Explanatory fictions—for real?

Abstract

In this article I assess Alisa Bokulich's idea that explanatory model fictions can be genuinely explanatory. I draw attention to a tension in her account between the claim that model fictions are explanatorily autonomous, and the demand that model fictions be justified in order for them to be genuinely explanatory. I also explore the consequences that arise from Bokulich's use of Woodward's account of counterfactual explanation and her abandonment of Woodward's notion of an intervention. As it stands, Bokulich's account must be deemed unworkable.

1 Introduction

Traditionally it has been assumed that only theories that are (approximately) true can provide genuine explanations. Aristotelian crystalline spheres or Cartesian vortices, for instance, do not explain the motion of the heavenly bodies, because they are simply too far from the truth. In a recent book and several papers Alisa Bokulich (2008a, 2008b, 2011, 2012) has departed from this tradition and argued that at least some fictitious entities can do genuine explanatory work. She goes even further than this: some models employing fictions can be even more explanatory than theories that we consider to be approximately true.

In seeking to come to terms with explanatory fictions in science, Bokulich heavily borrows from James Woodward's counterfactual account of causal explanation which has enjoyed great popularity in recent years. But, since Woodward's account is an account of *causal* explanation, it seems Bokulich cannot adopt it wholesale. In order to free herself of any causal commitments, she abandons one of the central notions of Woodward's account: the notion of an intervention.

The purpose of this paper is it to critically assess Bokulich's account of explanatory fictions. In Section 2 I present Bokulich's account. In Section 3 I point to a tension in Bokulich's account which appears to be hard to ease. In Section 4 I assess Bokulich's attempt to apply Woodward's counterfactual account of explanation to model fictions. I conclude this paper in Section 5.

2 Bokulich's account of explanatory fictions

Bokulich presents her account of model explanations in terms of the following three tenets:

1. the explanans must make reference to a scientific model,

2. 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, and
3. there must be a 'justificatory step' that shows that the model in question is a 'good' (explanatory) model.

These three tenets are supposed to cover any kind of model explanation. The kinds of model explanations in which Bokulich is particularly interested—and arguably has received most attention for—are model explanations involving fictions (see next section). Bokulich provides no specific account for explanatory fictions. However, an account can be reconstructed from her writings:

1*. There is a set of scientific models M (forming a subset of all explanatory scientific models). Models of M explain by virtue of counterfactual dependencies between relevant features of the explanandum phenomenon and structures represented in models of M, whereby these structures are fictions;
2*. Models of M are explanatorily autonomous: they sometimes provide insight into phenomena where true theories don't;
3*. Models of M are licensed as being *genuinely* explanatory by the so-called 'justificatory step'.

Note that (1*) and (3*) capture the essence of (2) and (3), respectively. Additionally they are more specific about explanatory model fictions than (1) and (3). This additional detail will be justified in the following subsections. Next, (1) appears to be logically weaker than (1*). Whereas the latter states that the scientific model does the explaining, (1) only states that the explanans must make reference to a scientific model. But this difference is inessential. In (2), it is clearly stated that it is the model that does the explaining (and not something else merely referring to the model), therefore clearly rendering the model the explanans. Further, all the explanations Bokulich discusses are explanations where the model does the explaining and not something else (merely referring to the model).

Lastly note that the most obvious difference between Bokulich's tenets and the tenets more specific to model-fictions is tenet (2*): the *explanatory autonomy* of model fictions. Much of the appeal and novelty of Bokulich's account, I believe, derives from this tenet that she nowhere spells out that explicitly. If made explicit however, I will argue in the following, a tension arises in Bokulich's account that cannot be resolved without giving up one of the other tenets (Section 3). Further, I will argue that tenet (1*) runs afoul of several desiderata of scientific explanation (Section 4). Let us however first put some flesh on the bones of the above tenets.

2.1 What's a model fiction?

Bokulich does not provide any explicit definition of what she considers to be a model fiction; she merely illustrates with an example. The Bohr model postulates that electrons orbit the

atomic nucleus on periodic trajectories. If, for example, the trajectories represent a series of points occupied successively and continuously by electrons, then they are false representations according to quantum mechanics. An electron on such a path would everywhere have a simultaneously well-defined position and momentum. But Heisenberg's uncertainty principle rules this out. So it's not that classical orbits somehow approximate or idealize the real electron behavior in the atom. They are just plain wrong. Still, Bokulich believes that they provide genuine explanations of quantum phenomena.

