veritist will maintain that any false proposition that helps to represent explanatorily relevant factors will not do so by “representing phenomena as if they had features they do not”, but instead through some non-representational function. Potochnik proposes three positive contributions for these idealizations. I argue that the first two contributions are genuine, but also perfectly consistent with veritism. The third contribution would undermine veritism, but I argue that it is not a legitimate requirement on explanation.

To appreciate Potochnik’s account of explanation, it is important to emphasize her commitment to explanatory pluralism. Potochnik supposes that all explanations are causal pattern explanations in a way that is supposed to fit with Woodward’s approach to causal explanation. So in this sense she is not an explanatory pluralist: all explanations exploit a single explanatory relevance relation of causal relevance. However, Potochnik does emphasize how different research programs approach a single explanatory target in incompatible ways. For example, one of her cases considers the color variation among a flock of Harris sparrows. One explanation of this variation is offered by game theorists via models that exploit frequency dependent selection. Another explanation of this same variation is presented by developmental biologists using models of gene-environment interactions. These different types of causal explanation cannot be easily combined or subsumed into some third type of causal explanation. One reason for this is that the two types of explanations are said to employ incompatible explanatory assumptions: “These different causal pattern explanations represent some of the same causal influences in different ways” (Potochnik 2017, 150). In this sense, then, Potochnik endorses a kind of causal explanatory pluralism, where different types of causal explanations are required to make sense of explanatory practice.

This kind of explanatory pluralism is naturally associated with Woodward's requirement that a set of variables be chosen prior to any causal explanatory claims being advanced. For the game theorists, "Coloration is a badge that helps the birds divide resources while avoiding unnecessary injury" (Potochnik 2017, 150). So, one causal variable connects coloration to behaviors that are in turn tied to the division of scarce resources like food or mating opportunities. The game theorists advance their causal explanation of the color variation with respect to this sort of variable set. By contrast, the developmental biologists select a different variable set. It will include factors such as how genes are distributed among the individuals in the population and how environmental factors influence gene expression. By starting with this variable set, the developmental biologists can arrive at different causal explanations of the very same color variation.

For each type of explanation to be presented, an agent must deploy propositions that (i) identify the target of the explanation and (ii) indicate what kind of explanation is being offered. It is clear how these propositions are essential to the practice of presenting explanations. At the same time, the veritist should also emphasize how neither sort of proposition is explanatorily relevant to the target of the explanation. To see why, recall the restricted notion of explanatory relevance that Woodward deploys: a proposition is explanatorily relevant to some target just in case (a) the proposition is a causal generalization tied to that target or (b) the proposition characterizes the actual values of some of these causal variables. If an agent says that the aim is to explain the color variation in a flock of Harris sparrows, this proposition does indicate the intended target of the explanation, but it is not explanatorily relevant to the target as it is not a proposition of type (a) or (b). Similarly, when an agent says that they are offering an explanation tied to developmental biology, then they are not advancing a proposition of type (a) or (b). So, no such proposition functions to represent an explanatorily relevant feature of the situation. For the veritist, then, no proposition that functions in way (i) or (ii) is part of the genuine explanation of the target.

Suppose, though, that an agent advances a proposition with the function of indicating what type of explanation is being offered, but where that proposition is an idealization. As Potochnik emphasizes, sometimes a game theorist studying Harris sparrows will indicate their preferred type of explanation by advancing a proposition that falsely discounts the causal significance of some other kind of causal factor. For example, a game theorist may say simply "that feather color is heritable" (Potochnik 2017, 151). This is a substantial and overly simplified claim about the causal variables that are the focus of developmental biologists. So it might seem, as Potochnik puts it, "The game theory explanation represents what is possibly environmentally mediated gene expression with the simple assumption that feather color is heritable" (Potochnik 2017, 151). This makes it seem as if the false proposition about heritability is part of the game theorists' explanation, i.e. that it functions as an explanatorily relevant, though false, proposition. Of course, the veritist must deny this reading of what is going on with such propositions. The proposal that I am defending is that the veritist should say that some false propositions about the causes of one type function only to enable wholly true explanations concerning causes of some other type. In this case, the function of the false proposition is to indicate what type of causal explanation is being offered. The point of distorting the heritability of the color variation is to enable a focus on the frequency dependent selection for that trait. Schematically, a false value for a variable from the variable set emphasized by an alternative research program can help to indicate that one’s research program is concerned with the true values of a variable from some other variable set.

