1. Introduction

The last two decades or so witnessed a major shift of focus in the philosophical literature on scientific explanation, wherein various model-based approaches to explanation have been proposed (Hughes 1997; Frisch 1998; Morgan and Morrison 1999; Batterman 2002; Parker 2003; Woodward 2003; Giere 2004; Craver 2006; Godfrey-Smith 2006; Contessa 2007; Bokulich 2008, 2011, 2012; Strevens 2008; Kennedy 2012; Weisberg 2013; King 2016).[1] Among many attempts to carve out such a model-based approach to scientific explanation (or "model explanation" for short), the counterfactual account of model explanation seems to be a promising route that many philosophers have started to pursue (e.g., Frisch 1998; Woodward 2003; Bokulich 2008, 2011, 2012; Rice 2015).

Perhaps James Woodward's interventionist account of explanation is the most influential version of the counterfactual account of model explanation (Woodward 2003), though Woodward initially intends his account to be a general account of scientific explanation rather than an account of model explanation. According to Woodward, scientific explanation is associated with whether, and to what extent, a generalization can be used to answer "what-if-things-had-been-different questions" ("w-questions" hereafter) (Woodward 1997, 2000, 2001, 2003, 2010). More specifically, an explanatory generalization is one that can tell us information about how changes in variables that figure in the explanans, typically under intervention, would be systematically associated with changes in variables that figure in the explanandum, i.e., provide us with information about patterns of counterfactual dependence between variables. One may find that it is not difficult to apply Woodward's general account of scientific explanation to model explanation: an explanatory model is one that tells us information about how changes in the explanandum would systematically depend on changes in the explanans. This time, the explanandum refers to the output, pattern or phenomenon to be explained in the target system (and also reproduced by the

[1] Early versions of model-based approaches can be found in McMullin (1978, 1984, 1985).

model), and the explanans refers to the mechanisms or dynamics represented by the model (and also encoded in the target system).

Although this is an attractive extension of Woodward’s account, he himself does not fully develop it. Fortunately, Alisa Bokulich picks out this thread, following Woodward’s track to explore what a counterfactual account of model explanation would look like. On her account, a model can explain its explanandum because the model shows “how the elements of the model correctly capture the pattern of counterfactual dependence of the target system” (Bokulich 2011, 39). However, I think her account fails, for it seems to assume that a model must bear some *substantive* representational relationship to its target system so as to be explanatory (we will see in Section 2 that Bokulich seems to hold *isomorphism* to be the representational relationship between a model and its target system). As many authors have pointed out, to be explanatory a model needs not bear such a substantive relationship to its target system. For example, a minimal model may bear no substantive representational relationship to its target system but can still offer explanation (Batterman 2002a, 2002b; Batterman and Rice 2014; Rice 2015; for similar views, see Morgan and Morrison 1999; Knuuttila 2005, 2011; Kennedy 2012).[2]

To avoid the problem faced by Bokulich’s account and in the meanwhile preserve the insight derived from Woodward that model explanation has something to do with the model’s counterfactual structure, this essay suggests an alternative account of model explanation based on Suárez’s *deflationary* approach

[2] Notice that some of these views are said to be non-representational because they totally dismiss the relevance of any representational relationship between the model and the target in scientific modeling. On the other hand, my account to be developed in what follows seems to be representational because it proposes that a model represents a target in virtue of some activities performed by the modeler involved. So, it appears that my account conflicts with these non-representational views. However, I think the conflict is only superficial. First, as will become clear, the term ‘representation’ employed in my account should be deflationarily construed. Second, my account shares with these non-representational views the core idea that it is not any substantive representational relationship between the model and the target that makes the model able to do the work it is supposed to do in scientific modeling, e.g., model explanation.

to scientific representation. Suárez's approach is deflationary in that it does not postulate any substantive representational relationship (e.g., isomorphism, partial isomorphism, similarity, etc.) between the model and its target (Suárez 2004, 2015, 2016).[3] This does not mean that there cannot be any relationship between the model and the target, but rather that the fact that a model can represent a target cannot be reduced to any substantive explanatory properties of the model, or the target, or their relations (Suárez 2015, 36). On this account, that a model can represent a target is due to the way the model is used by the modeler to make *inferences* about the target—hence this account is also called the inferential account of scientific representation. Due to two reasons, I think this path is worth pursuing. First, because of its distancing itself from positing any substantive relationship between the model and the target, it directly circumvents thorny problems many substantive accounts of representation suffer (e.g., for problems faced by the isomorphism view see Downes 1992; Suárez 2003; Frigg 2006; Odenbaugh 2008; Weisberg 2013). Second, and more importantly, it goes in concert with scientific practice since it rightly points to the way of how scientific models are used by modelers to explain phenomena in practice, that is, how models are used as inferential means in the context of scientific explanation.