2.2 Counterfactual dependence

The core of Bokulich's account of explanatory fictions consists of a modified version of Woodward (2003)'s counterfactual account of explanation. According to Bokulich

> [T]he [false] model explains the explanandum by showing how there is a pattern of counterfactual dependence of the relevant features of the target system [viz. the explanandum] on the structures represented in the model (2008a, 226).

We gain understanding of the phenomena through explanatory fictions, just as in Woodward's account (2003, 11), by being able to answer *what-if-things-had-been-different*-questions (for short, w-questions). The explanatory fictions thus tell us "how the target system would behave, if the structures represented in the model [i.e., the explanatory fictions] were changed in various ways" (ibid.). Bokulich illustrates this with several examples. Let me list those examples under the heading of the phenomenon they allegedly explain. All of the relevant explanations make reference to the model fiction of the electron orbit.

- Balmer series of the spectral lines of hydrogen:
  - "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; added emphasis).
- Morphology of "scarred" wavefunctions in descriptions of the so-called "stadium billiard" system, where the probability density of many wavefunctions is localized around rare classical periodic orbits—and hence "scarred"—rather than chaotic:
  - "[…] this pattern of dependence [between the model fiction and the phenomenon] allows one to say precisely how the quantum wavefunction morphology *would change if, for example, the classical periodic orbit had been different*" (2008a, 227; added emphasis)
- Rydberg atoms of barium, i.e. barium atoms with their outermost electron in a highly excited state, when placed in a strong magnetic field paradoxically continue to yield absorption peaks long after their ionization energy has been reached and passed. The resonances in the observed absorption spectrum constitute the explanandum:

– "There is a pattern of counterfactual dependence [between the model fiction and the phenomenon] in that *one can say precisely how the quantum absorption spectrum would have been different if the classical closed orbits had been changed*" (2008b, 147; added emphasis).

Since Woodward's account is an account of *causal* explanation and since explanatory fictions are only uncomfortably viewed as causes of the explanandum phenomena, it may seem that Bokulich cannot adopt Woodward's account of counterfactual explanation wholesale. Precisely for this reason Bokulich makes a crucial modification to Woodward's account:

> Making use of Woodward's account of causal explanation, we can see that, while there is a clear pattern of counterfactual dependence [between the explanatory fiction and the explanandum phenomenon], it does not make sense to construe dependence as a set of possible interventions or manipulations. That is, it does not make sense physically to talk about intervening in the classical trajectories to change the quantum wavefunctions. (2008a, 228)

That is, by dropping Woodward's notion of interventions from his counterfactual account of explanation, Bokulich hopes to rid the account of its restriction to causal explanatory relations. Whether or not this move does what Bokulich hopes it does and whether it is reasonable to make this move will be discussed in Section 4.

2.3 Autonomy

According to Bokulich, semiclassical explanations in quantum mechanics, i.e. quantum mechanical explanations employing false classical representations, are *not* merely "calculational tools" that ease the use of complex quantum mechanical calculations. Rather semiclassical models employing fictions offer "genuine physical insight" (2008a, 230; see also 2008b, 138f.). What is more, Bokulich holds that semiclassical explanations are *autonomous* in the sense that they provide even *more* insight than pure quantum mechanical ones. With regard to her main example of explanatory fictional classical trajectories in the quantum realm (see above), Bokulich writes

> My claim is not that there is no purely quantum mechanical explanation for this phenomenon, but rather that such an explanation that omits reference to classical periodic orbits *is deficient*. Although one can 'deduce' the phenomenon of wavefunction scarring by numerically solving the Schrödinger equation, such an explanation *fails* to provide adequate *understanding* of this phenomenon. (2008a, 230, altered emphasis)

Appealing to Hitchcock and Woodward (2003)'s account of explanatory depth, Bokulich claims that "the semiclassical model allows one to answer a *wider* range of w-questions [than quantum mechanics] about how the system would behave if certain parameters [in the model fiction] were changed" (2008a, 233; added emphasis).

2.4 Justification

Bokulich is no explanatory egalitarian. That is, she does not treat *all* model fictions as being capable of providing *genuine* scientific explanations. There are some model fictions that Bokulich wants to rule out as providing genuine explanations (such as the device of epicycles in Ptolemaic astronomy; see (Bokulich 2012)). In order to distinguish genuinely explanatory from non-explanatory fictions, Bokulich requires that genuinely explanatory fictions be justified. She calls this the "justificatory step".

