Synthese (2021) 199:3219–3244
https://doi.org/10.1007/s11229-020-02931-6

UNREALISTIC MODELS

Do fictions explain?

James Nguyen[1,2,3]

Received: 1 April 2019 / Accepted: 20 October 2020 / Published online: 25 November 2020

Abstract
I argue that fictional models, construed as models that misrepresent certain ontological aspects of their target systems, can nevertheless explain why the latter exhibit certain behaviour. They can do this by accurately representing whatever it is that that behaviour counterfactually depends on. However, we should be sufficiently sensitive to different explanatory questions, i.e., 'why does certain behaviour occur?' versus 'why does the counterfactual dependency invoked to answer that question actually hold?'. With this distinction in mind, I argue that whilst fictional models can answer the first sort of question, they do so in an unmysterious way (contra to what one might initially think about such models). Moreover, I claim that the second question poses a dilemma for the defender of the idea that fictions can explain: either these models cannot answer these sorts of explanatory questions, precisely because they are fictional; or they can, but in a way that requires reinterpreting them such that they end up accurately representing the ontological basis of the counterfactual dependency, i.e., reinterpreting them so as to rob them of their fictional status. Thus, the existence of explanatory fictions does not put pressure on the idea that accurate representation of some aspect of a target system is a necessary condition on explaining that aspect.

Keywords Models · Fictions · Representation · Explanation

1 Introduction

Suppose someone asks you why the difference between high and low tide, the tidal range, changes throughout the lunar month. You might answer that it's the relative positions of the sun, the moon, and the earth that explain this difference. Depending

✉ James Nguyen
james.nguyen@sas.ac.uk

1 Institute of Philosophy, University of London, London, UK

2 Department of Philosophy, University College London, London, UK

3 Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, UK

on the lunar cycle, either the sun and the moon are positioned in such a way as to ensure that their gravitational forces align, thus producing spring tides (the tidal effect of the sun and the moon reinforce each other), or their force vectors are at right-angles to one-another, thereby producing neap tides (the smaller solar tidal effect is orthogonal to the larger lunar effect). Spring tides are higher (at high tide, and lower at low tide) than neap tides, so during spring tides the tidal range is greater than during neap tides. I think that this explains why the tidal range varies across the lunar month.

But there is a complication. This answer involves a 'fiction'. There's no such thing as Newtonian gravitation. From our current perspective the ocean isn't acted on by a gravitational force; it's just trying to 'go straight in a crooked world' (Bokulich 2016, p. 273).[1] Nevertheless, it doesn't seem detrimental to the model that it involves such a force, indeed involves such a force essentially (without the force there is no such model and no explanation), and moreover there doesn't seem to be any scientific or philosophical pressure to attempt to replace the explanation based on the force model with one that doesn't involve such a fiction, i.e., one that made reference to the curvature of a spacetime manifold instead (if this could be done at all).

To a philosopher, this might seem puzzling. The model that explains why the tidal range changes throughout the month seems to represent it as being induced by something that we know isn't there in the world. In this sense the model is an explanatory fiction. But how can fictions explain? And if they do, do we need to develop a philosophical account of explanation that gives up on the requirement that explanations are accurate representations?

I argue that this puzzle arises from two conflicting intuitions or observations. The first is that accurate representation is necessary for explanation. The second is that fictional models, which on the face of it are inaccurate in crucial respects, nevertheless provide explanations. In this paper I show how this conflict can be dissolved. First by disambiguating the request for a 'first-order' explanation (in this case: 'why does the tidal range vary across the lunar month?') from the request for a 'second-order' explanation ('why does what explains such a variance play the explanatory role that it does?'). Then, if fictional models are taken to provide first-order explanations (only), they do so without being inaccurate in the relevant respects. Second, if fictional models are also taken to provide second-order explanations (which is much less clear), then we should reconsider their fictional status. A closer look at the explanatory uses of fictional models will demonstrate that they do not conflict with the idea that accuracy is necessary for explanation. So whilst this paper is framed in terms of whether or not fictional models explain, its broader target is a defence of the idea that accurate representation is necessary for explanation.

The structure of the paper is as follows. In Sect. 2 I clarify two notions of 'fiction' that have been invoked in the literature on scientific representation and explanation. I distinguish between the sense of fiction as ontology (a work of fiction), and the sense of fiction as inaccuracy (fiction as misrepresentation). It's the latter which is primarily relevant to the question of whether or not fictions can explain in some philosophically interesting sense. In Sect. 3 I introduce Bokulich's (2008a; 2008b; 2009; 2011; 2012; 2018a; 2018b) account of model explanation, which is explicitly designed to allow that

[1] Bokulich reports that she owes this expression to John Stachel.

(inaccurate) fictional models explain, and thus puts pressure on the idea that accurate representation is necessary for explanation.[2] The crucial aspect of this account is that a model explains some phenomenon by accurately representing its modal profile, i.e. how it would change were various other features to change (and that this can be done even if the model inaccurately represents the ontology of the target). In Sect. 4 I draw on a distinction between first-order and second-order explanations (*cf.* Skow 2016, 2017) and claim that Bokulich's examples are ambiguous between them (in doing so I also show how Skow's distinction can be utilised without adopting his account wholesale). In Sect. 4.1 I argue that (inaccurate) fictional models can offer first-order explanations of certain phenomena, but they do so in such a way that their fictional nature doesn't feature in the explanation itself. In Sect. 4.2 I pose a dilemma for a staunch defender of the view fictional models can explain: either they can provide second-order explanations, but they do so because they are interpreted such that they end up accurately representing the ontological basis of the counterfactual dependency, i.e., reinterpreting them so as to rob them of their fictional status; or they cannot provide such explanations, precisely because they are fictional, i.e., they are interpreted literally and therefore misrepresent said ontological basis. Section 5 concludes by emphasising that accurate representation, at least of both the features of the target that are to be explained, the explananda, and the features that do the explaining, the explanantia, remains a necessary condition on scientific explanation (or at least that fictional models do not challenge this).

2 Fictional models

At least some scientific models are fictions. What does this mean? For the likes of Godfrey-Smith (2006), Frigg (2010), Frigg and Nguyen (2016a), and others, this is an attempt to analyse the ontology of scientific models. The idea is that, ontologically speaking at least, scientific models should be thought of as akin to the situations described by works of fiction. Just as we can talk about a farmyard populated by animals—a pig called 'Napoleon', a horse called 'Boxer', and so on—who interact with each other in various ways—enacting different governing polices and engaging in revolutions—we can also talk about fictional systems like idealised pendula, celestial bodies subject to gravitational forces, and populations of animals with unlimited food. Advocates of this view of fictional models urge us to think of the sorts of systems that are described in scientific textbooks and research papers as akin to the sorts of systems that are described in works of fiction.

Of course without an account of the nature of fictional situations, this approach risks analysing the already confusing (what models are) in terms of something just as obscure (what fictions are). Luckily, philosophers of science can appeal to various different accounts offered in discussions of the ontology and metaphysics of fiction in order to cash out the analogy. Some prefer to think that this means that scientific models

[2] I focus on Bokulich's work throughout since, to my mind, she provides the clearest expression of the tension between fictional models in science and the role of accuracy in explanation. For other useful discussions of fictional models see the papers collected in Suárez (2009a) and Woods (2010), and the references in Sect. 2.

are abstract artefacts created by humans but existing as abstracta independently of us (Thomasson 2020). Others adopt a more deflationary view and take it to mean that scientific models are imaginary entities that are associated with Waltonian 'games of make-believe' (Frigg 2010; Frigg and Nguyen 2016a). And in principle any position concerning the nature of fictional entities could be utilised in service of developing an account of what scientific models are, ontologically speaking.

However this is worked out, it's crucial to note that when we shift from the ontological question (what models are) to the semantic question (how they represent their target systems), or functional question (what role they play in scientific practice, which is the focus of most contributors to Suárez (2009a)), nothing in that version of the fictional account demands that 'fictional' models in this sense are in any way 'false' or 'inaccurate' representations of their actual target systems. I believe that the world of Orwell's *Animal Farm* is an accurate representation of the political pressures faced by the USSR. But even if this is disputed (e.g. because the Politburo didn't convene in a farmyard), there are plenty of other works of fiction where the situations described in the work are present in the real world too. For example, Cambridge, as described in Faulk's *Englby*, matches Cambridge in the real world, and the Dublin described in Rooney's *Normal People* matches the real Irish capital. Actual scenarios can be described in works of fiction, and those scenarios needn't be 'false' or 'inaccurate' representations of the real world. Fictional models in this sense do not need to raise any novel explanatory questions.