Integrating the counterfactual account of model explanation with the inferential approach to scientific representation, an inferential account of model explanation is suggested in this essay. According to this account, model explanation as a key part of scientific practice proceeds in a two-step manner: (i) the modeler first entertains the counterfactual structure of the model in various ways such that she can build a whole range of counterfactual statements about the model, and (ii) she then infers from the model to the target by making a range of hypothetical statements that transfer over claims derived from the model onto claims about the target.

[3] For a discussion of the isomorphism view, see Sneed (1971), Stegmüller (1976), Suppe (1977), Suppes (1962, 1967), Van Fraassen (1970, 1972); the partial isomorphism view, see Bueno (1997), Bueno et al (2002, 2012), Da Costa and French (2003), French (1997, 2003), French and Ladyman (1998); and the similarity view, see Giere (1988), Godfrey-Smith (2006), Weisberg (2013).

Note that the account to be developed in this essay falls within the broad category of counterfactual account of explanation, since it shares the core with many other counterfactual accounts: offering explanation has something to do with providing counterfactual information (Woodward 2003; Bokulich 2011; Rice 2015). However, some might doubt why the counterfactual account better suits model explanation than other accounts, such as the Deductive-Nomological account (Hempel 1965), the causal account (Salmon 1984; Woodward 2003; Strevens 2008), the non-causal account (Walsh et al. 2002; Lange 2013; Ariew et al. 2015; Rice 2015; Baron, Colyvan and Ripley, 2017), the mechanistic account (Craver 2007), etc. The first reason is that, except for the Deductive-Nomological account,[4] the counterfactual account is not a competitor against the other views but can work hand in hand with them. For example, Woodward's work features both a counterfactual and a causal characteristic (Woodward 2003), Baron, Colyvan and Ripley propose a non-causal counterfactual account (Baron, Colyvan and Ripley 2017), and Craver's mechanistic account is compatible with the counterfactual account in the way that his causal-mechanical explanations can be rephrased in counterfactual expressions without sacrificing their own explanatoriness (Craver 2007). Second, as will be shown in the following sections, a counterfactual account can best suit model explanation because it reflects the inferential and hypothetical nature of the surrogate reasoning occurring in model explanation, that is, using a model—i.e., a surrogate—to explain its target system.

To elaborate on that account, the essay goes as follows. Section 2 will first briefly outline Bokulich's counterfactual account of model explanation, followed by Section 3 describing Suárez's inferential account of scientific representation. Then, in Section 4, based on Bokulich's and Suárez's accounts, an inferential account of model explanation will be developed. To show how the inferential

[4] I think the Deductive-Nomological account is implausible because, first, its requirement for laws cannot be met in model explanation—many models do not invoke laws to explain, and second, many model explanations do not work in a deductive way but involving empirically finding explanatory (causal) variables.

account really works, an agent-based simulation model drawn from biology will be examined in Section 5.

2. Bokulich's Counterfactual Account of Model Explanation[5]

The idea that a model can be explanatory because it captures the pattern of counterfactual dependence of its target system can at least be traced back to Margaret Morrison's work on models, where she claims that "The reason models are explanatory is that in representing these systems, they exhibit certain kinds of structural dependencies" (Morrison 1999, 63; Cf. Bokulich 2008, 255). Yet, Morrison does not develop this sound idea into a philosophical account of model explanation.

Partly due to the popularity of Woodward's interventionist view of scientific explanation, as a version of the counterfactual account of explanation, the development of Morrison's idea has become possible very recently. Alisa Bokulich is one author who tries to bring Morrison's idea into a flesh-and-blood form based on Woodward's counterfactual account of scientific explanation. Bokulich's basic idea is that a model "explains the explanandum by showing how the elements of the model correctly capture the pattern of counterfactual dependence of the target system" (Bokulich 2011, 39).