> This justificatory step is intended to call explicit attention to the detailed empirical or theoretical process of demonstrating the *domain of applicability* of the model. In other words, it involves showing that the model in question is a *good model*, able to *adequately capture the relevant features of the world* (where 'relevant' is determined by which specific questions the model is trying to answer). (ibid., 228; emphasis added)

In principle, there are two ways in which a model fiction can be justified as providing an adequate explanation of the explanandum phenomenon of interest. It can proceed either "top-down" or from the "ground up". That is, either there is an "overarching theory" that specifies the phenomenological domain in which the model fiction is explanatorily adequately, or there are idealizations that relate the target to the model system "smoothly" (ibid., 226-7). Bokulich claims that the latter kind of justificatory step is most common for models in general. However, when it comes to her particular examples of alleged model fictions, Bokulich rules out bottom-up idealizations. For example:

> In the case of wavefunction scarring, this justificatory step cannot proceed via a de-idealization analysis, because the classical trajectories are not properly thought of as an 'idealization' of the quantum dynamics. In other words, one does not asymptotically recover the quantum wavefunctions by deidealizing and 'adding something back in' to the classical trajectories. Instead this justificatory step proceeds top-down, by way of Gutzwiller's periodic orbit theory, which specifies precisely how the classical trajectories can properly be used to model certain features of the quantum dynamics. (2008a, 227-8)

In this example, it is the Gutzwiller periodic orbit theory that provides a top-down justification for the model fiction being genuinely explanatory. The Gutzwiller periodic orbit theory, for Bokulich, is thus the 'overarching theory' providing justification (see fn 18 on p. 227). But how does this justification proceed precisely? The Gutzwiller theory is an instance of semiclassical mechanics which "involve[s] a thorough blending of classical and quantum ideas" (Bokulich 2008b, 121). Indeed semiclassical theories, according to Bokulich, establish "correspondence relations" between classical concepts´, such as periods and stabilities of classical periodic orbits, and quantum mechanical ones, such as average quantum density of states (ibid., 120; 2008a, 230). In her latest publication on the topic, Bokulich has accordingly referred to semiclassical mechanics (such as the Gutzwiller's periodic orbit theory) as a "well defined *translation key*, whereby statements about classical trajectories can be translated into true conclusions about [for example] the actual morphology of the wave function of the

quantum dot" (Bokulich 2012, 735). Bokulich makes it quite clear that semiclassical mechanics does not merely link fictions to the explanandum phenomena, but rather translates "statements about the fictions to statements about the *underlying structures* or *causes* of the explanandum phenomenon" (ibid.; emphasis added), i.e., quantum mechanics.

Bokulich also refers to her account of model fictions as an account of *structural* model explanation, for the counterfactual dependence between the model and the target system is a "consequence of the structural features of the theory (or theories) employed in the model" (230; added emphasis). That is, this dependence is a consequence of the structural features of semi-classical theories like the Gutzwiller theory. But again, it is not the Gutzwiller theory alone that accomplishes the justification. It is the *linking* of the model fiction to real quantum behavior through semiclassical theories such as Gutzwiller's that justifies model fictions as genuine explanatory devices. And it is of course quantum mechanics that accurately describes real quantum behavior. It is thus ultimately quantum mechanics that justifies model fictions as being genuinely explanatory. Not without reason Bokulich concludes that "realism comes in to distinguish explanatory fictions from nonexplanatory fictions" (Bokulich 2012, 735).

3 A tension in Bokulich's account

As mentioned above, there is a tension in Bokulich's account between her tenet of the explanatory autonomy of model fictions and her tenet of model fictions having to be justified in order to be *genuinely* explanatory. The tension is this: either model fictions are justified or they are not. If they are not, they provide no genuine explanations. As Bokulich (2012) puts it, "it is the 'justificatory step' that must do the heavy lifting in distinguishing explanatory from nonexplanatory fictions" (ibid., 736). But if model fictions are justified, i.e., if they are linked (in a very precise manner) to quantum mechanics through semi-classical theory (like the Gutzwiller's periodic orbit theory or Delos and Du's closed orbit theory), how can model fictions be claimed to be explanatorily autonomous? To see the problem more clearly, consider the following.

As we saw in section 2.2, Bokulich claims that there is a counterfactual dependence of the following sort

C1: Had the relevant structures in the model (e.g., shape of the electron orbit) been different, then the relevant features in the explanandum phenomenon (e.g. in the hydrogen spectrum) would have been different.