Potochnik also identifies a third function for idealizations in explanation. This is tied to Potochnik’s account of how genuine scientific explanations afford understanding. According to Potochnik, “Idealizations facilitate understanding, but not truth, when they enable the representation of cognitively valuable connections and patterns that more accurate portrayals would miss” (2017, 97). In particular, for an explanation to afford understanding, the scope of the causal pattern must be appreciated by the agent who grasps the explanation. So, for Potochnik, “the explanatory dependence relations – causal patterns – are not simply causal dependencies but causal dependencies embodied by some range of phenomena” (2017, 140). As the scope of the pattern is an intrinsic feature of the pattern, the scope must be represented by some element of a genuine explanation. But scientists typically have no way to delimit the scope of such a causal pattern without some recourse to idealization. For example, a scientist may say that, assuming that feather color is heritable, some kind of frequency dependent selection is embodied in the Harris sparrow population. The proviso here is essential to the explanation but the scientist is aware that it is overly simplified and a distortion of some other causally relevant pattern.

When it comes to this function for idealizations in explanation the veritist should follow Woodward and distinguish the causal generalization from its scope. The causal explanation is just the causal generalization and a characterization of the actual values of some of the causal variables. Every causal generalization *has* a scope in the various senses of “scope” identified by Woodward (Woodward 2006). For example, a causal generalization will be invariant to some degree when it continues to hold true for a range of values of its causal variables. And such a causal generalization will also be insensitive to some degree when it continues to hold, or fails, for a range of values of other causal variables. A good explanation will deploy a causal generalization that has a wide scope in either the sense of invariance or insensitivity. But a characterization of this scope is not some further element of the explanation. Strictly speaking, the scope of the generalization is not explanatorily relevant to the target: a claim concerning the scope is not an element of the explanation. On this notion of explanatory relevance, it is enough for the causal generalization to hold of the target. When a causal generalization holds of some target, the explanation will provide answers to "what if things had been different?" questions. The range of such questions that can be answered will of course depend on the scope of the causal generalization.

Potochnik may reply that this veritist response gives up the core assumption that grasping a genuine explanation affords scientific understanding. Here the veritist has two options to pursue. First, they could develop an account of scientific understanding that retains this link to explanation and then argue that understanding does not require appreciating the scope of the explanatory causal pattern (Strevens 2017). Second, they could give up the assumption that grasping a scientific explanation is sufficient for understanding. On this alternative, scientific understanding is a cognitive achievement that requires more than grasping the wholly true scientific explanation. The elements that Potochnik argues are required for understanding may thus be required for understanding without impacting the character of explanation. On either option, veritism about explanation is preserved.

IV. Commitments to Underlying Truths

IV.1. Instrumentalism

Scientists tend to value the explanations that they obtain, and they also distinguish between models that afford such explanations and models that have some other function such as prediction or exploration. Following Rohwer and Rice, I will call a model “explanatory” when it generates a genuine explanation for some target in the right way (Rohwer & Rice 2013, 346).[6] However, even among veritists, there does not seem to be any consensus on what this “right way” amounts to. Until this is clarified, the reflective scientist who wishes to remain a veritist about explanation should be uneasy.

One clear solution to these difficulties would be take an instrumentalist attitude towards explanatory models: their role is simply to generate a genuine explanation, and if they accomplish this, then there is nothing more to ask. This is one way to interpret Craver’s “ontic” approach to explanation. In the primary sense, “the term [‘]explanation[’] refers to an objective portion of the causal structure of the world, to the set of factors that produce, underlie, or are otherwise responsible for a phenomenon” (Craver 2014, 40). Such explanations “are not true or false” (Craver 2014, 40), but when we present models, the models may represent systems in ways that are true or false. Such models can be explanatory when “they bring to light aspects of the system under investigation that are difficult to see unless one makes false assumptions ... one reveals aspects of the ontic structure of the system that would otherwise be occluded” (Craver 2014, 49-50). In the tides case, it is difficult for us to grasp the causal generalization at issue. We are free to develop a model that dramatically distorts the causally relevant ontic structures as long as the model enables us to grasp some true causal generalizations that would have otherwise eluded us.

Instrumentalism about explanatory models gives up any epistemic role for the model in relation to the explanation. One of the strengths of Newton's model is *how* it enables us to derive this causal generalization, not just *that* it enables that derivation. Craver can only appeal to the model as a kind of cognitive aid that allows us to grasp some true proposition that would have otherwise eluded us. An instrumentalist can evaluate the efficiency of an instrument in producing a desired outcome. But if we follow the instrumentalist strategy we have no properly epistemic dimension of evaluation that we could deploy. Newton's model is better than some of its competitors because it allows us to derive this causal generalization in a superior way. A non-instrumentalist version of veritism, then, must identify this aspect of these explanatory models.