For Bokulich, what makes an explanation an example of model explanation is that the explanans in question must make appeal to a scientific model (Bokulich 2008, 145). Given such, then a general account of model explanation must explain how a model can be genuinely explanatory, and must be able to demarcate models that are explanatory from those that are not. Her answer has been mentioned above, that is,

> That model explains the explanandum by showing how the elements of the model

[5] Note that Bokulich names her account as an account of *structural model explanation* (2011, 40). However, given that Woodward's counterfactual account of scientific explanation resides in the heart of her account, I view her account as a version of the counterfactual account.

> correctly capture the pattern of counterfactual dependence of the target system. More precisely, in order for a model M to explain a given phenomenon P, we require that the counterfactual structure of M be isomorphic in the relevant respects to the counterfactual structure of P. That is, the elements of the model can, in a very loose sense, be said to 'reproduce' the relevant features of the explanandum phenomenon. Furthermore, as the counterfactual condition implies, the model should also be able to give information about how the target system would behave, if the elements described in the model were changed in various ways. (Bokulich 2011, 39)

Here we see both Morrison and Woodward's influence on Bokulich. For one thing, following Morrison, Bokulich states that the explanatory power of a model has something to do with the model's structural dependences. For another, largely inspired by Woodward's account, she thinks that the way a model explains is closely associated with how the model can be used to answer various "what-if-things-had-been-different questions". Interestingly, we can also see the influence of the semantic view on Bokulich from her presentation (e.g., "we require that the counterfactual structure of M be isomorphic in the relevant respects to the counterfactual structure of P").

Finally, for Bokulich an adequate account must satisfy a further 'justificatory step', which 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" (Bokulich 2011, 39). It will turn out that the justificatory step, which somehow resembles Hughes's interpretation step (Hughes 1997), is of tremendous importance to Bokulich's account, though she does not fully develop it. I will return to this point in Section 4.

To see how Bokulich's account works, let us consider her example: Niels Bohr's model of the hydrogen atom:

> According to Bohr's model, the electron can orbit the nucleus only in a discrete series of allowed classical trajectories known as stationary states. While in a stationary state the energy of the electron is constant, and the electron can only gain or lose energy by jumping from one stationary state to another. When such a

> transition or "quantum jump" occurs, a single photon of a given frequency is emitted (or absorbed). The frequency of the photon is given by the difference in energy of the two allowed orbits. The spectrum of hydrogen […] is built up out of the photons being emitted in these jumps between stationary states, where only those frequencies (or wavelengths) corresponding to allowed quantum jumps occur, and the intensity (or brightness) of a spectral line is given by the probability of that jump occurring. (Bokulich 2011, 41)

Bokulich says her account of model explanation can cast light on why Bohr's model is genuinely explanatory. To begin with, the explanans makes appeal to an idealized model, i.e., Bohr's model, hence it is clearly an example of model explanation. Moreover, the model explains the explanandum by showing that "the counterfactual structure of the model is isomorphic (in the relevant respects) to the counterfactual structure of the phenomenon" (Bokulich 2011, 43). This means that the model is able to answer a wide range of "w-questions" about its target system. For example, the model is able to answer questions such as "how the spectrum would change if the orbits were elliptical rather than circular, or how the spectral lines would change if the hydrogen atom were placed in an external electric field" (*Ibid.*, 43), etc. Finally, there is a justificatory step, "specifying what the domain of applicability of the model is, and showing that the phenomenon in the real world to be explained falls within that domain" (*Ibid.*, 39). In particular, "modern semiclassical mechanics provides a top-down justificatory step showing that Bohr's model—despite failing as a literal description—is nonetheless a legitimate guide to quantum phenomena in certain domains" (*Ibid.*, 43).

So far so good. A model can explain because the model can *capture* the counterfactual structure of its counterpart in the target. However, it seems Bokulich ties her account of model explanation to some substantive representational relationship too closely, because—as least from her presentation—her account seems to imply that for a model to be explanatory it must first bear some relationship such as isomorphism to its target system. We will see in the following sections that bearing such a substantive relationship to a target system is not a necessary condition for an explanatory model.

3. Suárez's Inferential Account of Scientific Representation

It should be clear from the outset that Suárez is not mainly concerned with the problem of model explanation. Rather, his focus is on how to make sense of scientific representation without postulating any substantive model-world relationship. Nevertheless, it will become evident in what follows that his inferential conception of scientific representation will lend dramatic support to developing an inferential account of model explanation.