However, 'fiction' has another reading, a reading that does connote inaccuracy. In this sense it might be considered fictional that the audience at Donald Trump's inauguration was the biggest ever, or fictional that he won the biggest margin in the US electoral college since Ronald Reagan. One way of understanding this sense of fiction is in terms of truth-values. A sentence is fictional in this sense only if it is false, or alternatively if its truth-value is not relevant to the function it is deployed for (Suárez 2009b, pp. 11–13). In the context of model-based science things are a little more complicated for two reasons. First, and most importantly, following Weisberg (2007b), and indeed the majority of the contemporary literature on scientific modelling, I characterise model-based science in terms of its indirect nature.[3] As such, model descriptions (which are linguistic) specify, or somehow describe, model systems, and it is the latter that are the units that represent the target systems of interest.[4] So on the indirect account, model descriptions are, strictly speaking, not the 'fictional' items involved in the explanations: their truth-values aren't evaluated with respect to target systems; they're evaluated with respect to the model systems they specify, and this isn't at issue when it comes to how 'fictional' (inaccurate) models, or indeed any of kind of model, explain(s).

[3] For dissenting voices see Toon (2010a, b, 2012) and Levy (2012, 2015).

[4] If we were to adopt the 'fiction view of models' discussed previously these models would be fictional scenarios, but I am not committed to this analysis here. Other options are that these model systems are mathematical structures (Suppes 1960/1969; van Fraassen 1980; Suppe 1989; van Fraassen 2008) or abstract objects (Giere 1988).

Since it's the model systems that are the primary units of representation, how should we understand them as providing 'fictional' (inaccurate) explanations? This depends on how the model systems are understood as representations, which brings us to our second complication. Model systems are not obviously truth-bearers.[5] They represent their targets in a manner analogous to the way in which maps, or concrete models—such as the Phillips–Newlyn machine (Frigg and Nguyen 2018), ball-and-stick models of molecules (Toon 2011), or the U.S. Army Corps of Engineers model of San Francisco Bay (Weisberg 2013)—represent their targets. As Giere (1994, p. 11), notes (with particular reference to maps), such representations 'have many of the representational features we need for understanding how scientists represent the world. There is no such thing as a universal map. Neither does it make sense to question whether a map is true or false'. So when models are fictional in the sense that Trumpian boasts are fictional, rather than saying they are false, it's better to say that such a model is an *inaccurate representation* of its target.[6]

But here one might object that all models are inaccurate in some respect (Teller 2001). And if this is the case then all models are fictional in this sense. That doesn't seem right. Bokulich (2008a, b, 2011, 2012, 2018a, b) draws a distinction between models that introduce kinds of entities, properties, states, or processes ('features' for short) known not to be present in the target system, and those that distort the actual features of the target. The former are appropriately dubbed 'fictional' models in the sense that they (seem to, see Sect. 4.2) inaccurately represent the kinds of ontological features present in the target system, whereas the latter accurately represent the kinds of ontological features that are present, but do so in a way that distorts their details.

Now, notice that these models do represent actual systems in the world, they just misrepresent certain ontological features of those systems. As such, these models are 'fictive' in Suárez (2009b, c) sense, and should be distinguished from what he calls 'fictional' models, i.e. representations of systems that don't exist.[7] Models that represent non-actual targets raise all sorts of interesting questions (as do models that don't represent any target whatsoever, non-actual or otherwise (*cf.* Weisberg 2013, Chapter 7)). However, if fictional models are to put pressure on the idea that accurate representation is necessary for explanation, the puzzle is most pressing when the thing that is to be explained is something in the actual world, not something that doesn't exist. Moreover, as we will see, Bokulich's preferred examples of explanatory fictions are precisely models that are targeted at actual systems (e.g. tidal behaviour). So from now on I will use the term 'fictional model' to refer to a model which (i) represents an actual target system; but (ii) misrepresents that system's ontology. We can try to clarify

[5] For a dissenting voice see Mäki (2011).

[6] Notice that the distinction drawn by Suárez (2009b, pp. 11–13) in terms of thinking about fictions in truth-conditional or functional terms, i.e. in terms of inaccuracy or in terms of the irrelevance of accuracy, carries over here. We can also talk about fictional models in the sense that they have aspects whose representational (in)accuracy is irrelevant to their function.

[7] So notice that, as a matter of unfortunate terminology, I'm using the term 'fictional' as he uses 'fictive', rather than as he uses 'fictional'. I address his work in Sect. 4.2.

this by (very briefly) recapping the examples that Bokulich uses to illustrate her account.

Newtonian tides

Bokulich (2016, Sect. V) cites the National Oceanic and Atmospheric Administration (NOAA) website for a standard explanation of the tides. The explanation is given by considering a Newtonian model of the moon-earth system. The bodies in the model are assumed to revolve together around their common centres of mass. They are held together by a gravitational attraction force, and simultaneously kept apart by an equal (only at the centres) and opposite centrifugal force resulting from their individual revolutions around their common centre of mass. On the surface, i.e. not the centre, of the earth (in the model) there is an imbalance between these forces: in the hemisphere of the earth closest to the moon (in the model) there is net tide-producing force acting in the direction of the moon's gravitational attraction, and on the hemisphere opposite the moon there is a net tide-producing force acting in the direction of the centrifugal force, i.e., away from the moon. These two forces result in two tidal bugles on opposite sides of the earth (high tides). We can introduce a third body to our model, the sun, which exerts an analogous differential force in its revolution around the earth–sun centre of mass.[8] However, since the tide-producing force (in the model) is inversely proportional to r^3, where r is the distance between the celestial bodies between which the force holds, and the sun is further from the earth than the moon, despite its mass the sun exerts only half of the force exerted by the moon on the earth resulting in a smaller tidal bulge.[9] When the sun and the moon are aligned their respective tidal bulges reinforce each other (spring tides), when they are orthogonal to one another there is destructive interference producing neap tides. This is what explains the variance in tidal range across a lunar month.

As Bokulich (2016) points out, force models are also used to explain vast numbers of other complications that arise with the tides (why some areas only have one high tide a day, what effect the depth of the ocean has on the tides, and so on). The crucial thing about all of these models (Bokulich argues) is that gravity (in the model) is identified with a classical force. The model's gravitational forces are Newtonian; proportional to the masses of the objects involved and the reciprocal of the square of their distance from one another. But, at least according to general relativity, we know that gravity is actually 'the curved geodesic structure of a 4-D spacetime manifold whose metric is determined, in accordance with the Einstein field equations, by the stress-energy tensor of the matter fields' (Bokulich 2016, p. 273). Gravity (in the world) is not a classical force; it's the curvature of spacetime. So gravity (in the model) is a fiction. It doesn't *distort* the known features of the target, the curvature of the manifold, it

[8] We needn't actually construct and calculate the details of a three-body model for the sort of explanation I am concerned with in this paper.

[9] It's worth noting here that the tide-producing force (in the model) is inversely proportional to r^3 even though the gravitational force (in the model) is inversely proportional to r^2. This is because the former concerns the *difference* between the forces at various positions on the earth's surface.

introduces a new kind of feature, a force vector between bodies in the model, which we know isn't actually in the target system. This makes the model fictional.

Quantum dot

Bokulich Bokulich (2008a, b, 2011, 2012) draws on a wealth of examples from semi-classical physics, both to develop a novel understanding of the relationship between classical and quantum physics, and to motivate her account of what fictional models are and how they explain (it's more accurate to say that she takes it for granted that these models explain and this requires developing a novel philosophical account of explanation to accommodate them). Here I'll focus on her discussion of quantum dots (Bokulich 2012), a specific kind of target system related to quantum scarring (which is discussed throughout her work). A quantum dot is a semiconductor in which electrons are confined to a very small 2D plane. When a quantum dot is weakly coupled to external leads there is the possibility of an electron tunnelling into the dot. At certain voltages one can compensate for the Coulomb repulsion of the electron already in the dot, and the charge in the dot will fluctuate between N and $N + 1$ electrons, resulting in a series of peaks in the dot's conductance. In order to understand the patterns of conductance we construct a model that matches the shape of the dot (and includes the the external leads and so on), and investigate which *classical* orbits would be allowed within the dot. Because of the dot's irregular shape the classical system is ergodic, which means almost all of the trajectories are not periodic. But there is an (infinite, but measure 0) set of unstable orbits that are periodic, and it's these that correspond to the peaks in the conductance of the dot:

> the period of modulation of the Coulomb-blockade peaks is determined by the periods of the classical orbit that intersects with the leads, the frequency of the oscillations is proportional to the area covered by the orbit, and the peak distribution is determined by the Lyapunov exponent of the classical orbit (Bokulich 2012, p. 729).

The model thus represents the patterns of conductance of the dot as being determined via the voltage and position of the external leads, understood quantum mechanically in terms of the initial conditions of an electron's wavefunction, and the shape of the dot itself (in terms of which of its internal orbits are periodic). However, the model does so by invoking the idea that electrons within the dot follow classical orbits, something we know isn't the case. In fact, the pattern of conductance peaks, and indeed quantum scarring more generally, is known to actually be determined by complex interference patterns in the electrons' wavefunctions, which are themselves properly thought of as spread out throughout the dot. So classical orbits (in the model of the dot) are a fiction. The model doesn't distort the known ontology of the dot, the wavefunction, it introduces a novel kind of feature, electrons following classical orbits, that we know isn't actually present in the target system.