The problem, then, is that the most natural defense of veritism is to consign models to a merely enabling role, but this gives up the epistemic distinctions that scientists deploy when they rank models as more or less explanatory. The obvious response to this problem is to focus on what the propositions generated by the model say about the target system. If these propositions somehow characterize why some phenomenon has the character that it does, despite their falsity, then we can see why scientists would value the model for its own sake. However, it turns out to be a delicate matter to pin down how the propositions generated by the model should relate to the explanatory target for the model to count as explanatory.

The most well-known non-instrumentalist characterizations of explanatory models seem too restrictive. Consider, for example, the influential proposal by McMullin. McMullin argues that "The implications of construct idealization, both formal and material, are … truth-bearing in a very strong sense" (McMullin 1985, 264). McMullin's construct idealizations involve representations, such as scientific models, that are deliberately simplified. When the simplification alters features that are believed to be explanatorily relevant, we have what McMullin calls formal idealization (McMullin 1985, 258). Material idealizations, by contrast, arise when the simplification turns on the omission of some features of the target system (McMullin 1985, 262). The idealized models that I have focused on thus relate only to McMullin's formal idealizations. He imposes a very demanding test for when formally idealized models can explain: the model is explanatory just in case there are "processes of self-correction and imaginative extension" that "are suggested by the model itself" (McMullin 1985, 264).[7] I agree with McMullin that when this condition is met, the model may be explanatory. However, a model can explain and not meet this demanding condition. In Newton's case, the model is explanatory and yet the model itself does not indicate how to arrive at a wholly true derivation of the explanatory generalization. More generally, a model can explain even though it is idealized and the scientists using the model do not know how to remove that idealization from the model.

By contrast, another influential approach, developed by Mäki, is too flexible. It is worth emphasizing that Mäki has long been a champion of the view that idealizations have a wide variety of non-representational functions, and that puzzles about idealization often arise because philosophers ignore these various functions. One of his examples considers a derivation arising from an economic model that exploits the idealization (C) that the economy is closed. Mäki says that his "strategy is to refrain from considering [C] and other such assumptions as factual assertions but rather to turn them into other sentences that are used to make factual assertions about the world or to make claims about them and their role in inquiry"

(Mäki 2012, 221). One such paraphrase would be to take the use of (C) to assert that such a factor is explanatorily irrelevant to the target in question.

Mäki also emphasizes that an agent may not aim to paraphrase an idealization. Instead, the agent may make "meta-claims about the original idealization" (Mäki 2012, 226). One such meta-claim involves a commitment to replace that falsehood in future theorizing: "the orientation is ... a matter of attitudes such as intention, prescription, hope, or even promise: it is intended, prescribed, hoped, or promised that the assumption in question indeed will be just an early-step assumption and will be relaxed at a later step" (Mäki 2012, 227). On this reading, the non-representational function of such false propositions is to express a commitment that the false proposition will be replaced at some point in the future.

Unlike McMullin, Mäki would have no difficulty counting Newton's model of the tides as explanatory. All that is needed is the plausible assumption that Newton and his followers committed themselves to various meta-claims about how that model could be improved in the future. However, it seems that Mäki has not offered any proposals for what sorts of improvements are consistent with counting the original model as explanatory. Some changes to a model would produce a completely different account of the phenomena that had nothing to do with the original proposal. In such cases, the adequacy of the revised account is not a sufficient reason to count the original model as explanatory.

The upshot of our brief examination of McMullin and Mäki is that an idealized model should count as explanatory when it gets at the truth about the phenomenon being explained, but it need not do so in a way that easily allows users of the model to formulate this truth. The same lesson can be extracted from Rice's more recent discussion of these issues.

Rice argues that idealizations "are often essential to the explanations provided by scientific models because they allow for the application of various (mathematical) modelling techniques" (Rice 2019, 196).[8] The only way for the model to explain is for it to generate some propositions, but this outcome is only obtainable using some mathematical techniques. However, to use these techniques, some idealizations must be exploited. Crucially, the veritist cannot reply that these idealizations "only distort irrelevant or unimportant features and do not 'get in the way' of the accurate parts of scientific models that do the real work" (Rice 2019, 198). So, we have no way to identify what is special about explanatory models that is tied to the veritist assumption that genuine explanations are wholly true.