Furthermore, if it is true that, as Bokulich claims, Gutzwiller's orbit theory and Delos and Du's closed orbit theory provide a precise translation key that lets us translate between statements about model fictions and "statements about the *underlying structures* or *causes* of the explanandum phenomenon", i.e., quantum mechanics, then, presumably, it must be the case that

C2: Had the relevant quantum mechanical features (e.g. the wave function) of the relevant quantum systems (e.g. the quantum dot, the quantum stadium billiard, the quantum density of states, etc.) been different, then the relevant structures in the model fiction (such as the shape of the electron orbit) would have been different.

If that wasn't so, then it could be the case that, if some features in the explanandum phenomenon were different, features in the model fiction would be different without the relevant quantum mechanical features (e.g. the wave function) being different. But then Bokulich's claim that there is a precise translation key between the model and quantum mechanics would be false. So either it is true that model fictions (via the translation key of semiclassical mechanics) lack justification, or, counterfactuals of the C2-kind must be true.

If it is true, however, that features in the explanandum phenomenon counterfactually depend on the structures in the model fiction (by C1) and, in turn, that the model structures counterfactually depend on features in quantum mechanics (by C2), then presumably the features of the explanandum phenomenon counterfactually depend on features in quantum mechanics. Given that Bokulich, with Woodward, takes counterfactual dependence to underlie explanation she thus needs an *additional* argument for why it is the model, rather than quantum mechanics, which does the real explanatory work! Without such an argument, Bokulich's account risks collapsing into the view that model fictions are not more than convenient devices that ease calculations — a view she rejects (2008, 138-9). She makes the much more ambitious claim that explanatory fictions (i) provide genuine insight into the physical phenomenon of interest, and that they (ii) sometimes even provide insight where quantum mechanics doesn't, allowing for the answering of a *wider* range of w-questions than quantum mechanics. But once again, if model fictions need to be justified via true theories such as quantum mechanics in order to be explanatory, what are the grounds for making such ambitious claims?

Unless Bokulich can come up with an argument establishing the autonomy of model fictions whilst ensuring their justification, her only option to ease this tension appears to be to surrender at least one of the tenets of her account. That is, she could either give up on her idea that fictions are explanatorily autonomous, or she could drop her demand that explanatory fictions need to be justified via true theories. Recently Bokulich has reconfirmed her espousal of the justificatory step requirement (Bokulich 2012). But, as mentioned above, giving up on the tenet of explanatory autonomy of fictions would risk her account collapsing into a much more modest position, which she has made clear she does not want to hold.

Apart from this hitherto unacknowledged tension in Bokulich's account, there is another issue that I want to focus on in the rest of this paper.

4 Counterfactuals and interventions

As we've seen above, Bokulich makes heavy use of Woodward's counterfactual account of explanation in order to spell out the first tenet of her account of explanatory fictions. Before discussing Bokulich's use of Woodward's account in more detail let us first outline the essentials of the latter.

In Woodward's account the notion of a (non-anthropocentric) intervention, which will be discussed in more detail below, plays a central role in evaluating the truth-values of counterfactual conditionals. In Woodward's account, a causal claim (e.g. X causes Y) is true if and only if the relevant *active* counterfactual conditional is true, where active counterfactuals are counterfactuals whose antecedents are made true by interventions (Woodward 2001, 199; Woodward 2003, 145). That is, a causal claim about X and Y is true iff it is true that had one *intervened* on X, Y would have been different. This *semantic function* of interventions in Woodward's account helps to pick out several important features of causation that are fundamental for Woodward's account of causal explanation. Thus, there is a *discriminatory*, a *gauging*, and *asymmetry-individuating function* of interventions that are all part of the *semantic function* of interventions. Let us consider these sub-functions in turn.

First, the *discriminatory function* of interventions allows Woodward to distinguish between causal relations and mere correlations. Consider the famous barometer-pressure-storm example and the following counterfactual conditionals.

C3: Had the value of the barometer reading been *changed* below a certain value, a storm would have occurred.

C4: Had the atmospheric pressure been *changed* below a certain value, a storm would have occurred.

Obviously the first active counterfactual conditional is false, whereas the second one is true. The (non-anthropocentric) notion of an intervention thus allows us to discriminate causation (such as the one between atmospheric pressure and occurrence of a storm) from correlation (such as the one between the barometer reading and the occurrence of a storm). However, compare this to what Woodward calls 'passive' counterfactual:

C5: Had the value of the barometer reading been below a certain value, a storm would have occurred.

Without the conceptual tool of interventions, C5 comes out true. Counterfactual dependencies also obtain for mere correlations. Interventions are thus essential for the *discriminatory function* of interventions.