So what these case studies demonstrate is that fictional models (seemingly, again see Sect. 4.2) misrepresent the ontological features present in a target in a way that involves introducing novel kinds of features—Newtonian forces, classical orbits—

known not to be present in the target system. In contrast, idealised models that are not fictions should be analysed in terms of distorting the features known to be present. For my current purposes this suffices to distinguish what I mean by fictional models.[10]

Now, whilst there are various treatments of the epistemic capabilities of idealised models that distort known target features (e.g. McMullin 1985; Jones 2005; Weisberg 2007a; Strevens 2008; Nguyen 2020), the question of how fictional models, models that invoke features known not to be present, work has not, to the best of my knowledge, been explicitly addressed in these terms (beyond Bokulich's work). And it's these sorts of models that seem to pose a particularly novel kind of philosophical puzzle. If a fictional model M of some target system T represents T's behaviour as being generated by a completely different ontological feature of T than the one we know is actually present, then how can M play an explanatory role? In contrast, representing T's behaviour as being generated by some feature of T that we know is there, but inaccurately representing the way that the feature generates the behaviour, seems to pose a different and (possibly) less threatening kind of puzzle (or at least a puzzle that many authors have already attempted to address). The aforementioned models are cases in point. The Newtonian model, if interpreted literally, represents the differences in tidal range as being determined by the relative positions/masses of the sun and the moon with respect to the earth, via the gravitational forces of each on the latter. The quantum dot model, if interpreted literally, represents the conductance patterns as being determined by the initial conditions (e.g., the shape of the dot), via the periodic classical orbits that they allow. But we know that there's no such thing as gravitational force and that electrons do not follow classical orbits. So how do these models explain? Before addressing that question it will prove useful to specify what it means for a model to explain. That is the task of the next section.

3 Model explanation

What it means for models to explain is a thorny topic. For my current purposes I'm focusing on models that explain via representing certain counterfactual dependencies that hold in the target. It's important to note that I'm not claiming that this exhausts all kinds of model explanations.[11] For my current purposes what matters is just that *some* models explain via representing dependencies, and that *some* of these models appear to be fictional in the sense discussed in the previous section. In the first instance it's these

[10] I'm not claiming that there is a strict clear distinction here. One might be able to reinterpret the distortion of a known feature as the introduction of a new (non-actual) ontological feature, and vice versa. This strikes me as analogous to the fact that we can reinterpret cases of abstraction (e.g. a model that ignores friction) as idealisation (e.g. a model that misrepresents a friction coefficient as 0) (Jones 2005). However, I do think we have enough of a pre-theoretical grasp of the fiction/non-fiction distinction to motivate the rest of the discussion.

[11] For example I'm leaving it open that: models can explain by providing comparison cases (Kennedy 2012; Jebeile and Kennedy 2015); by demonstrating that previously held necessity or impossibility hypotheses are false (Grüne-Yanoff 2009); by demonstrating various features of the theories in which they are embedded (Luczak 2017); via renormalisation group transformations (Batterman and Rice 2014); and other forms of non-causal explanation (Reutlinger and Saatsi 2018), and that models might offer these kinds of explanation in a manner that doesn't involve the representation of counterfactuals.

kinds of model explanations that seem to threaten the idea that explanation requires accurate representation. So in order to get a handle on how fictional models explain it's useful to analyse cases where the notion of explanation is as straightforward as possible, to ensure that we keep in focus the fictional aspects of models that play such explanatory roles. As such we can set aside the complications that arise from more exotic forms of model explanations.

Bokulich (2008a, b, 2009, 2011, 2012, 2018a, b) develops an account of model explanation that is explicitly geared to allow for fictional models to explain in this way. The account involves three main claims: first, the explanation must make essential reference to a scientific model; second,

> that model explains the explanandum by showing how there is a pattern of counterfactual dependence of the relevant features of the target system on the structures represented in the model. That is, the elements of the model can, in a very loose sense, be said to 'reproduce' the relevant features of the explanandum. Furthermore, as the counterfactual condition implies, the model should also be able to give information about how the target system would behave, if the structures represented in the model were changed in various ways (Bokulich 2008a, p. 226; *cf.* Bokulich 2011, p. 39);

and third, the model explanation must satisfy a 'justificatory step' that specifies what the domain of applicability of the model is, and shows that the phenomenon in the real world to be explained falls within that domain.[12]

My primary focus here is on the second condition (although I return to the third in Sect. 4.2). Drawing on Woodward (2003), Bokulich argues that this condition requires that models explain in virtue of answering 'what-if-things-had-been-different' questions, or *w*-questions.[13] As Woodward puts it, such an explanation provides 'information about a pattern of counterfactual dependence between explanans and explanandum' (2003 p. 11), and:

> an explanation ought to be such that it can be used to answer what I call a what-if-things-had-been-different question: the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways. We can also think of this as information about a pattern of counterfactual dependence between explanans and explanandum (*ibid.*).

The conclusion that Bokulich draws from this, combined with her case studies as evidence, is that a model can answer *w*-questions even whilst misrepresenting the ontology of the system, i.e., being fictional. If correct, this would seem to put pressure on the idea that accurate representation is a necessary condition for explanation.

[12] Bokulich (2018b, p. 143) offers a fourth condition to allow for multiple explanations of the same explanandum. I agree with this condition. I won't talk about it here.

[13] Bokulich departs from Woodward in not spelling this condition out in terms of manipulability or intervention. This is deliberate in the sense that it is supposed to allow her account to capture non-causal explanations. For discussion of this aspect of her account see Saatsi and Pexton (2013) and Schindler (2014). For my current purposes I'm focusing on how it works when the explanations in question are causal, but this doesn't make too much of a difference to the issues I'm discussing.

The problem, however, is that when stated like this, the counterfactual condition doesn't fully specify the relationship between the 'things which could have been different' and the fictional aspects of the model. In particular, I think that this condition is ambiguous between two different readings, and as we will see, it's this ambiguity which makes the fact that fictional models can explain seem mysterious.

The first reading can be specified as follows. We will say that: a model M explains a target's behaviour, A (the explanandum), only if M accurately represents A and accurately represents the counterfactual dependency of A on some other target feature, B (the explanans). More precisely, taking care to keep reference to features of the model and features of the target distinct:

Counterfactual Model Explanation Condition: for a model M of a target T, where T has a feature A which can take values in $\{A_1, A_2, ..., A_n\}$; M explains A only if: M has a feature P which can take values in $\{P_1, P_2, ..., P_{n'}\}$; P accurately represents A; M has a feature Q which can take values in $\{Q_1, Q_2, ..., Q_{m'}\}$; Q accurately represents some feature in the target B which can take values in $\{B_1, B_2, ..., B_m\}$; and the dependencies of the values of P on the values of Q accurately represents how the values of A depend on the values of B.[14]

In the simplest case we might have bijections between each of these sets of values in such a way that specifying a value Q_i in the model serves to both represent a value of B_j, a feature in the target, and also fixes a value of P_k in the model such that P_k accurately represents the value A_l in the target that would arise were B_j the actual value of B in the target.[15] So, in order for a model to explain some behaviour A of the target system, it needs to accurately represent that behaviour, the explanandum (via the feature P of the model), and accurately represent the explanans of that behaviour B (via the feature Q of the model), in the sense that the model accurately represents how A would change, were B to change.[16]

I take it that this captures what Bokulich means when she says that in cases of model explanation 'we require that the counterfactual structure of [the model] be isomorphic in the relevant respects to the counterfactual structure of [the phenomenon to be explained]' (2011, pp. 39, 43) (she admits that she is using isomorphism in a loose sense), and that in cases of model explanation 'the elements of the model can, in a very loose sense, be said to 'reproduce' the relevant features of the explanandum' (Bokulich 2008a, p. 226; *cf.* Bokulich 2011, p. 39).[17] Moreover, I take it that this is suggested in discussions such as the following:

> the semiclassical model allows one to answer a wider variety of *w*-questions about how the system would behave if certain parameters were changed-and pro-

[14] Although this condition is phrased as a necessary condition, I restrict its scope to cases where models explain by accurately representing a target's counterfactual behaviour. I'm open to other ways in which models can perform explanatory roles, *cf.* fn. 11.

[15] By 'simplest case' I mean simple in the sense of simple to philosophically analyse, not simple in the sense that the explanation is scientifically simple.

[16] It's worth noting here that I'm allowing the variables A and B to range over observable and unobservable features of the target system. I'm grateful to Juha Saatsi for encouraging me to be explicit about this.

[17] This is also how Fang (2019) interprets her account.