We can illustrate Rice's worries using Newton's model of the tides. The model-based explanation that we have focused on relies on a causal generalization that connects the rate of rotation of the earth on its own axis to the rate with which high tides occur in places like Cape Town. However, it seems important to consider how this generalization is arrived at using propositions generated by the model. A crucial step is the proposition that the waters of the earth instantly respond to the impressed tidal forces. This is a false proposition, when taken to be about the actual waters of the earth, as any such body of water will take some time to respond due to inertia. Still, using this proposition is the only way for Newton to mathematically derive a proposition that characterizes the "spheroid" shape of the waters once the tidal forces are considered. We have identified a proposition, then, that is both essential to the model-based derivation of our causal generalization, and also false of a causally relevant feature of the process being explained. How quickly the waters respond is clearly causally relevant to how often the tides occur in these places.

Rice's proposed strategy for addressing this issue is tied to universality classes. A universality class is a class of actual and possible systems that realize some property of interest. In our case, we might have a universality class whose members each realize the property that the tides occur twice at many locations in a twenty-four period. One member of this class is the actual earth, while a wide variety of other physically possible planets with tides are included in the class as well. One way to link the model that involves water with a response time of 0 seconds to the actual earth is to note that many of the systems in the universality class will have a non-zero but constrained response time. This range of variation within the universality class suggests that the specific response time found on the actual earth is not causally relevant to the target property. As Rice summarizes his proposal, "the 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" (Rice 2018, 2816). Explanatory models are distinguished, then, by their relationship to the members of an appropriately delimited universality class.

The notion of a universality class is a useful one, but it is not clear how it can be deployed to address the reflective puzzles of a working scientist like Newton. In particular, Rice allows a model to be explanatory when it generates a causal generalization that is shared with other members of some universality class, even when the basis for the causal generalization in the model is different from the basis in the target system: "many of the patterns of counterfactual dependence that hold in the pervasively distorted model system will be similar to those of its real-world target system(s) – those counterfactual relationships will just hold for (perhaps very) different reasons in the model and perhaps only in limiting cases" (Rice 2018, 2808). This flexibility risks a kind of instrumentalism about explanatory models. If we are free to place the model in any universality class that includes the target, then the only test for a model being explanatory is that it generate the right causal generalization. But if the generalization holds for different reasons across these systems, then it is not clear how looking at one system can explain what is occurring in another system. The basis for the generalization must somehow be restricted if we are to avoid instrumentalism about explanatory models. Again, the lesson is that we must connect what the model says is going on in the target system with what is actually responsible for the explanatory generalization that is generated by the model.[9]

IV.2. Cohesion

The epistemic value of an idealized model, when that model is used to generate an explanation, should be tied to how that model generates the propositions that make up the explanation. In the simplest cases, the model generates a true causal generalization that in fact applies to the target of the explanation. But if we consider how that generation works, we can distinguish between cases where the derivation reflects what is really going on with the target and cases where the derivation is wrongheaded. Our focus must be, then, on the subject matter of the derivation of this causal generalization and how this subject matter relates to what is really going on in the target.

Strevens offers what is perhaps the most comprehensive recipe for addressing these issues. He begins by identifying a positive non-representational function for idealizations when they appear as part of a purported explanation: “the role of an idealization is to assert the explanatory irrelevance, that is, the failure to make a difference, of a salient causal factor” (Strevens 2008, 318). Strevens’ main example involves various scientific models that generate

Boyle's Law PV = k. This law says that, for some amount of gas at some temperature, pressure times volume is a constant. One model whose propositions allow one to derive the law does so using the proposition that "the molecules do not collide with one another" (Strevens 2008, 308). This proposition is false, when taken to be about the gases at issue. For Strevens, the point of presenting this proposition is not to truly represent how often the molecules collide with one another. The function of deploying that proposition is to assert some *other* proposition, namely that the actual rate of molecular collision is explanatorily irrelevant to the target of the explanation. This other proposition is true. So, in an extended sense, the veritist position is vindicated: no false proposition in a genuine explanation functions to represent some explanatorily relevant features of the target. False propositions serve only to direct one to some true and explanatorily important proposition.[10] Unlike McMullin, the explanatory model need not indicate how to arrive at the relevant, true proposition. And, as we will see, unlike Mäki, a restrictive notion of subject matter ties the derivation to what is really going on in the target system.