The *discriminatory function* of interventions is used by Woodward to further distinguish explanations from *apparent* explanations. Whereas an explanation only referring

to a change in the barometer reading for why a storm occurred is deficient, an explanation of the same phenomenon referring to the drop in air pressure is a good one, for only the latter refers to an actual causal relationship. Hence the *discriminatory function* also serves an explanatory purpose.

Second, the *gauging function* of interventions allows one to evaluate the range of a causal claim and the *depth* of the relevant explanatory claim. Causal relations identified in science, Woodward points out, are often 'valid' only within a certain range. For instance, the ideal gas law is correct only within a certain range of values of temperature and pressure. Outside of that range another generalization, the van-der-Waals equation, is the more accurate one. In that case, a counterfactual like the following is no longer true.

C6: Had the gas pressure had value x, the temperature of the gas would have been y, as demanded by the ideal gas law.

But in such cases the ideal gas law is no longer *explanatory* either. That is, the ideal gas law will no longer be explanatory in cases in which e.g. C6 comes out false. Again, the semantic function of interventions is put to an explanatory use.

Third, the *asymmetry-individuating function* of interventions allows us to discern the asymmetry between cause and effect, and accordingly, the asymmetry between explanans and explanandum in causal explanations. In the famous flagpole example, the height of the flagpole explains the length of the shadow cast by the flagpole, but the length of the shadow does not explain the height of the flagpole. Consider the following counterfactual:

C7: Had the length of the shadow been different, the height of the flagpole would have been different.

Woodward (2003) points out that, when read in the 'passive' mode (without the notion of intervention), C7 comes out as correct. Not so in the 'active' mode: although we can intervene to vary the height of the flagpole, we cannot intervene on the shadow so as to vary its length (without changing the height of the flagpole). The length of the shadow is not the cause for the height of the flagpole. It is for this reason that the former cannot be the explanans of the latter.

In sum, Woodward's notion of an intervention helps to pick out important causal characteristics that in turn accommodate some important explanatory desiderata. But because interventions in Woodward's account individuate causal relations, and because fictions don't cause anything in the real world, Bokulich, it seems, cannot adopt Woodward's account without significant modifications. That is why she decides to drop Woodward's notion of an intervention entirely. But that has undesirable repercussions which I will discuss in the next section.

4.1 Counterfactuals without interventions?

Firstly, it has been pointed out by Belot and Jansson (2010) that Bokulich's abandonment of the notion of intervention results in the loss of the *discriminatory function*. In particular, Belot and Jansson criticize Bokulich for being unable, without interventions, to distinguish "common cause cases" and "cases where we intuitively have a prediction but do not have an explanation" from proper explanations (83-4). Their point seems to be that proper explanations are those that exhibit causal characteristics which, in turn, we can identify by means of interventions. But this is begging the question against Bokulich: she does not believe that causation is the right way of thinking about explanatory fictions.

Secondly, it would seem that Bokulich also robs her account of the *gauging function* of interventions by dropping the notion of an intervention. After all, Bokulich now has no means for determining *under which conditions* and *in what range* of values for the relevant variables counterfactuals like C6 or C1 come out true or false.

Thirdly, perhaps the most important function Bokulich loses when eschewing Woodward's notion of an intervention is the *asymmetry-individuating function*. It appears that Bokulich, without the notion of an intervention, has no way of telling what the explanans and the explanandum is supposed to be in any given situation. Take the spectral lines of hydrogen and the explanation provided by the Bohr model. According to Bokulich there is a counterfactual dependence between that phenomenon and the model fiction of the following form.

C8: Had the structures in the model been different, the spectral lines would have been different.

Presumably this counterfactual dependence does *not* represent a causal relationship—as Bokulich herself stresses (see Section 2.2). When counterfactual dependencies do not represent causal relationships, what else is it that they can represent? In the literature on causation, there appears to be only one other class of relationships: correlations. C5 is a case in point. Crucially, given that both the occurrence of a storm and barometer reading are caused by a change in pressure, they co-vary. So if C5 is true, it will also be true that

C9: Had a storm occurred, the value of the barometer reading would have been below a certain value.

Assuming (with Bokulich) that the counterfactual dependence between the model fiction and the explanandum phenomenon does not represent a causal relation, and given that the only other kinds of relationships counterfactual dependencies are normally taken to represent are correlations, and lastly, given that counterfactuals representing correlations

(such as C5) have what one might call 'reverse counterfactuals' (such as C9), it follows that for C8 there will be a reverse counterfactual of the following form.[1]

C10: Had the spectral lines been different, the structures in the model would have been different.

It is unclear how Bokulich could rule out such a reverse counterfactual. However if she can't, she won't be able to identify the structures in the model as the explanans and the phenomena as the explanandum, rather than the other way around.