In approaching the notion of scientific representation in a deflationary spirit, Suárez characterizes it as a two-part activity involving "its essential directionality and its capacity to allow surrogate reasoning and inference" (Suárez 2004, 767). More specifically, it involves

> [T]he exercise of the inferential capacities of the model source (with respect to the target), and the setting of what I call representational force of the source towards the target. Both components are elements of practice and ensue in relations only in those contexts in which the practice's outputs include the establishment of a particular match or comparison between source and target. (Suárez 2015, 41)

The first part of the activity is called representational force, concerning how a source A, i.e., a model, points to a target system B, and the second part is called inferential capacity, concerning how A allows a modeler to draw inferences about B. Putting the two parts together, we may say that

> A represents B only if (i) the representational force of A points towards B, and (ii) A allows competent and informed agents to draw specific inferences regarding B. (Suárez 2004, 773)

Although the representational force of a source has something to do with the internal structure of the source, it "is essentially linked to a practice of interpreting features of a [source] as standing for features of a [target]” (Suárez 2015, 43). Here, the pragmatic dimension comes into play, namely, the essential role played by the agents and the purposes of modeling:

> First, the establishing and maintaining of representational force in (i) requires some agent’s intended uses to be in place; and these will be driven by pragmatic considerations. Second, the type and level of competence and information required in (ii) for an agent to draw inferences regarding *B* on the basis of reasoning about *A* is a pragmatic skill that depends on the aim and context of the particular inquiry. (Suárez 2004, 773)

Therefore, it is due to the competent modeler who draws inferences regarding the target based on the source. The inferences take a form of transferring over the *claims* derived from the source onto the *claims* about the target, and there are inference generation rules that guarantee this form of transferring: “such rules are complex features of the practice that involve carrying out demonstrations or modifications of the source in order to guide our beliefs regarding the behaviour of the target” (Suárez 2015, 45–46). These rules can be grouped into two kinds, depending on whether they concern the source or the relation between the source and the target. The first kind of rules are related with the internal structure of the source, from which we can demonstrate or derive various results about the source itself (e.g., “what would happen to this variable if we were to change the value of that variable in the source?”). The second kind of rules are about connecting the source with the target by interpreting features of the source as referring to features of the target.

The inference from the source to the target can be performed using a number of *means*, as long as they allow the modelers to draw conclusions about the aspects of the target system in terms of the corresponding aspects of the source. In other words, the inference does not require any particular kind of means, since “it just requires that there be some means or other” (Suárez 2004, 775). Therefore, the means in some cases might take the form of similarity, isomorphism, partial isomorphism, or whatever, so long as they allow the modelers to make inferences about the target in terms of the source. Notice that means are different from *constituents* of scientific representation, for the latter constitute individually necessary and collectively sufficient conditions for scientific representation while the former do not (Suárez 2015, 46). Analogously put, we may say that means are whatever tools a person might use in undertaking a specific task, and, though employed by the person, they are not themselves part of the task.

An example of Suárez may help us to illustrate the nitty-gritty of his account. The North British Railway Company was in charge of building a rail bridge across the main estuaries of the Firth of Forth in the east of Scotland in the 1870s. It was a very challenging project because it must overcome the problem associated with the stronger side winds in that area (Ibid., 42). The project was further challenged by an event that happened in 1879: the collapse of the Tay Bridge, a nearby estuary bridge, which resulted in breaking down a train with 79 passengers in it (Ibid., 42). Facing these serious challenges, the chief engineer, Benjamin Baker, decided to use a cantilever design instead of a girder design used in the Tay Bridge (Ibid., 42). The principle of the cantilever design involves tension-compression:

> In a cantilever bridge the lower arm of each lever is compressed while the upper arm is correspondingly in tension. In the central pier by contrast the lower girder is compressed while the upper one is in tension. This led Baker to choose different kinds of design for the different arms of the levers—those in tension would be built as lattices, while those in compression were tubular. […]. Lattices minimize resistance to wind pressure, while tubes maximize resistance to compression shears. (Ibid., 42)

According to Suárez, the representational force of the source, i.e., the graphs of the bridge written on papers, is towards ‘a very particular bridge’ capable of withstanding the strong side winds and avoiding the structural defects of the Tay Bridge (Ibid., 43). The inferential capacity of the source, therefore, is “geared towards showing clearly how […] a bridge built as designed would indeed withstand such shears and stresses” (Ibid., 43). More specifically, the representational force of the bridge model involves an understanding of the principles of the cantilever design, and of the various parts of the model (e.g., tubes, girders, lattices, etc.) (Ibid., 43); these elements are then interpreted as referring to their counterparts in the target. The inferential capacity of the model involves not only the principles of the cantilever design but also many other principles associated with torsion, compression, tension, etc., which are needed in calculating how a bridge could resist the strength of the side winds (Ibid., 43); outcomes of reasoning based on these principles and calculations are then transferred to claims about the target.