Is something like Strevens' approach the best way to identify explanatory models while preserving veritism? To answer this question, we must examine in more detail how Strevens can argue that idealizations function to make claims about explanatorily irrelevant features. Strevens' account of explanatory relevance relies on two assumptions. First, there is some fundamental relation of causation that we aim to respect in our explanations. Second, what is explanatorily relevant to some target proposition p is determined by abstracting away from the specifics of this fundamental relation as much as possible while preserving the type of causation that actually made p obtain. For example, in the case of some window that is broken by a rock that is thrown through the window, there is some threshold t for the momentum of the rock that is sufficient for that throw to break the window. The explanatorily relevant feature of the rock's trajectory, then, is just that it meet the window with a momentum that is greater than t. A sound wave with sufficient intensity would also break that window, so an even more abstract characterization of what occurred is that some rock with a momentum greater than t or some sound wave with intensity greater than i impacted the window. But this latter characterization of what occurred is too abstract to be explanatorily relevant as this disjunctive proposition allows for two disjoint types of causal processes. Strevens insists, then, that a proposition e is explanatorily relevant to the target p just in case e's truth involves the most abstract features of the fundamental causal process that are "cohesive" with what actually occurred, where cohesion restricts how e can be true to one type of causal process.

In the case of Boyle's law, the explanatory derivation that we have considered includes the proposition that the molecules do not collide. For Strevens what is explanatorily relevant to Boyle's law is that the rate of collision be below some threshold. Similarly, in the tides case, the explanatory derivation tied to Newton's model includes the proposition that the waters respond instantaneously. For Strevens what is explanatory relevant to this target is that the response time is below some threshold. For both cases, then, presenting a false proposition is a way of indicating that the actual quantity is below some threshold. And as that quantity is in fact below that threshold, the actual value of this quantity is explanatorily irrelevant to the target proposition. So, in this roundabout way, presenting a false proposition improves the explanation by indicating that some quantity is not explanatorily relevant. This is an improvement because some person involved may have mistakenly assumed that the actual value of this quantity was explanatorily relevant. In working through the point of this idealization's being included in the explanation, such a person's illusions are dispelled.

This short summary illustrates three contentious aspects of Strevens' approach that I believe the veritist should avoid. First, Strevens requires a fundamental relation of causation, while the veritist should be neutral on the existence of such a relation. Second, there is only one notion of explanatory relevance for Strevens that is fixed by his abstraction with cohesion algorithm. By contrast, I maintain that the veritist should allow for more than one explanatory relevance relation. These relations may all be species of causal explanatory relevance, as with Potochnik and Woodward, or they may include some non-causal explanatory relevance relations.

A third contentious element of Strevens' proposal brings in issues tied to how a scientist like Newton can endorse their model and also concede that the future of science may lead the model change in fairly dramatic ways. Strevens has a very rigid account of the relationship between the idealization that appears in the explanatory derivation and the truth that is gestured at through the use of the idealization. The truth must be expressed in the very same scientific vocabulary that was used to present the idealization. As we have just seen, for a scientist to work out that some idealization is functioning to indicate that some specific quantity is explanatorily irrelevant, they must reason using the vocabulary at their disposal. However, I argue below that this condition proves too demanding to accommodate cases from the history of science that involve changes in scientific vocabulary. The basic question, then, is can we preserve something like Strevens' basic approach to idealization without reliance on fundamental causation and cohesion?[11]

Here is an alternative characterization of the non-representational function of such false propositions that avoids these three features of Strevens' approach. A model is explanatory just in case (i) the model generates an explanatory generalization and (ii) each idealization in the derivation is partially true so that (iii) there is a wholly true derivation of that explanatory generalization that goes via these underlying truths. I follow Yablo, and say that p is *partially true* when for some r, p entails r, r is true and the subject matter of r is part of the subject matter of p.[12] I will say that such an r is a truth *underlying* p. For example, it is false to say that the response time of the water is 0 seconds. But this proposition is partially true because (i) that the response time of the water is 0 seconds entails that the response time of the water is less than some threshold of t seconds, (ii) it is true that the response time of the water is less than this threshold of t seconds and (iii) the latter proposition's subject matter is part of the subject matter of the former proposition. That is, no new elements of the world are relevant to how the truth of the latter proposition is fixed. So, that the response time of the waters is less than this threshold of t seconds is a truth that underlies the false proposition that the response time of the waters is 0 seconds.[13]

If we adopt this account of model explanation, then we get a corresponding proposal for what an agent is doing when they put forward a model as explanatory. An agent who presents an idealized model as explanatory *commits* themselves to (i) the truth of the explanatory generalization that is generated by the model as well as (ii) the existence of a wholly true derivation that goes through only the truths underlying the partial truths of the model-based derivation. So, the non-representational function of the idealizations that appear in the initial derivation is to signal the commitment of the agent to the partial truth of these false propositions, along with the existence of an explanatory derivation that includes only those underlying truths. This proposal echoes Mäki's point on "meta-claims", but adds additional constraints tied to the subject matter of the derivation.