> vides this information without having to explicitly carry out the tedious quantum calculations for each possible case (Bokulich 2008a, p. 233),

and when citing (Narimanov et al. 2001):

> the semiclassical model allows physicists to answer a wide range of what-if-things-had-been-different questions. As Narimanov et al. write, from this model they now understand "how as a system parameter varies [such as] the magnetic field, for instance, or the number of electrons in the dot ([as] controlled by varying a gate voltage) - the interference around each periodic orbit oscillates ... When the interference is constructive for those periodic orbits which come close to the leads used to contact the dot, the wavefunction is enhanced near the leads, the dot-lead coupling is stronger, and so the conductance is larger" (2001, p. 2) (Bokulich 2012, p. 731).[18]

In the case of the Newtonian model of the tides both the explanandum, *A*: the tidal range in the target, and the explanans, *B*: the relative positions and masses of the sun, moon, and earth, are accurately represented by the model. (If you're concerned that gravity isn't in the explanans here, bear with me for now; this is addressed below.) Such a model works to explain the tidal range because it specifies how the tides would vary were the positions/masses of the celestial bodies varied (I don't just mean 'vary' in the sense in which they follow their orbits, the model also works to specify how the tides would vary if the bodies were in positions outside of their orbits too, or different masses). In the case of the quantum scarring model, again, both the explanandum, *A*: the patterns in the conductance peaks that are observed, and the explanans, *B*: the initial condition of the system, including the parameters referenced by Narimanov et al. (2001) in the above quotation, are accurately represented by the model.

The models in question then provide answers to *w*-questions, where the things that could have been different are the positions/masses of the celestial bodies, or the initial conditions of the wavefunction (for example). Notice though, that when described in these terms, the model in question is an accurate representation of both the explanans and the explanandum in each case: in particular, neither model is fictional with respect to either the positions of the celestial bodies, or the initial conditions of the wavefunction (and both models accurately represent the target of the explanation, the tidal range and conductance peaks respectively).

There is, however, another reading of Bokulich's second condition of model explanation. Recall that she requires that 'the model should also be able to give information about how the target system would behave, if the structures represented in the model were changed in various ways' (Bokulich 2008a, p. 226; *cf.* Bokulich 2011, p. 39). The question is: what does the phrase 'the structures represented in the model' refer to? Under one reading (the one just discussed), it refers to whatever it is that the relevant target behaviour counterfactually depends upon. But under another reading it refers to the whatever it is that underpins this counterfactual relationship, which in the

[18] Here I assume that the 'parameters' referred to in these quotes are parameters that the target system that interest actually has; whether or not its in a magnetic field; the shape of the quantum dot; the initial conditions of the wavefunction; the voltage; the resulting pattern of conductance peaks; and so on. Another interpretation of this phrase is offered below.

models in question correspond to their *fictional* aspects. Under this reading the model answers *w*-questions where the things that could have been different are fictional: if gravitational force had been proportional to r^4 rather than r^2, or if the classical orbits within the quantum dot had been different. This reading is also suggested at various points throughout Bokulich's work:

> this pattern of dependence allows one to say precisely how the quantum wave-function morphology would change if, for example, the classical periodic orbit had been different, or if the Lyapunov exponent of that same orbit had taken on another value (Bokulich 2008a, p. 227; *cf.* Bokulich 2008b, p. 157)

and

> Bohr's model is able to correctly answer a number of 'what-if-things-had-been-different questions,' such as how the spectrum would change if the orbits were elliptical rather than circular (Bokulich 2011, p. 43).

Under this alternative reading, the explanatory questions being answered by the model are different. The question is no longer the first-order 'why does *T* exhibit *A*?' (that question is answered by a model accurately representing *A*'s dependence on *B*). The question is now the second-order 'why does *T*'s behaviour *A* depend on *B*?' The latter kind of question is now answered by the following:

Second-Order Model Explanation Condition: *M* explains *A*'s dependence on *B* (in the target) only if *M* accurately represents that *A* depends on *B* (in the target) (by meeting the **Counterfactual Model Explanation Condition**) *and* has a feature *R* that accurately represents whatever target feature *C*, it is that *A*'s dependence on *B* itself depends upon.[19]

In the case of the tidal model, again *A* is the tidal range and *B* is the relative positions and masses of the sun, moon, and earth, but the explanandum has shifted. Now, rather than trying to explain *A* itself, a model that meets the second condition attempts to provide an explanatory answer to the question 'why does *A* depend on *B*' and the explanans given invokes *R*, the gravitational force (in the model) that is supposed to represent whatever it is that provides the basis for the counterfactual dependency of *A* on *B* (and in the model, does in fact provide the basis for the counterfactual dependence of *P* on *Q*). Similarly, in the case of the quantum scarring model, the model no longer attempts to explain *A*, the conductance patterns, via representing its dependency on *B*, the initial conditions, but rather attempts to explain why *A* depends on *B*. And it does so by invoking *R*, the classical orbits that are supposed to represent whatever it is that provide the basis for why *A* or more generally, quantum scarring, depends on *B*. Let's now investigate the distinction between first-order and second-order explanations and *w*-questions in more detail.

[19] For the cases relevant here *C* will itself be another aspect of the target system. But as I discuss in the following section this might not be the case for explanations in general.

4 Levels of explanation

Once we grant that models explain by answering *w*-questions, we can distinguish between different kinds of *w*-questions in a way that corresponds to two different kinds of explanation that models can provide:

First-order:

i 'Why does *T* have behaviour *A*?'
ii *M* explains *T*'s having *A* only if *M* accurately represents *A's* counterfactual dependence on feature *B* of the target (**Counterfactual Model Explanation Condition**)
iii *M* provides answers to what-if-*B*-had-been-different questions.

Second-order:

i 'Why does *A* depend on *B*?'
ii *M* explains why *A* depends on *B* only if *M* accurately represents *C*, which *A*'s dependence on *B*, itself depends on (**Second-Order Model Explanation Condition**).
iii *M* provides answers to what-if-*C*-had-been-different questions.

The shift from a first-order explanation to a second-order explanation concerns the explanatory depth of the model in question. There are at least two ways of thinking about 'explanatory depth'. First, one could consider a 'chain' of causes (or dependencies): one explains something by representing its cause, and then provides more explanatory depth by representing the cause of the cause, and so on. Another way of thinking about explanatory depth however, is to explain why the initial explanation explains in the way it does. In this case, one explains something by representing its cause (or dependencies), and then provides more depth by explaining why the causal relationship initially invoked holds in the way that it does. Figuratively, the first kind of explanatory depth concerns extending a chain of dependencies, the second concerns providing more detail about a section of the chain itself. It is the latter that I'm interested in in here. The explanans provided by a model to a first-order question becomes the explanandum of the second-order question: if a model answers the question 'why *A*?' with '*A* depends on *B*', then this first-order explanans becomes the second-order explanandum, whose explanans is an answer to the second-order question 'why does *A* depend on *B*?'.[20] With this in mind we can now turn to how the fictional models discussed above answer first-order and second-order explanatory questions.

4.1 Fictions and first-order explanations

I hope that by now it is straightforward to see how the Newtonian model of the tides and the quantum dot model provide first-order explanations for tidal behaviour and the patterns of conductance respectively. They do so by accurately representing those

[20] This distinction loosely matches Skow's distinction between what he calls first-order and second-order reasons for some phenomena (Skow 2017, p. 907; *cf.* Skow 2016, Chapter 4.2), although note that I'm talking about explanations rather than reasons, and I'm not attempting to provide a 'universalist' account of explanation; I'm focusing on cases where the explanations in question are given in terms of causal counterfactuals, without claiming that this is how *all* explanations work.

features of their targets, and what it is that they counterfactually depend on, namely the positions/masses of the sun, moon, and earth, and the initial conditions of the wavefunction in quantum dots. Now the crucial thing to note here is that although the models are fictional (in the sense that they misrepresent their targets in certain ways), and although these fictional aspects play an essential role in the first-order explanations they offer (without them we wouldn't have the models that represent the relevant dependencies), the fictional aspects themselves don't feature in the first-order explanations. The models explain by representing how changes to the explanans yield changes to the explananda, and they do so accurately despite the fact that in the model the features which represent these aspects of the target are connected to each other via a feature (gravity, classical orbits) that itself misrepresents the features of the target. In both cases, the model provides an explanation of the behaviour A by answering certain *w*-questions: namely, questions of the form 'what would happen to A if B were different?'. These answers explain A by showing that it depends on B. And when answering these questions, the fact that the model contains fictional elements (in the sense that they inaccurately represent what underpins the counterfactual dependence) which play an essential role in *generating* the explanation—since the fictional elements mediate between P and Q in the model, which respectively accurately represent A and B in the target—the misrepresentation does not feature in the explanation itself, since it concerns A and B alone. In this sense, the models are not fictional with respect to the explanans, B, they accurately represent those features of their targets, and how varying those features impacts the explanandum, A, which is also accurately represented.

Now, notice that these explanations do not provide answers to questions like: 'what if gravity had been different?', or 'what if classical orbits had been different?'. These latter *w*-questions, and their potential answers, seem to, at least on the face of it, concern what if the features of the *models* were different, rather than what if features of the *targets* were different. If the models are supposed to provide first-order explanations by answering 'why A?' questions by invoking the fact that A depends on gravity, or classical orbits, one shifts from talking about A, to the feature of the model, P, that (accurately) represents A in the target. Of course one can answer 'why P?' with 'P depends on R' *in the model*—in the case of the tidal behaviour in the model, it depends on the gravitational force in the model, and in the case of the patterns of conductance in the model, they depends on the classical orbits in the model—but it is unclear what these answers are supposed to refer to if we are talking about the behaviour of the target itself. After all, we know that Newtonian gravity isn't part of the ontological furniture of the world, and we know that electrons aren't the sorts of things that follow classical orbits. Thus, if we ask what if these features were varied, we're no longer talking about the target systems in question. So, whilst Bokulich is right that fictional models can offer first-order explanations, this observation in itself doesn't put pressure on the idea that in order to explain a certain phenomenon a model has to accurately represent it and (at least some of) its dependencies.