In sum, Suárez's inferential account of scientific representation features how competent modelers draws inferences about the target based on the source, wherein the inferences often take the form of transferring over the claims derived from the source onto the claims about the target. There is no need to postulate any substantive model-world relationship other than acknowledging the fact that different means of representation may be employed in different cases.

4. An Inferential Account of Model Explanation

In Section 2 we mentioned that Bokulich's account of model explanation ties too closely to a substantive representational relationship between a model and its target in the manner that for a model to be explanatory it must correctly represent the counterfactual structure of its target system. With the aegis of Suárez's inferential account of scientific representation, we are now in a good position to develop an alternative account of model explanation that does not rely on any substantive representational relationship between the model and its target: an inferential account of model explanation. According to such an account, rather than making appeal to any substantive model-world relationship, it is the modeler who *hypothesizes* that the counterfactual structure of the model captures it counterpart in the target.

Based on this idea, my claim about model explanation (ME) boils down to the following statement:

> (ME) It is the modeler who (i) entertains the counterfactual structure of the model by asking Woodward's w-questions, and then (ii) hypothesizes that the claims derived from the counterfactual structure of the model may be applied to its target system.

ME consists of two steps. First, since a model can be described as a structure (Weisberg 2013), i.e., a dependence relationship,[6] it follows that variables in the model counterfactually depend on each other. More specifically, changes (or interventions) in explanans variables that figure in the model can be systematically associated with changes in explanandum variables that sometimes take the form of outputs of the model (note that the explanandum variables, represented in the model, are supposed to describe or reproduce their counterparts in the world). As such, the model can be used to answer Woodward's w-questions about itself: we can ask how one variable in the model would change as a result of intervention on another variable in the model. Second, the modeler then hypothesizes that—based on her background knowledge, modeling goals, conceptualization of the target, etc.—the counterfactual dependence relationships derived from the model may be applied to their counterparts in the target. In other words, a kind of inferential relationship can be hypothesized between the model and its target.

The first step of ME typically takes the form of making *counterfactual statements* (CS for short), for instance:

> (CS) In the model M, had the variable X taken such-and-such a value x_i, then the variable Y would have taken such-and-such a value y_j.

The modeler can play the counterfactual structure of the model in whatever ways,

[6] Christopher Pincock proposes a similar idea ("objective dependence relations") when discussing non-causal explanations. According to him, in addition to causal dependence relations there are abstract dependence relations that can also be used to do explanatory work (Pincock 2015, 878).

but there must be a small set of CSs out of the whole set of all possible CSs that especially interest her. Which set of CSs would particularly interest her largely depends on what kind of questions she would ask and what the modeling goal is in her mind. Once a model has been built, more often than not the modeler will only concentrate on a few number of focal variables and a few number of relations among them, entertaining them in various but constrained ways. For example, the modeler must be aware that a variable may only have physical meanings within a certain range of values, or her might know prior to building the model that a variable is highly likely to be a cause of another variable but not vice versa, etc. The first step should remind us of Suárez's first kind of rules that are related with the internal structure of the source, from which we can demonstrate or derive various results about the source itself (see Section 3).

The second step of ME usually takes the form of making *hypothetical statements* (HS), namely *hypotheses* that transfer over claims derived from the counterfactual structure of the model onto claims about the target. Essentially, this involves the modeler assuming that the counterfactual structure of the model also holds in the target system; this assumption often goes quite quick and unnoticed by the modeler, especially when the model employed is well-developed and widely accepted, e.g., the Lotka-Volterra model. More specifically, the hypothesis usually takes the following form:[7]

> (HS):
>
> (i) If a model M has such-and-such attributes, patterns, or mechanisms,
>
> (ii) then, *hypothetically*, its target system T would also have such-and-such attributes, patterns, or mechanisms.