Both elements of this proposal are required to make sense of the contrast between explanatory and non-explanatory models. Element (i) focuses on what the model adds to the propositional explanation of the model target. If we make (i) sufficient for a model to be explanatory, then we simply have instrumentalism about explanatory models. Condition (ii) is one way to pin down how the model generates its explanatory generalization. If there is a match between the partial truths found in the model and the truths underlying these partial truths, then it is clear that the idealized model is correctly characterizing what is going on in the target system. This is something like what McMullin had in mind, but it does not require that the scientists be able to actually formulate a wholly true model-based derivation of the explanatory generalization. And, crucially, an agent can provide evidence that both conditions are met without being in a position to carry out the reasoning that Strevens requires. This makes the underlying truth proposal less demanding that what Strevens has proposed.

For any proposed explanation that includes an idealization, the veritist can consider the best interpretation of the non-representational function of that idealization. One possibility is that the function is to assert the explanatory irrelevance of some factor, in Strevens' sense, while another possibility is that the function is to commit the agent to the existence of an explanatory derivation that avoids the idealization and that exploits only the truth underlying the idealization. I claim that even when the truth underlying the idealization is closely associated with the explanatory irrelevance of some factor, the veritist should adopt the underlying truth approach, and not Strevens' approach. This is because a commitment to partial truth does not require taking a stand on the nature or existence of any fundamental causal relation. I maintain that no past or current scientist is in a position to know of the existence of any fundamental causal relation. So, it would not be reasonable for any scientist to put forward a proposition with the non-representational function that Strevens singles out. By contrast, scientists are often in a position to commit themselves to the partial truth of the idealizations they use.

The advantages of the underlying truth approach are tied up with what I have called the autonomy of explanation. Perhaps the most important place where the autonomy of explanation does philosophical work is when we consider cases of radical change in the history of science. One striking feature of our tides case is that we continue to judge Newton's model to be explanatory even though our current theory of gravitation rejects gravitational forces and Newton's law of universal gravitation. Despite our rejection of Newton's account of gravitation, we still judge his model to be superior to other purported explanations of the tides, such as those that appealed to mechanical or magnetic processes. Some like Bokulich take this aspect of our practice to show that explanations do not have to be true (Bokulich 2016, section V). Others use it to motivate some kind of instrumentalism about explanatory models. I maintain that the best defense of veritism will legitimate this aspect of our scientific practice by indicating the basis for our different treatment of these models from the history of science. Partial truth offers a promising angle here. Consider again the counterfactual generalization that Newton extracted from his model: were the earth to rotate much more slowly, then the tides would only occur once a day at these locations. We continue to endorse this generalization using our current understanding of gravitation. But we cannot endorse the appeal to gravitational forces that played an essential role in Newton's derivation of this generalization. For us, invoking the pattern of tidal forces that is determined by the gravitational forces at work was a mistake on Newton's part. Despite this radical change in our understanding of gravity, there is still a claim that is generated by Newton's model that we can endorse. This is the claim that characterizes the tides in gravitational terms. And notice that even though the claim that the tides are caused by gravitational forces is false, this claim is partially true. For one entailment of the claim that the tides are caused by gravitational forces is that the tides can be characterized in gravitational terms.[14] And this latter claim does not add to the subject matter of the former claim. This means that Yablo's conditions for partial truth are met. That is, Newton's claim that the tides are caused by gravitational forces is partially true.[15]

Consider, then, Newton's proposed explanation of the tides from Newton's own perspective. As already noted, Newton should be neutral on the existence and character of any fundamental relation of causation. In addition, Newton should commit himself to the partial truth of the idealizations that are exploited in the explanatory derivation of his causal generalization. Finally, Newton should be aware that his whole theory of gravitational forces may be false. Still, he can rationally endorse his model as explanatory if he has reason to suppose that the claims that he makes about gravitational forces in relation to the tides are partially true. If so, he can rationally commit himself to the existence of a wholly true successor derivation of his explanatory generalization that will proceed through the truths that underlie these falsehoods. This successor may involve a radically different theory, with different vocabulary. But this kind of dramatic change is consistent with the original derivation mapping onto the successor derivation in the way that the underlying truth proposal requires.