This problem can be rendered even more acute. Suppose now that the wavefunction morphology, as Bokulich also claims, counterfactually depends on the shape of the electron orbit in the Bohr model (see Section 2.2). Then, it would be true that, had the spectral lines of hydrogen been different, the shape of the electron orbit would have been different (by C10) and thus, the wavefunction morphology would have been different. Thus, we would have a counterfactual dependence of the wavefunction morphology on the spectral lines of hydrogen. And if, as Bokulich assumes with Woodward, counterfactual dependence underlies explanation, then what we get is an explanandum supposedly explaining another explanandum. That can't be right.[2]

In order to escape the problem of explanatory asymmetry, Bokulich needs to block the inference from C8 to C9. There appear to be just three options for Bokulich to achieve that. First, she could insist that the counterfactual C8 does not represent either a causal relationship or a correlative one. But it's not easy to see what kind of relationship that could be. Second, she could revise her beliefs and allow for causation by fictions after all. Third, and relatedly, she could give up on her insistence that the notion of an intervention is misplaced in the context of explanatory fictions. I shall explore the second and third option in the next section.

4.2 A way out?

There is reason to think that Bokulich might have given up on the notion of an intervention too quickly. Recall, Bokulich abandons interventions because "it does not make sense physically to talk about intervening" in fictions (2008a, 228). However, that alone cannot be a reason for eschewing Woodward's notion of an intervention, as should be clear from the following.

Woodward (2003, 128) offers a stronger and a weaker notion of physical possibility. In the former, an intervention is physically possible, if it is consistent with some set of *actual* initial conditions and the *actual* laws. In the latter, an intervention is physically possible if it is consistent with some set of *possible* initial conditions and the *actual* laws. When joined with determinism, Woodward points out, the weaker notion implies that interventions are

[1] Alternatively, one may here also appeal to the motivation for C2 provided in Section 3.
[2] I owe this extension to an anonymous referee.

impossible unless they *actually* occur. Since such a notion is counterproductive to a *counterfactual* account of causation, Woodward dismisses it. In the weaker notion, an intervention is impossible if and only if it is incompatible with the laws of nature. Since the classical laws of physics are approximately true, interventions on structures in the semiclassical models Bokulich discusses should not be ruled out prematurely. *Prima facie*, at least, physically possible interventions thus seem to be perfectly compatible with explanatory fictions. In fact, Woodward dismisses even weak physical possibility as constraint on interventions. For him, interventions merely have to be *logically* or *conceptually* possible (p. 132).[3]

Given this extremely weak notion of an intervention, there seems no *prima facie* reason why interventions should not be applicable to fictions—contra Bokulich. The crucial question now of course must be whether interventions suffice to fulfill the three functions mentioned in Section 4 in the context of fictions.

In evaluating counterfactuals, Woodward appeals to the whole paraphernalia of causal models (see Woodward 2003, 139-145). Causal models—despite their name—encode (so-called non-backtracking) counterfactual (rather than causal) information and therefore ought to be compatible, at least in principle, with counterfactuals pertaining to fictions.[4]

Consider two variables, X and Y. Suppose Y counterfactually depends on X. Then, in a causal model, Y is represented as a mathematical function (more specifically a so-called structural equation) which can be depicted as a causal graph where a "directed edge" (i.e., an arrow) is being drawn from X to Y.[5] Now, in such a model, if we were to intervene on X so as to change its value, the value of Y would change, but not vice versa. If that were the case, Woodward (amongst many others) would take this to be indication of X causing Y.

Now consider counterfactuals involving fictions such as "Had Sherlock Holmes not stopped the murderer, another person would have died" (SH), or "Had the Cartesian vortex changed its direction, the planets would revolve in the opposite direction around the sun" (CV), or finally, counterfactual C8. In the standard literature, such counterfactuals *do not have a representation in causal models*. Rather, standardly causal models represent only those counterfactuals that are considered to be *true in the actual world*.[6] And in the actual

[3] Woodward makes this step after considering various counterexamples to the weak physical possibility constraint. The details of this step are beyond the current discussion.

[4] Backtracking counterfactuals are counterfactuals where one "back-tracks" in time from an effect to its cause so that "had e been different, c would have been different". More appropriately, perhaps, one should here speak of the counterfactual "backtracking" from the explanandum to the explanans. For explicit statements that causal models encode (only non-backtracking) counterfactual information see (Woodward 2003, 43) and especially (Hitchcock 2001).