However, let's suppose for the time being that this is the only explanatory role that they play (in particular, let's suppose that they don't successfully answer the second-order explanatory questions). I want to highlight that even if this is the case, this observation goes relatively far in capturing the true but unspecific claim that these fictional models are, in some sense, explanatory. In particular, I want to highlight

that by playing this first-order explanatory role, the models already take on a status beyond being 'phenomenological models' or 'calculation devices', at least where these monikers are used in a pejorative sense (see Bokulich 2008a, p. 227; 2008b, p. 138 ; 2011, p. 44–45 ; 2012). There are three things to note in this regard.

First, in their aforementioned roles, the models in question are not just used to simply calculate the value of their behaviour, *A*, in some particular case. Both the Newtonian model of the tides and the quantum scarring model provide a wealth of counterfactual information between their respective behaviour (*A*: tidal behaviour, conductance patterns) and what it depends on (*B*: the positions/masses of celestial bodies, the initial conditions of the wavefunction and the shape of the quantum dots); they don't just represent particular values of *A*, they represent how these values change with values of *B*. Second, neither of these models answer the first-order explanatory questions in a way that involves them being constructed via an ad hoc fitting of the model to the empirical data (Bokulich 2011, pp. 44–45). Third, neither of these models are 'independent of theory' in any way that would justify calling them 'phenomenological' in that sense (McMullin 1968). The Newtonian model of the tides obviously draws upon Newtonian mechanics, and as Bokulich discusses extensively, the quantum dot model, and models of quantum scarring more generally, draws on a rich interplay between classical and quantum theory (and with respect to this aspect of her discussion I am in complete agreement). These considerations, I think, show that the models in question are not *just* calculation devices. They do provide explanations, albeit first-order ones. However, if this is all they do, there is the worry that:

> they do not purport to give us any *genuine insight into the way the world is*. An explanatory model, by contrast, does aim to *give genuine insight into the way the world is* (Bokulich 2011, p. 44), italics added).

Of course if what I argued above is true, they do give us genuine insight into the way some aspects of the world are (the relationship between *A* and *B*). However, one might still worry that they do not exhibit 'enough' explanatory depth; even though the model might explain *A* by representing its modal profile with respect to *B*, this may strike some as a relatively 'shallow' explanation.[21] So the next question is whether or not they give us genuine insight into another aspect of the way the world is (*why A* and *B* exhibit the dependency they do). And this brings us to the question of whether the fictional models have the explanatory depth to answer second-order explanatory questions.

[21] Obviously the request for explanatory depth cannot be pushed indefinitely: at some point one has to accept that a model cannot answer every explanatory question about a target. But nevertheless, it does seem reasonable to ask whether the models in question provide more than the first-order explanations presented here.

4.2 Fictions and second-order explanations

Recall the second-order questions relevant to the explanations offered by the fictional models discussed in this paper:

Q1: Why do the tides depend on the positions/masses of the celestial bodies?
A1: Because the celestial bodies exert a certain classical force on one another.
Q2: Why do the patterns of quantum scarring depend on the initial conditions of the wave function, the shape of the quantum dot, and so on?
A2: Because electrons follow classical trajectories.

The way that the models answer Q1 and Q2 respectively is by invoking a reason why the dependencies in the target system behave the way they do. The answers invoke a basis for the counterfactual dependencies. These play the role of *R* (which is supposed to accurately represent *C*) in the **Second-Order Model Explanation Condition**. The problem is that in the case of fictional models, *R* (seems to) inaccurately represent the actual basis of the counterfactual dependencies in the model: the proposed explanans isn't present in the target.[22] It's not gravitational force that determines why the target tides depend on the positions/masses of the target celestial bodies, it's rather to do with the curvature of the spacetime manifold (and we know that this is the case). Of course there is a relevant counterfactual dependency in the model between gravity and the relationship between the tides and the positions/masses of the target celestial bodies. There is also a relevant counterfactual dependency in the target between spacetime curvature and the relationship between the tides and the positions/masses of the target celestial bodies. But the dependency in the model is not the same as the dependency in the target (the former stems from gravity as a force, the latter from spacetime curvature). It's not classical orbits that determine why the target conductance peaks or different quantum scarring patterns arise in different set ups of quantum dots, it's interference patterns in the electrons' wavefunctions, which are spread out throughout the dot (and again, we know that this is the case, and the respective dependencies in the model and target are mismatched).

So what should we think about how the models answer the second-order *w*-questions, like 'why do the tides depend on the positions/masses of the celestial bodies'? If the answers concern what-if-gravity-had-been-different, or what-if-the-classical-orbits-had-been-different, what is the relevant feature in the world that is varying? Presumably it's not gravity, or classical orbits, since we know that the former isn't an ontological feature of our world, and we know that electrons aren't the sorts of things which follow classical orbits. Of course we can vary the values of gravity, or classical orbits, *in the model*. But when we're asking *w*-questions, we're not (primarily) interested in what if things had been different in the model, we're interested in what if things had been different in the target. In fact, it's *precisely* the fact that these features (those I have labelled *R*) in the models in question seem to misrepresent

[22] Or alternatively the explanans given is inaccurate. Which reading one prefers here depends on whether one prefers to adopt an ontic or epistemic account of explanation, i.e. whether one prefers to think of explanans as things in the world, or our representations of such things. This is relevant to how to phrase the puzzle I'm discussing here, in terms non-existent explanans or inaccurate explanans, but beyond this it doesn't matter too much for my current purposes. See Illari (2013) for a useful discussion.

what's actually going on (features I have labelled *C*, in the cases in question these are spacetime curvature and wavefunction interactions) that makes the models fictional in the first place (recall the discussion in Sect. 2).

So, under this reading there is no explanans to vary, and thus these models don't successfully answer second-order explanations of the relevant counterfactual dependencies, because they *misrepresent* those dependencies. We can ask what-if-*R*-had-been-different questions about the model, but in order to generate the second-order explanation, we need answers to what-if-*C*-had been different. And if *R* is what makes the model a fiction, in the sense of misrepresenting the ontology of *C*, it is not obvious how these latter questions can be answered, and thus no second-order explanation is provided (I take it that Bokulich agrees with this, since because 'gravity' and 'classical orbits' aren't features of the targets, they do not enter into counterfactual dependencies with other features of the targets, tidal range and the behaviour of the dot, which is, by her own lights, what is required for explanation). So, and this is the first horn of the dilemma for the defender of the idea that fictional models explain qua fictions: they don't provide second-order explanations precisely because they are fictional in the relevant sense (and thus the only explanations that they do provide, the first-order ones, are not misrepresentations).

However, an account of explanation according to which such models fail to provide second-order explanations is fairly conservative. One might argue that the Newtonian model doesn't just explain the tidal range, it also explains why the tidal range depends on the positions/masses of the relevant celestial bodies, despite, or even in virtue of, the fact that a purported explanation offered by the model invokes forces that we know don't exist. One might claim that the model of quantum dots doesn't just explain the conductance patterns, it also explains why those patterns depends on the shape of the dot and the initial conditions of the set up, despite, or even in virtue of, the fact that this explanation invokes the idea that electrons travel in classical orbits. One might demand a more liberal account of explanation in order to allow for the pre-theoretical intuition that such models don't just offer first-order explanations by accurately representing counterfactual dependencies of the target system, they also provide second-order explanations of the counterfactual dependencies themselves even though they misrepresent the ontological basis that gives rise to them.[23] I admit, when I'm feeling particularly open minded about explanation I feel the pull of this in certain cases.

Whilst Bokulich is not explicit about the precise explanatory questions that the models she investigates can successfully answer, I think when pressed she would agree that they can provide successful second-order explanations, since it's precisely here where their fictional nature seems to lead to the philosophical question concerning how *fictional* (in the sense of misrepresenting) models can explain.[24] As we have seen,

[23] It's not clear to me that this pre-theoretical intuition should be granted, especially since it can be explained away by the idea that it is motivated by a confusion between first-order and second-order explanations, i.e. because such models provide first-order explanations we expect them to provide second-order explanations too. But why expect that?