Strevens' official approach to idealization has little to say about these sorts of radical changes. However, he does briefly discuss how we should currently evaluate Newton's explanation of Kepler's laws, given that it deploys "a force acting directly between objects rather than by way of mass's effect on space time" (Strevens 2008, 328). Newton's derivation can be recast in a contemporary form by simply deleting the problematic aspects of Newton's assumptions about gravity. What results is a derivation with a "black box" that fails to specify how gravity connects the masses to the required trajectories. This derivation cannot count as explanatory for the resulting assumptions fail Strevens' cohesion requirement: a wide variety of types of fundamental causal processes are consistent with these assumptions. These derivations only count as explanatory, then, with respect to what Strevens calls an explanatory framework. A framework involves adding an implicit "given that ..." prefix to an explanatory claim (Strevens 2008, 150). So, while it is wrong to say that X died because X did not receive an antidote, it is right to say that, given that X has ingested a poison, X died because X did not receive an antidote. The framework enables a genuine explanation, although the explanation is not as good as a framework-independent genuine explanation. For Strevens' Newton case, the framework condition is "given an inverse-square dependence" (Strevens 2008, 329). This obviates the lack of cohesion of the derivation, but only at the cost of relativizing the explanation to this framework.[16]

The choice, then, is between imposing a cohesion requirement based on fundamental causation or opting for the more open-ended underlying truth test for explanatory models.

Both options endorse a non-representational function for idealizations in a model-based derivation of an explanatory generalization. But the underlying truth test is arguably more faithful to scientists' reluctance to interpret their models in fundamental terms. Autonomous explanations that aim to remain neutral on what is fundamental seem more likely to be true, and so a veritist is better served by our more flexible underlying truth test.

V. Conclusion

Scientists often explain using models that incorporate idealizations. This feature of scientific practice has been used to argue that explanations need not be wholly true. I have argued that the best way to defend veritism is to maintain that a model is explanatory when it generates an explanatory generalization in the right way. This right way is consistent with the presence of falsehoods in the derivation of the generalization so long as there is an appropriate truth underlying each falsehood. A scientist can then rationally claim that their model is explanatory if they have evidence that these underlying truths exist. It remains to be seen how this evidence is to be assembled or what form of scientific realism is consistent with this approach to model-based explanation.

Acknowledgments

This paper has benefitted enormously from comments and questions by a number of audiences: (i) the Philosophy Colloquium of Oakland University (March 2019), (ii) the 11th Annual Auburn Philosophy Conference (April 2019), (iii) Idealization Across the Sciences, Institute of Philosophy, Prague, Czech Republic (June 2019), and (iv) Mathematics and its Applications: Philosophical Issues, University of Leeds (Sept. 2019). I am especially grateful to

Alisa Bokulich, Joyce Havstad, James Nguyen, Angela Potochnik, Collin Rice, Juha Saatsi, and Michael Strevens for critical feedback. Two anonymous referees for this journal also provided very helpful suggestions for improvement.

References

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

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

Bokulich, A. (2012). Distinguishing explanatory from nonexplanatory fictions. *Philosophy of Science* 79: 725—737.

Bokulich, A. (2016). Fiction as a vehicle for truth: Moving beyond the ontic conception. *The Monist* 99: 260—279.

Bokulich, A. (2017). Models and explanation. In L. Magnani & T. W. Bertolotti (eds.), *Springer Handbook of Model-Based Science*. Springer, pp. 103—118.

Bokulich, A. (2018). Representing and explaining: The eikonic conception of scientific explanation. *Philosophy of Science* 85: 793—805.

Craver, C. (2014). The ontic account of scientific explanation. In M. Kaiser, O. Scholz, D. Plenge & A. Hütterman (eds.), *Explanation in the Special Sciences*. Springer, pp. 27—52.

Fletcher, S. (2019). On the reduction of general relativity to Newtonian gravitation. *Studies in the History and Philosophy of Modern Physics* 68: 1—15.

Lawler, I. & E. Sullivan (forthcoming). Model explanation versus model-induced explanation. *Foundations of Science*.

Mäki, U. (2012). The truth of false idealizations in modeling. In P. Humphreys & C. Imbert (eds.), *Models, Simulations, and Representations*. Routledge, pp. 216—233.

McMullin, E. (1985). Galilean idealization. *Studies in the History and Philosophy of Science* 16: 247—273.

Newton, I. (1999). *The Principia: Mathematical principles of natural philosophy*. I. B. Cohen, A. Whitman & J. Budenz (trans.). Oakland, CA: University of California Press.

Potochnik, A. (2011). Explanation and understanding: An alternative to Strevens' *Depth*. *European Journal for the Philosophy of Science* 1: 29—38.

Potochnik, A. (2015a). Causal patterns and adequate explanations. *Philosophical Studies* 172: 1163—1182.

Potochnik, A. (2015b). The diverse aims of science. *Studies in the History and Philosophy of Science*, Part A 53: 71—80.

Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science* 83: 721—732.

Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.

Potochnik, A. (forthcoming). Idealization and many aims. *Philosophy of Science (Symposium Proceedings)*.