Notice that making HSs does not require that the model in question must be a

[7] I thank XX and YY for alerting me to know that the inference from the model to its target (and vice versa) is in fact a hypothesis: because the model behaves in such-and-such a way we hypothesize that the target would also behave in such-and-such a way. The formation of the HS is indebted to many colleagues, including XX, XY, YY, and ZZ.

good model. The absence of this requirement has two implications. First, it leaves room for the possibility that a not-so-good model may also offer some explanation—though the explanation might be inaccurate. This in turn implies that offering explanation is not an all-or-nothing issue but a matter of degree, i.e., a model may be more or less better than the other in explaining a phenomenon, and this goes in accordance with the fact that the explanatory power of different models can be compared (Woodward 1997, 2000; Strevens 2004; Weisberg 2004; Weslake 2010). Second, it also leaves room for the possibility that models with different explanatory power are employed by the modelers *in the same way* in explaining phenomena—that is, regardless of whether the model is explanatorily good or not-so-good, it is put into the same two-step practice by the modeler when explaining. Therefore, whether a model is explanatorily good or better than the other seems to be a problem independent of how a model explains, and we need to set extra conditions for determining when a model is explanatorily good or better than the other. However, for the limitations of space, I will leave that problem to another occasion.

Also note that the act of making HSs is insensitive to what kind of model-world relationship is operating, be it isomorphism, partial isomorphism, structural similarity or something the like. This is not denying that there might be such a model-world relation; rather, I think whether a model possesses such a model-world relation or not is orthogonal to the problem of how a model explains. The idea is that there might be a number of distinct relations that different kinds of models bear to their targets, but the fact that all these different kinds of models can be employed to explain essentially converges on one common ground: they all help the modelers to draw conclusions about their targets based on claims derived from these models.

One may find that the HS can run in many different ways: from the model to its target, or from the target to its model, or back and forth. Moreover, through the HS we can see the clear link between prediction and explanation. For any established model M, the fact that M has such-and-such attributes, patterns, or mechanisms (attributes for short) can lead to the *prediction* that its target T may also have such-and-such attributes. In contrast, exploring the specific way the model produces such-and-such attributes leads to the *explanation* of why *T* manifests such-and-such attributes. Therefore, the reason why a model can explain is spelt out by the fact that we can explain an explanandum (or an output, a pattern, a phenomenon) in the target system in terms of the counterfactual structure derived from the model. In other words, by appealing to the counterfactual structure derived from the model but hypothetically extrapolated to the target, we are able to explain why an explanandum would appear in virtue of how the explanans would change.

Ultimately, the inference from the model to its target consists in the *hypothesis* that the counterfactual structure of the model may be applied to the target. In light of the HS, the modelers infer that the same interventions or changes on the variables of the model and the target will lead to the same changes in the outcomes of both the model and the target; that is, the outcomes we get from the model by changing certain variables should also be manifested in the target by changing the corresponding variables. The inference in the second step of ME reminds us of Suárez's second kind of rules that are associated with connecting the source to the target by interpreting features of the source as referring to features of the target. Note that the second step does not necessarily come after the first step—though they can be conceptually distinguished, they often go hand in hand in practice.

Accordingly, an inferential account of model explanation offers a novel way of understanding when a model is explanatorily good and when it is not. For such an account, an explanatorily impoverished model is one that leads a competent modeler to draw wrong conclusions about the target. A good case in point is the phlogiston model: it explains the phenomenon of combustion by postulating that combustible materials incorporate an element called phlogiston, and that when stuff is burning phlogiston is released into the air. This model is explanatorily poor because it inevitably leads a competent modeler to draw conclusions about the combustible materials in reality that, first, do not contain such an element, and second, do not proceed in the way described by the model when burning. Note that saying that an explanatorily poor model can lead a competent modeler to draw wrong conclusions differs from saying that the model is wrong or false, nor does it imply that an explanatorily good model is *true* with respect to the reality it is supposed to represent. As William Wimsatt famously puts, all models are literally false though false models usually lead to truths about the world (Wimsatt 2007).

However, there are other ways to draw wrong conclusions based on models aside from the one described above. For example, an incompetent modeler might draw mistaken conclusions about the target on the basis of an explanatorily good model. This is partly due to the myriad of inferences that can be drawn from a model, and partly due to the modeler's incompetence in distinguishing inferences that are appropriate given the modeling goal from those that are not. In addition, an incompetent modeler might unluckily come up with an explanatorily poor model, resulting in inferences that are wrong-headed compared with the target system in question. All in all, making good explanation is therefore partly due to the model and partly due the modeler entertaining the model.