[24] Although this may depend on the details of the case. For example, in her (2011, p. 44). She writes 'Bohr's model does genuinely explain the Balmer series, though the explanation it offers may not be as deep as that offered by modern quantum mechanics, and moreover, the explanation offered by modern (nonrelativistic)

whilst the fictional features of the models in question play an essential, but instrumental, role in generating the relevant first-order explanations, they do not feature in those explanations themselves, which only make reference to the modal profile of the explanans and explanandum, which are accurately represented by the models in question. So if the presence of fictional models in science is supposed to challenge the idea that explanation requires accurate representation, then we should turn to whether or not the fictional elements themselves play a direct explanatory role.[25] The next step then is to investigate how these models can play such a role, which requires investigating the relationship between their fictional aspects (*R* features), and the actual ontology of their targets (*C* features). It's here where Bokulich faces the other horn of the dilemma: these models can provide second-order explanations, but to do so involves reinterpreting them as not being fictional; interpreting them in such a way that they don't radically misrepresent the ontological basis that generates the relevant counterfactual dependencies, or so I will argue. This undercuts the philosophically novel aspects of her discussions in the sense that it blunts the threat that explanatory fictional models might have on the idea that accurate representation is necessary for explanation.

Recall the distinction between fictional and non-fictional idealisations discussed in Sect. 2. Non-fictional idealisations were taken to get the ontology of their targets right, but to distort that ontology in a way that didn't involve representing the target as having some feature known not to be present in the target. Fictional models misrepresented the very ontology present in the target system. The puzzle arises if fictional models are taken to provide second-order explanations whilst misrepresenting what it is that actually underpins the relevant dependencies. In the case of the tidal model, there are Newtonian forces *in the model* which determine the counterfactual dependencies of the tides (in the model) on the celestial positions/masses (in the model). And it was assumed that the model thereby represented counterfactual dependency of tidal behaviour *in the target* on the celestial positions/masses (in the target) as being determined by Newtonian forces. Thus, the model is fictional and yet still provides what might feel like a successful second-order explanation despite not meeting the **Second-Order Model Explanation Condition** (in virtue of misrepresenting the counterfactuals' ontological basis). But, just because those forces are present in the model, it needn't be the case that the model *represents* those forces as being present in the target. Just because a model has some feature doesn't mean it represents its target as having that feature, any more than statue being made of bronze represents its subject as being made of bronze.

This brings us to the topic of how models represent. It's commonplace in the literature on scientific representation that there is a central role for model users in interpreting which features of their models play a representational role, and indeed what proposed target features those model-features are supposed to represent (for reviews of this literature see Frigg and Nguyen 2016b, 2017, 2020). According to Frigg and Nguyen

Footnote 24 continued
quantum mechanics may not be as deep as that offered by quantum field theory', suggesting that she grants that, in at least some cases, fictional models may not provide deep, i,e. second-order, explanations.

[25] An alternative possibility, which I discuss in the conclusion, is that these fictional elements feature in some alternative first-order explanations. However, as I argue there, for them to do so, they still have to be interpreted in such a way as to rob them of their fictional status.

(2016a, 2018) models come with 'keys' that specify which features of a model are associated with which features the model exports to the target system. According to Suárez (2004, 2015), models represent their targets only if they allow competent and informed agents to draw specific inferences about their targets.[26] According to Hughes (1997), models are such that results obtained by 'demonstrations' on the model can be 'interpreted' in terms of their targets. What all of these accounts have in common is that the inferences that a user draws about the target can, but need not, be of the form 'the target has similar features to the model'; that is, they needn't be of the form 'if a model has a relevant feature P, the target has P, or a feature similar to P.[27] The 'key' that connects model-features with features to be exported to their targets, or the inferences/interpretations made by competent and well informed agents, may allow for a mismatch between model-features and features that the model represents the target as having, whilst still allowing for accurate representation of the latter. A caricature of a person with a large nose doesn't have to represent its subject as having a large nose. In the appropriate context the large nose plays the role of representing the target as a liar, or as having a nosey character (Elgin 2017, Chapter 12). In these cases the key takes a feature of the representation and transforms it into another to be exported to the target. Or alternatively, the competent and well informed agent knows to interpret the the feature of the caricature—having a large nose—as representing a different feature of the target, namely that s/he is a liar, or is overly interested in other peoples' business. Either way, the feature of the model isn't carried over to the target directly, there is a mismatch between the model-features and the feature the model represents the target as having.

One way of putting this is that the representational content of a model is a function of both the model-features *and* the key or inferential rules used to interpret it. With this in mind, just because Newtonian gravity is a feature of the model this doesn't mean that it is part of the representational content of the model. The key or rule that is used to interpret the model, when applied to the features of the model involving fictional forces, can translate them to entirely different features to be exported to the target. And the representational content of the model will involve the claim that the target has the exported features rather than the model-features, just as the representational content of the caricature is that the person is a liar or has a nosey character, rather than having a large nose.

Now, earlier I said that a model is a fiction if it misrepresents the ontology of the target in such a way as to represent the target as having features which we know are not there. But as applied to the Newtonian model of tides, that relies on the idea that

[26] Here I'm assuming that the competent and well informed agents use some sort of inferential rule to draw these inferences. I take it that these rules correspond to what Suárez (2010) describes as the 'means' of representation. In keeping with Suárez's 'deflationary' way of thinking, I'm not assuming that these rules, or means, are the same across all instances of scientific representation, just that in the relevant cases in question, they allow competent model users to infer truths about their targets from features of the models that don't match features of their targets. For example, just because, in the Newtonian model, the dependency of tidal behaviour on celestial positions depends on gravity, a competent and well informed agent needn't use the model to infer that in the world the dependency of tidal behaviour on celestial positions depends on a classical gravitational force, they can infer that the former dependency itself depends on something else (i.e., to preempt what's to come, spacetime curvature).

[27] See in particular Frigg (2006) and Suárez (2003) for arguments against thinking about representation in terms of similarity.

the model represents the counterfactual dependence of tidal variance on celestial positions/masses in the target as being determined by gravitational forces. But the model doesn't have to be, and in fact, I argue, typically isn't (these days), interpreted that way. One way of interpreting the model is to explicitly *not* export gravitational forces to the target, but rather to export something like 'being determined by spacetime curvature in such a way that the effect will approximate the effect of classical gravitational forces'. As far as I can see, even once we accept that the gravitational model offers a first-order explanation in terms representing the tides' counterfactual dependence on the positions/masses of the sun, earth, and moon, the intuition that the gravitational model additionally provides a second-order explanation of *why* tidal variance depends on these positions/masses, is not because it represents the actual tides as being determined by Newtonian gravitational forces, but rather because our interpretation of the model is embedded in a broader theoretical framework where we know that at the appropriate speeds and masses any actual system will by approximated by a Newtonian system. So, suitably interpreted, the model doesn't represent the relevant tidal counterfactual relationships as being determined by Newtonian gravity, but represents them as being determined by something which is approximated by Newtonian gravity in the appropriate regimes.[28]

The same applies to the quantum dot model. Rather than exporting 'the electrons follow classical orbits', we export 'the electrons' quantum behaviour is approximated by those classical orbits'. Again, if we want the quantum dot model to provide a second-order explanation, it is not because it represents the actual electrons as classical objects, but rather because our interpretation of the model is embedded in a broader theoretical framework relating quantum and classical physics in a semiclassical way. In fact, in both cases, it's because the models are embedded in their respective theoretical frameworks (i.e., we know how to connect the, if literally interpreted then inaccurate, models, with theories that we take to be accurate), that the models can be seen to provide second-order explanations. And what these connections allow for is the possibility of providing the interpretations that explicitly do not export the (inaccurate) ontological bases for the counterfactual behaviour in question, but rather export the actual ontology of the system with the proviso that it is approximated by the model's one in the appropriate regimes.[29]

In a sense then, my argument here is in agreement with Suárez's (2009c) treatment of 'fictive' models.[30] As noted at the beginning of this subsection, according to his account of representation, models have to licence inferences about their targets. And in

[28] To further motivate this claim, it might be useful to think of a world where we hadn't discovered general relativity, but we nevertheless knew that gravitational forces didn't exist in the Newtonian sense. In such a world I take it that the key sketched here wouldn't be applied to the model. Therefore, I think that in such a world the model *wouldn't* explain the second-order question of why tidal variance depends on the positions/masses of the earth, sun, and the moon (although I still think it would explain tidal variance itself). Turning this point on its head, we can consider the impact of the discovery of inter-theory relations between relativistic and Newtonian theories. My claim here is that such discoveries can contribute to the explanatory power of Newtonian models, in virtue of impacting how they are interpreted.

[29] At this point one could ask yet another explanatory question: why does the actual ontology approximate the ontology of the model? Here I take it the answer would be provided by the connections between the theoretical frameworks in which they are embedded.

[30] I'm grateful to an anonymous reviewer for encouraging me to be explicit about this.

the case of fictive models, these inferences can have true conclusions, even though the models and their targets are not similar, or isomorphic, to one another in the relevant sense. However, in such cases, the sense in which the models should be considered fictive is no longer obvious. Even though the descriptions used to present those models, if evaluated with respect to the target system rather than the model, are false (or are such that their truth-value is irrelevant to their function), it doesn't follow that the 'fictive' models themselves should be considered misrepresentations of the relevant aspects of their targets. In the context of the indirect view of modelling, these descriptions should not be evaluated with respect to the target system; they serve to describe the model, and the question is how the model represents (recall the discussion in Sect. 2). And in fact, on the natural way of understanding these models, interpreted in the manner I have discussed, they are no longer 'fictive' in the sense that they no longer misrepresent the features of the targets they licence inferences about, since the inferences they end up licensing are, in the relevant cases, true.