Price, H. (2013). *Expressivism, pragmatism and representationalism*. Cambridge University Press.

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

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

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

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

Saatsi, J. (2012). Idealized models as inferentially veridical representations: A conceptual framework. In P. Humphreys & C. Imbert (eds.), *Models, Simulations, and Representations*. Routledge, pp. 234—249.

Saatsi, J. (2016). On the 'indispensable explanatory role of mathematics'. *Mind* 125: 1045—1070.

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

Strevens, M. (2012). Replies to Weatherson, Hall, and Lange. *Philosophy and Phenomenological Research* 84: 492—505.

Strevens, M. (2017). How idealizations provide understanding. In S. Grimm, C. Baumberger & S. Ammon (eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science*. Routledge, pp. 37—49.

Strevens, M. (2019). The structure of asymptotic idealization. *Synthese* 196: 1713—1731.

Weatherson, B. (2012). Explanation, idealisation and the Goldilocks problem. *Philosophy and Phenomenological Research* 84: 461—473.

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

Woodward, J. (2003b). Experimentation, causal inference, and instrumental realism. In H. Radder (ed.), *The philosophy of scientific experimentation*, University of Pittsburgh Press, 87—118.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review* 115: 1—50.

Yablo, S. (2014). *Aboutness*. Princeton University Press.

Yablo, S. (2020). Models and reality. In A. Levy & P. Godfrey-Smith (eds.), *The Scientific Imagination*. Oxford University Press, pp. 128—153.

Endnotes

[1] Explanations of the tides have been recently discussed by Bokulich 2016, section V and Saatsi 2016, 1062. Saatsi emphasizes how explanations are autonomous in this case, drawing on Woodward 2003a, 224 and Woodward 2003b. I return to this feature in section IV.

[2] See also Bokulich 2012, 2016, 2017, and 2018.

[3] In addition, Bokulich argues that some of these explanations are not causal explanations. .

[4] Of course, not all models are mathematical models. For other kinds of models, other kinds of specifications will be involved.

[5] See Potochnik 2015a, 2015b, 2016, 2017 and forthcoming.

[6] See also Rohwer & Rice 2016. I do not consider here their requirement that an explanatory model be conducive to understanding. See also Lawler & Sullivan (forthcoming) for another useful discussion of how models can "induce, enable, or generate explanations" (23). However, Lawler & Sullivan do not offer any test for when such models should count as explanatory.

[7] See also McMullin 1985, 261, emphasized by Bokulich 2011, 37 and 2017, 104.

[8] See also Batterman & Rice 2014 and Rice 2018.

[9] See also Saatsi 2012 for a discussion of what he calls "inferentially veridical representations".

[10] For Strevens, an idealization can enhance the state of understanding of an agent who grasps the explanation. This point is developed in more detail in Strevens 2017. I must postpone consideration of the complex debates about the relationship between truth, explanation and understanding due to reasons of space.

[11] Despite its importance, Strevens' approach to idealization has received little critical discussion. An important exception is Weatherson 2012. Weatherson focuses on the problems that cohesion creates for making sense of much of our explanatory practice. See Strevens 2012 for a short reply.

[12] Yablo 2014.

[13] Yablo uses his notion of subject matter to isolate a role for statements S that are generated by a model ω but that are false of the actual world α: "S's truth in ω signals its truth in α about a subject matter m that ω and α agree on" (2020, 144). However, it is not clear what Yablo would say about our question concerning explanatory models. Also, there are various ways to identify the subject matter of a proposition that have implications for debates about models and explanations. I must reserve an examination of these difficult issues for future work.

[14] See Fletcher (2019) for a promising examination of how to connect Newtonian gravitation with general relativity. Fletcher's conclusion is that "if only in retrospective rational reconstruction, the transition to relativity theory from Newtonian physics involves much more conceptual continuity than is usually emphasized" (Fletcher 2019, 13).

[15] This should make clear that I ultimately agree with Rice and Strevens that many idealizations are useful because they make overly specific, false claims, whose corresponding less specific claims are true. My disagreement with these authors thus turns on how this diagnosis of the usefulness of idealizations is deployed in an analysis of when a model is explanatory.

[16] See Potochnik 2011 for an objection to Strevens based on his use of explanatory frameworks. In a recent paper Strevens has developed an account of "asymptotic idealization" that is considerably more flexible than the "simple idealization" considered in his earlier work (Strevens 2019). In this paper Strevens is clear that he aims to offer "a rational reconstruction" of what scientists have in mind when advancing explanations with models (2019, 1724). So one way to summarize my concerns is that Strevens' reliance on cohesion is not consistent with the aim of rational reconstruction.