[5] Note that it may also be the case that, vice versa, X counterfactually depends on Y. That would be the back-tracking counterfactual of the counterfactual mentioned in the text. However the model is "blind" to such dependencies. It merely represents *non*-backtracking counterfactuals.

[6] That causal models, as used by Woodward, normally encode information about counterfactuals that are true *in the actual world* becomes particularly clear in his comparison of his account to Lewis's (133-45). There he

world, counterfactuals involving fictitious entities in their antecedents are not true. Now one may of course want to depart from the standard practice and incorporate counterfactuals involving fictions in a causal model. Then, in such a model, intervening on a variable X that is connected to another variable Y via a directed edge, and thereby changing Y may be interpreted as "fictional causation" of Y by X.

It is however unclear how this departure from standard practice could help Bokulich resurrect her account. There are two problems. First, making room in causal models for the representation of counterfactuals involving fictions would entail that causal models no longer give information about only *actual* causation. Although this is something that would not be welcomed by Woodward and others, it may be something that Bokulich would be happy to accept. Second, and more importantly, what Bokulich needs is the ability to distinguish counterfactual CV from C8 in order to be able to say that the latter, but not the former, is genuinely explanatory. Yet causal models do not provide Bokulich with the resources of doing so. Rather, it seems, one would have to—in an ad hoc fashion—represent only those counterfactuals involving fictions that one takes to be genuinely explanatory. But this would of course simply beg the question.

Bokulich might want to consider adopting the standard account of the semantics of counterfactuals by David Lewis's. This account, contrary Woodward's, is not restricted to actual world counterfactuals. In Lewis (1979)'s account counterfactuals are evaluated by means of a similarity metric for comparing possible worlds. In a nutshell one evaluates counterfactuals of the form "Had c occurred, e would have occurred" by comparing the actual with possible worlds in which both c and e occur. If those possible worlds are closer to our (actual) world than those worlds in which c but not e occurs, the counterfactual is to be deemed true. The "closeness" of possible to our actual world is judged on the basis of a list of similarity criteria that involves considerations of the size of the miracles that are required to make the cause occur in possible worlds,[7] and the match of spatio-temporal region between the actual and the possible world.

For there to exist a counterfactual dependence between a model fiction and the explanandum of that model, in Lewis's account, the following would have to hold. First, it would have to be the case that there exists a possible world in which the model and the explanandum are true. This is not the actual world for we know that the model is false. Call this possible world the M-world. Second, there must be another possible world where features in the model fiction are different and features in the explanandum phenomenon

---

judges the relevance of possible worlds to the evaluation of a counterfactual *on the basis of a causal model that encodes counterfactual information of the actual world.* This aspect is not to be confused with the fact that causal models represent possible worlds in terms of variable values. Overall, one may therefore say that causal models <V,E> represent possible worlds in the variables V and the actual world in the structural equations E relating these variables.

[7] Miracles in Lewis's account are analogous to Woodward's notion of an intervention. See Glynn (forthcoming) for a thorough comparison of Woodward's and Lewis' accounts.

are different. Call this the M*-world. Third, the M* world must be closer to the M-world than any other possible world in which the explanandum phenomenon is different, but not the model fiction. In that case, we could say that, in that possible word M*, the fiction "causes" the explanandum phenomenon. We would therefore have something like "fictional causation".[8] The trouble is just that we can imagine many possible worlds in which something might fictionally cause really just *any* phenomenon. For such a potential explanation to become an actual explanation what we would have to require, fourthly, is that such M*-worlds be "reasonably" close to our actual world. But are they? Again, electron orbits do not exist. If they would, the Heisenberg uncertainty relation would be false. Bokulich would therefore once more face the task of having to tell apart explanatory and non-(or only potentially) explanatory fictions.

Bokulich may insist at this point that those possible worlds are the relevant ones to ours for which a correspondence can be established between the model fiction in the M world and the relevant true theory in the actual world. However she then still faces the issue discussed in Section 3.