Returning back to Bokulich, so far I have been suppressing a crucial aspect of her account of model explanation, a condition that is highly relevant to the question of the ways in which fictional models explain. She has another condition on model explanation; a *justificatory step* that specifies 'what the domain of applicability of the model is, and show[s] that the phenomenon in the real world to be explained falls within that domain' (Bokulich 2008b, p. 226). In the current context I'm going to interpret this step as a justification for how (some) fictional models can provide second-order explanations (the step could also reasonably be interpreted in terms of providing first-order explanations but I have already discussed that above). In the case of fictional model explanations she claims that this justificatory step is performed in a 'top-down' manner; the relevant theories involved (in the cases in question these theories include classical mechanics and general relativity, and our understanding of the relationship between the two, and classical mechanics and quantum mechanics, and our understanding of the relationship between the two) specify that the models in question are applicable to the target systems in question (Bokulich 2008a, p. 239; 2008b, p. 146; 2011, p. 39; 2012, p. 736; 2016, pp. 273–274). Since we're in the business of talking about application in terms of second-order explanations, it seems like the best way of interpreting how they do this is to accept that they provide information about how variances in R correspond to variances in the actual ontology (C) of the system, thereby allowing us to translate between what-if-R-had-been-different-questions-and-answers and what-if-C-had-been-different-questions-and-answers, where R is the ontological basis of the counterfactuals in the model and C is the ontological basis of the counterfactuals in the target.

This point, that fictional models can be interpreted via a key, or inferential rule, that translates their ontological basis into a different ontological basis to be exported to the target, is made in Bokulich's own discussion of how the quantum dot model explains:

> the theory of semiclassical mechanics provides physicists with what we might call a well-defined translation key, whereby statements about classical trajectories can be translated into true conclusions about the actual morphology of the wave function of the quantum dot. Note that the translation key given by semiclassical mechanics [...] is not from the empirical predictions generated by the

> fictions to the empirical predictions generated by the true description [...] Rather the translation key is from statements about the fictions to statements about the underlying structures or causes of the explanandum phenomenon (Bokulich 2012, p. 735).

But once she accepts this, then in combination with the aforementioned way of thinking about scientific representation, she undercuts her own claim that fictional models explain *qua* fictions. And this pulls the rug from under her project of developing an account of scientific explanation that drops the idea that accurate representation is necessary for explanation.[31] Once the fictional models are interpreted in this way, they are no longer fictions in the sense of misrepresenting the ontology of their targets. These translation keys or inferential rules specify the representational content of the models. And if one thinks that the models can generate successful second-order explanations, then a key or rule needs to be in place according to which the models don't misrepresent said ontology. By paying due attention to the ways in which competent and informed model users draw inferences from models to their target, or alternatively, the keys that accompany them, either of which is in part generated by the theoretical context in which they are embedded, models that explain the basis of the causal dependencies in question need no longer be seen as fictional.[32]

This, I think, poses a dilemma for a staunch defender of the view that fictional models can explain. The dilemma concerns whether or not they can answer second-order explanatory questions. If they can't, I think this is precisely because they are fictional in the specific respect at issue. If they can, the details of such an explanation require reinterpreting them so as to ensure that they don't misrepresent the ontological basis for the counterfactual dependence. Either way, once the explanatory question is suitably specified, there is no philosophical puzzle concerning how a fictional misrepresentation of some target feature can explain that specific feature.

5 Conclusion

Before concluding then, it is worth revisiting the role of fictions in first-order explanations. Recall that a model provides a first-order explanation of some target behaviour

[31] There are other philosophical projects where the same reasoning applies. See Frigg and Nguyen (2019) for a discussion of how the possibility of interpreting models in a non-literal manner blunts the need for an epistemology of science based on the idea that models provide 'felicitous falsehoods', rather than truths (*cf.* Elgin 2017). Saatsi (2019) provides a useful discussion of related issues in the context of thinking about scientific realism.

[32] It's worth briefly mentioning how this aspect of my discussion relates to Schindler's (2014) criticism of Bokulich's account. In the terminology of this paper, he argues that because the 'top-level' theories allow us to generate the actual ontological counterfactuals involving C from the model's counterfactuals involving R, Bokulich requires an additional argument 'for why it is the model, rather than quantum mechanics [or in the case of the tides, presumably general relatively], which does the real explanatory work' (Schindler 2014, p. 1747). Here I agree with Bokulich that fictional models can still play a crucial explanatory role; just because the models require interpretation via the theories in question, this doesn't mean that the theories themselves, without the fictional models, would suffice to generate the explanations offered by the models even when the latter are interpreted via an appeal to the theories. See Bokulich (2008a, pp. 232–233; 2008b, Chapter 6) for argument to this effect.

A only if it accurately represents *A*'s modal profile with respect to some explanans *B*. In the discussion in Sect. 4.1, I argued that in the explanations in question, those explanans were things like the positions/masses of the celestial bodies and the initial conditions of the quantum dot, features that the models also accurately represent (when interpreted literally). However, it's plausible that the features that were the explanans in the second-order explanations discussed in the previous section (e.g. spacetime curvature, which approximates Newtonian gravity in the appropriate regime), can also feature in first-order explanations as well.[33]

One could think about this in two ways. First, one could argue that without invoking 'gravity' or 'classical orbits', the first-order explanations discussed in Sect. 4.1 are unsatisfactory: yes tidal range depends on the positions/masses of celestial bodies *but only because* the positions/masses make a difference to gravitational attraction, and thus gravity cannot be excluded from those first-order explanations. This is suggested, for example, by Bokulich herself who states that physical oceanographers 'are interested in how gravity interacts with other factors to produce the complex tidal phenomena that they are trying to explain and understand' (Bokulich 2016, p. 273). Without understanding how the positions/masses of the celestial bodies interact with gravity, they do not provide a sufficient first-order explanation. Second, and relatedly, one could argue that 'gravity' and 'classical orbits' themselves provide us with first-order explanations. After all, we might also be able to use the model to answer questions like 'how would changes to "gravity" affect the tidal positions?', and thus also invoke 'gravity' as an alternative first-order explanation to the tidal range.

With respect to the first claim, I take it that whether or not the positions/masses of the celestial bodies suffices for a first-order explanation depends on the level of explanatory depth required by the context. If the first-order explanation isn't sufficient, and one demands to know why the tides depends on the celestial positions/masses, then one is, in effect, demanding a second-order explanation. But notice that this is not to say that in contexts where the first-order explanation is sufficient, that the fictions don't appear at all. As I have discussed throughout, the fictional aspects of the model are still *essential* in the generation of the explanation, because they play an essential role in structuring the counterfactual dependencies in the model, they just don't feature in the content of the explanation itself (*cf.* Lawler 2019). With respect to the second claim, that the fictional aspects of models may themselves provide alternative first-order explanations, notice that in order to make sense of the counterfactual dependency of the tidal range on 'gravity', in the sense of answering questions concerning how the tidal range (in the world) would differ were gravity to differ, we would also have to interpret 'gravity' (or more accurately, the gravitational force in the model) in a non-literal way. As previously discussed, asking how the tidal range would change were (Newtonian) gravity different, amounts to an infelicitous shift between talk of the target and talk of the model. If we are to talk solely in terms of the target, then, to be precise, we have to ask how the tidal range would change, were what in the target that approximates Newtonian gravity in the appropriate regimes to change. And again, to answer these sorts of questions the models have to be interpreted in such a way that they are not misrepresentations. So, if the fictional elements of models are taken to

[33] I'm grateful to an anonymous reviewer for pushing me on this point.

provide first-order explanations, then they can only do so by being interpreted in such a way that they are no longer fictional in the relevant respects.

To conclude. I have argued that fictional models, in the sense of models that appear to misrepresent the ontology of their target systems (rather than distorting the actual ontology), can explain certain kinds of features of their targets' behaviour. Namely, I have argued that a fictional model can explain a certain target feature by accurately representing whatever it is that the feature counterfactually depends on. For these kinds of explanations, the fictional aspects of the model play an essential role in generating the explanation, but do not feature in the explanation itself. However, I have argued that there is a second kind of explanatory question that one might also think fictional models can answer. It's in answering these sorts of questions that I think the real philosophical puzzle concerning fictional models comes to the fore, and I have suggested that either fictional models should not be taken to successfully answer such questions, precisely because they are fictional, or if one thinks that they do offer these sorts of explanations, then they do so in a way that robs them of their fictional status.

The crucial take home message of this paper then is that fictional models *qua* fictions—i.e., interpreted in such a way that makes them drastic misrepresentations of their targets—do not explain those features that they misrepresent, again precisely because they are fictional. However, if they are interpreted differently, by competent and well informed agents, using a key that allows for model-features to be associated with the actual features of the target, then they can play such an explanatory role, but this undercuts the idea that they are fictional (at least in the way that Bokulich uses the term). So the puzzle that seems to arise from models that drastically misrepresent the ontology of their targets, and yet still appear to be explanatory, dissolves. If I am right, then the existence of explanatory fictions in science gives us no reason to give up on the idea that the accurate representation of some target feature remains a necessary condition on explaining that specific feature.