One merit of this inferential account of model explanation is that it provides a unified theory to accommodate many (if not all) different kinds of models. A wide variety of scientific models are capable of doing explanatory work, e.g., mechanistic models, dynamic models, structural models, agent-based simulation models (the next section will scrutinize how this kind of models explain), etc. A mechanistic model explains by showing how the components of a system interact with one another in producing a certain mechanism, a dynamic model explains by revealing how a system's states change over time, a structural model explains by manifesting how a phenomenon or an explanandum can be derived from the structure of a theory (e.g., explaining the Pauli exclusion principle in fundamental physics) (Bokulich 2011, 40). Admittedly, these are different kinds of explanations. Nevertheless, they share one thing in common: all these model explanations involve drawing inferences that transfer over claims derived from the model onto claims about the target. Similar to what discussed above, the activity of drawing inferences in these explanations proceeds in a two-step manner: making counterfactual statements about the model and building hypothetical statements between the model and the target. For example, a mechanistic model explains by, first, specifying how changing key parts of the system would lead to changes in the system—i.e., specifying the counterfactual structure of the model—and second, extrapolating the results drawn from the model to the target.

Before concluding this section, it is worth briefly discussing the ‘justificatory step’ of Bokulich’s account. Bokulich’s justificatory step serves the purpose of “specifying what the domain of applicability of the model is, and showing that the phenomenon in the real world to be explained falls within that domain” (Bokulich 2011, 39). My discussion of how the model might be linked to its target through the HSs can be viewed as an extension of Bokulich’s justificatory step. This is because, on the one hand, the HSs built by modelers concern aspects of the model (and its target) that fall within the intended domain of applicability of the model. This is the very reason why the modelers bother hypothesizing and entertaining these statements. On the other hand, the HSs do not only concern the intended domain of applicability of the model in general, but more importantly manifest how particular aspects of the model within that domain might correspond to its counterparts in the target. More specifically, they manifest how the set of counterfactual dependence relationships, i.e., the particular aspects of the model within its domain of applicability, can be extrapolated to its counterparts in the target. This second dimension of the justificatory step, which renders my view distinct from Bokulich’s, is what has not been fleshed out in her account.

To sum up, the inferential account of model explanation holds that model explanation is essentially a two-step activity, in which the first step involves making counterfactual statements about the model and the second involves making hypothetical statements linking the model to the target. In all these steps, the modelers—rather than any substantive representational relationship between the model and its target—are of paramount importance to the explanation practice. The next section will illustrate the inferential account with a concrete model.

5. A Case Study: An Agent-Based Simulation Model

The model is drawn from Senior et al. (2015). We may observe a common pattern in many arthropods (e.g., spiders, burying beetles, etc.), such as "the effects of contest competition and the number and composition of foods in the nutritional environment on the evolution of individual nutritional strategies" (Senior et al. 2015, 4–5) (the pattern has been shown in Figure 1 below). The observed common pattern can be described as having to do with three focal variables: intra-specific competition for food, food composition, and the evolution of animals' nutritional strategies (namely nutritional latitude in what follows). Then we might attempt to understand how the former two variables might affect the last variable. According to Lihoreau et al.,

> An individual's nutritional strategy was governed by the fixed global parameter K, which we refer to here as 'nutritional latitude'. When eating a food that will not guide its nutritional state to the IT an individual has some probability of leaving, which is both a function of the balance of nutrients in the food being consumed, and K. Here, a high K means an individual is likely to consume the same imbalanced food until reaching a point of nutritional compromise (at which point it then seeks an alternative). In contrast, a low K corresponds to a low probability that an individual will continue feeding on a food rail that will not guide its nutritional state directly to the IT. (Lihoreau et al. 2014; Cf. Senior et al. 2015, 4)

The IT (Intake Target) mentioned above refers to a coordinate or a region within the nutrient space, which denotes "the optimal amount and blend of nutrients that the animal requires over a specified period in its life" (Senior et al. 2015, 3). Now consider an agent-based simulation model used to explore how the observed common pattern could arise based on the three variables mentioned above.[8] The