Something else she might be tempted to do instead is to appeal to something like the No-Miracles-Argument in establishing that M-worlds ought to be considered closer to our world than, say, worlds in which Aristotelian crystalline spheres exist. That is, one could try to argue that given the empirical success that semiclassical mechanics has (by employing the electron orbit) and given that the Aristotelian crystalline spheres did not yield much in terms of empirical success, we have good reason to think that the former is somewhat closer to the truth than the latter. However, it is well known that this argument fails in the face of the history of science: there have been a number of theories (ether theories, theories of the caloric, in particular) which had reasonably good empirical success, but which must be considered to be utterly false—just like electron orbits. In order to save the NMA, philosophers have applied various divide-et-impera moves that separate out those false posits from posits that actually were responsible for the theory's success (see e.g. Harker 2013). That move applied to semiclassical mechanics would arguably mean that it ought to be the quantum mechanical, rather than the classical, posits that are responsible for the empirical success of semiclassical mechanics. Whether or not that is the case, it cannot be decided here. Neither possible outcome can help Bokulich, however. Either it is quantum mechanics that is responsible for the empirical success of the explanatory fiction or it isn't. If it is, explanatory fictions, once more, need to be shown to be explanatorily autonomous. If it isn't, the explanatory fiction, once more, goes unjustified (because the M-world has not been shown to be close to the actual world). We thus find ourselves once more thrown back to the problem discussed in Section 3.

[8] Note that it has been debated whether Lewis's account is capable of successfully mapping the time-asymmetry of causation and therefore explanatory asymmetry (Elga 2001; Kment 2006; Wasserman 2006; Dunn 2011).

5 Conclusion

The discussion in this paper revealed two problems in Bokulich's account of explanatory fictions: a tension between two of her account's three tenets on the one hand, and the violation of several explanatory desiderata on the other hand. The tension arises between her demand that explanatory fictions need to be justified and the claim that explanatory fictions are explanatorily autonomous. But as it stands, Bokulich cannot have her cake and eat it too. Bokulich's account runs afoul of violating several explanatory desiderata because she eschews Woodward's notion of an intervention. She does so prematurely, for the latter is indeed compatible with fictions. Still the application of Woodward's account to fictions leads to problems that are irreconcilable with the idea that explanatory fictions provide actual rather than just potential explanations.

It is unclear how Bokulich's account of explanatory model fictions could be saved without abandoning one of its core components (i.e., explanatory autonomy and the 'justificatory step', in particular). What is more, it is unclear how *any* account of explanatory fictions which appeals to counterfactual dependence could be rendered workable without the provision of a radically new semantics of counterfactual dependence. Since this will be a big task, proponents of explanatorily autonomous model fictions (if there really are such things) might want to look elsewhere for alternatives.

Acknowledgements

I thank the audiences at the Causality in the Sciences (CaitS) conference in Ghent, at the Annual Meeting of the British Society for the Philosophy of Science in Bristol (both in 2011), and at the Department of Philosophy at Aarhus University for their feedback. I also thank Brian Hepburn and Franz Huber for reading earlier versions and for providing helpful comments. I'm particularly indebted to an anonymous referee of this journal for detailed and challenging remarks.

6 References

Belot, Gordon, and Lina Jansson. 2010. Alisa Bokulich, Reexamining the Quantum-Classical Relation: Beyond Reductionism and Pluralism , Cambridge University Press, Cambridge (2008) ISBN 978-0-521-85720-8 pp. x+195. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* 41 (1):81-83.

Bokulich, Alisa. 2008a. Can Classical Structures Explain Quantum Phenomena? *The British Journal for the Philosophy of Science* 59 (2):217-235.

— — —. 2008b. *Reexamining the quantum-classical relation : beyond reductionism and pluralism*. Cambridge: Cambridge University Press.

— — —. 2011. How scientific models can explain. *Synthese* 180 (1):33-45.

— — —. 2012. Distinguishing Explanatory from Nonexplanatory Fictions. *Philosophy of Science* 79 (5):725-737.
Dunn, Jeffrey. 2011. Fried eggs, thermodynamics, and the special sciences. *The British Journal for the Philosophy of Science* 62 (1):71-98.
Elga, Adam. 2001. Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science*:S313-S324.
Glynn, Luke. forthcoming. Of Miracles and Interventions. *Erkenntnis*:1-22.
Harker, David. 2013. How to Split a Theory: Defending Selective Realism and Convergence without Proximity. *The British Journal for the Philosophy of Science* 64 (1):79-106.
Hitchcock, C, and J Woodward. 2003. Explanatory generalizations, part II: Plumbing explanatory depth. *Nous* 37 (2):181-199.
Hitchcock, Christopher. 2001. The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy* 98 (6):273-299.
Kment, Boris. 2006. Counterfactuals and explanation. *Mind* 115 (458):261-310.
Lewis, D. 1979. Counterfactual dependence and time's arrow. *Nous* 13 (4):455-476.
Wasserman, Ryan. 2006. The future similarity objection revisited. *Synthese* 150 (1):57-67.
Woodward, J. 2003. *Making things happen: a theory of causal explanation*. Oxford: Oxford University Press.