Acknowledgements Extended discussion with two anonymous referees greatly improved this paper. I am also grateful to audiences at the universities of Edinburgh, Florida State, Leeds, and Birmingham for useful discussions. Thanks also to Alisa Bokulich, Corey Dethier, Roman Frigg, Ashton Green, and Sebastián Murgueitio Ramírez for helpful comments on earlier drafts. The Jacobsen Fellowship at the University of London allowed me to write this paper, and I am also grateful to the British Academy, for a Rising Star Engagement Award, and the Jeffrey Rubinoff Sculpture Park, for further support

References

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

Bokulich, A. (2008a). Can classical structures explain quantum phenomena? *The British Journal for the Philosophy of Science*, *59*(2), 217–235.
Bokulich, A. (2008b). *Reexamining the quantum-classical relation: Beyond reductionism and pluralism*. Cambridge: Cambridge University Press.
Bokulich, A. (2009). Explanatory fictions. In M. Suárez (Ed.), *Fictions in science. Philosophical essays on modelling and idealization* (pp. 91–109). London: Routledge.
Bokulich, A. (2011). How scientific models can explain. *Synthese*, *180*(1), 33–45.
Bokulich, A. (2012). Distinguishing explanatory from nonexplanatory fictions. *Philosophy of Science*, *79*(5), 725–737.
Bokulich, A. (2016). Fiction as a vehicle for truth: Moving beyond the ontic conception. *The Monist*, *99*(3), 260–279.
Bokulich, A. (2018a). Representing and explaining: The eikonic conception of scientific explanation. *Philosophy of Science*, *85*, 793–805.
Bokulich, A. (2018b). Searching for noncausal explanations in a sea of causes, forthcoming. In A. Reutlinger & J. Saatsi (Eds.), *Explanation beyond causation philosophical perspectives on non-causal explanations* (Chapter 7). Oxford: Oxford University Press.
Elgin, C. (2017). *True enough*. Cambridge, MA: MIT Press.
Fang, W. (2019). An inferential account of model explanation. *Philosophia*, *47*(1), 99–116.
Frigg, R. (2006). Scientific representation and the semantic view of theories. *Theoria*, *55*(1), 49–65.
Frigg, R. (2010). Models and fiction. *Synthese*, *172*(2), 251–268.
Frigg, R., & Nguyen, J. (2016a). The fiction view of models reloaded. *The Monist*, *99*(3), 225–242.
Frigg, R., & Nguyen, J. (2016b). Scientific representation. In N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Metaphysics Research Lab, Stanford University. **(winter 2016 edition)**.
Frigg, R., & Nguyen, J. (2017). Models and representation. In L. Magnani & T. Bertolotti (Eds.), *Springer Handbook of model-based science* (pp. 49–102). Berlin: Springer.
Frigg, R., & Nguyen, J. (2018). The turn of the valve: Representing with material models. *European Journal for Philosophy of Science*, *8*(2), 205–224.
Frigg, R., & Nguyen, J. (2019). Mirrors without warnings. *Synthese*. **(forthcoming)**.
Frigg, R., & Nguyen, J. (2020). *Modelling nature: An opinionated introduction to scientific representation*. Cham: Springer.
Giere, R. N. (1988). *Explaining science: A cognitive approach*. Chicago: Chicago University Press.
Giere, R. N. (1994). *Viewing science*. In PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association (Vol. 1994, No. 2, pp. 3–16).
Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, *21*(5), 725–740.
Grüne-Yanoff, T. (2009). Learning from minimal economic models. *Erkenntnis*, *70*(1), 81–99.
Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science*, *64*, S325–S336.
Illari, P. (2013). Mechanistic explanation: Integrating the ontic and epistemic. *Erkenntnis*, *78*(2), 237–255.
Jebeile, J., & Kennedy, A. G. (2015). Explaining with models: The role of idealizations. *International Studies in the Philosophy of Science*, *29*(4), 383–392.
Jones, M. R. (2005). Idealization and abstraction: A framework. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *86*(1), 173–218.
Kennedy, A. G. (2012). A non representationalist view of model explanation. *Studies in History and Philosophy of Science*, *43*(2), 326–332.
Lawler, I. (2019). Scientific understanding and felicitous legitimate falsehoods. *Synthese*. **(forthcoming)**.
Levy, A. (2012). Models, fictions, and realism: Two packages. *Philosophy of Science*, *79*(5), 738–748.
Levy, A. (2015). Modeling without models. *Philosophical Studies*, *152*(3), 781–798.
Luczak, J. (2017). Talk about toy models. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *57*, 1–7.
Mäki, U. (2011). Models and the locus of their truth. *Synthese*, *180*(1), 47–63.
McMullin, E. (1968). What do physical models tell us? In B. V. Rootselaar & J. Staal (Eds.), *Logic, methodology and philosophy of science III, studies in logic and the foundations of mathematics* (52nd ed., pp. 385–396). Amsterdam: Elsevier.
McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A*, *16*(3), 247–273.
Narimanov, E. E., Baranger, H. U., Cerruti, N. R., & Tomsovic, S. (2001). Semiclassical theory of coulomb blockade peak heights in chaotic quantum dots. *Physical Review B*, *64*, 235–329.

Nguyen, J. (2020). It's not a game: Accurate representation with toy models. *The British Journal for the Philosophy of Science*, *71*(3), 1013–1041.
Reutlinger, A., & Saatsi, J. (Eds.). (2018). *Explanation beyond causation: Philosophical perspectives on non-causal explanations*. Oxford: Oxford University Press.
Saatsi, J. (2019). Realism and explanatory perspectives. In M. Massimi & C. D. McCoy (Eds.), *Understanding perspectivism: Scientific challenges and methodological prospects* (pp. 65–84). New York: Routledge.
Saatsi, J., & Pexton, M. (2013). Reassessing Woodward's account of explanation: Regularities, counterfactuals, and noncausal explanations. *Philosophy of Science*, *80*(5), 613–624.
Schindler, S. (2014). Explanatory fictions-for real? *Synthese*, *191*(8), 1741–1755.
Skow, B. (2016). *Reasons why*. Oxford: Oxford University Press.
Skow, B. (2017). Levels of reasons and causal explanation. *Philosophy of Science*, *84*(5), 905–915.
Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.
Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science*, *17*(3), 225–244.
Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science*, *71*(Supplement), 767–779.
Suárez, M. (Ed.). (2009a). *Fictions in science. Philosophical essays on modelling and idealization*. London: Routledge.
Suárez, M. (2009b). Fictions in scientific practice. In M. Suárez (Ed.), *Fictions in science. Philosophical essays on modelling and idealization* (pp. 3–18). London: Routledge.
Suárez, M. (2009c). Scientific fictions as rules of inference. In M. Suárez (Ed.), *Fictions in science. Philosophical essays on modelling and idealization* (pp. 158–178). London: Routledge.
Suárez, M. (2010). Scientific representation. *Philosophy Compass*, *5*(1), 91–101.
Suárez, M. (2015). Deflationary representation, inference, and practice. *Studies in History and Philosophy of Science*, *49*, 36–47.
Suppe, F. (1989). *The semantic conception of theories and scientific realism*. Urbana: University of Illinois Press.
Suppes, P. (1960/1969). A comparison of the meaning and uses of models in mathematics and the empirical sciences. In P. Suppes (Ed.), *Studies in the methodology and foundations of science: Selected papers from 1951 to 1969* (pp. 10–23). Dordrecht: Reidel.
Teller, P. (2001). Twilight of the perfect model model. *Erkenntnis*, *55*(3), 393–415.
Thomasson, A. L. (2020). If models were fictions, then what would they be? In P. Godfrey-Smith & A. Levy (Eds.), *The scientific imagination* (Chapter 2). Oxford: Oxford University Press.
Toon, A. (2010a). Models as make-believe. In R. Frigg & M. Hunter (Eds.), *Beyond mimesis and convention: Representation in art and science* (pp. 71–96). Berlin: Springer.
Toon, A. (2010b). The ontology of theoretical modelling: Models as make-believe. *Synthese*, *172*(2), 301–315.
Toon, A. (2011). Playing with molecules. *Studies in History and Philosophy of Science*, *42*, 580–589.
Toon, A. (2012). *Models as make-believe. Imagination, fiction and scientific representation*. Basingstoke: Palgrave Macmillan.
van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.
van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.
Weisberg, M. (2007a). Three kinds of idealization. *Journal of Philosophy*, *104*(12), 639–659.
Weisberg, M. (2007b). Who is a modeler? *The British Journal for the Philosophy of Science*, *58*(2), 207–233.
Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford: Oxford University Press.
Woods, J. (Ed.). (2010). *Fictions and models: New essays*. Munich: Philosophia Verlag.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.