[8] "Agent-based modelling (ABM) is a computational modelling paradigm that enables us to describe how any agent will behave" (Wilensky and Rand 2015, 22). By the word *agent*, "we mean an autonomous individual element of a computer simulation. These individual elements have properties, states, and behaviors" (Ibid., 22).

following is the model description:

> Each generation consists of 150 individuals that must attain a certain level of fitness (i.e. nutritional state) within a fixed number of model iterations for it to be considered fit enough to breed. Fitness-proportionate selection then operates among those individuals fit enough to breed, with proximity to the IT (optimal point of nutrient intake in the nutrient space) determining its fitness. We allowed K to evolve 1000 generations under varying levels of competition and in differing nutritional environment (i.e. different abundance and nutritional compositions of food). In doing so, we aimed to explore the effects of contest competition and the number and composition of foods in the nutritional environment on the evolution of individual nutritional strategies. (Ibid., 4-5)

Suppose the population under consideration only feeds on three kinds of food, and we perform the model runs under different intensities of competition, c, which is bounded at 0 and 1. The population mean nutritional latitude K obtained from each model run is also bounded at 0 and 1 (Ibid., 5). Suppose, for simplicity, the environment contains one nutritionally balanced food (e.g., it contains the same amount of protein and carbohydrate), and two imbalanced but complementary foods (e.g., one might contain 40% protein and 60% carbohydrate while the other might contain 60% protein and 40% carbohydrate). For the latter two complementary foods, we can vary the extent of their nutritional imbalance to be either moderate or extreme (see Figure 1 below).

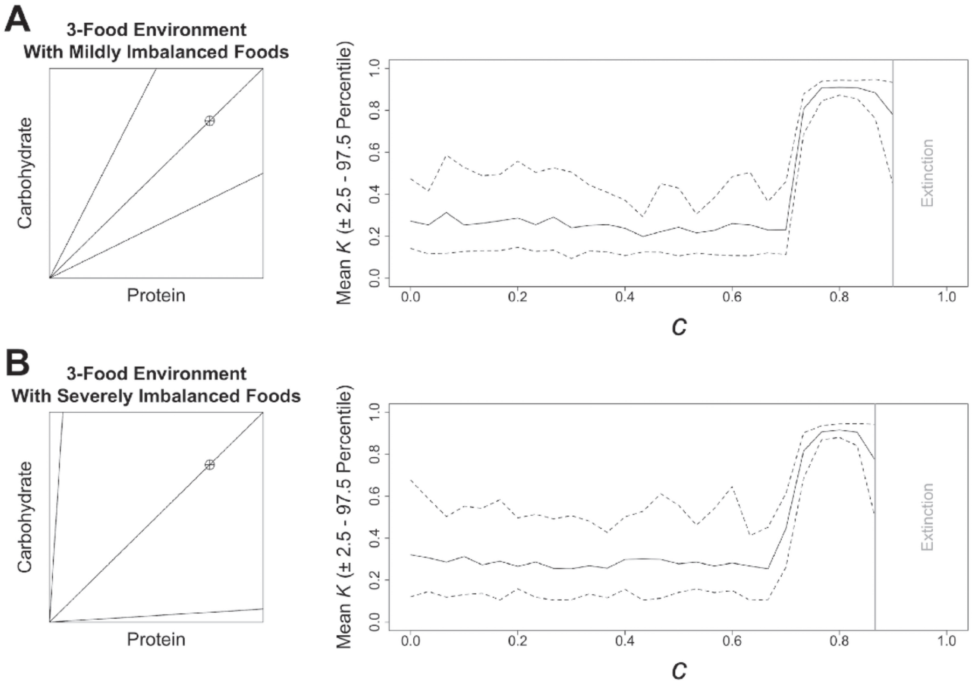

Figure 1. Results for the agent-based simulation model. Lines and crosshairs describe food rails and the intake target. This figure comes from Senior et al. (2015, 7).

The results in Figure 1 are summarized as follows:

> In these environments when $c = 0$, K was stable at a range of values […]. The high variance in stable values of K suggests that no one level of nutritional latitude is optimal where competition is weak, but most low levels are equally fit. In the face of increasing c, K was relatively stable up to a point. With mildly imbalanced foods at $c = 0.7$, and with extremely imbalanced foods at $c = 0.67$, K increased sharply to above 0.91. […] At very high c the population could not support itself as no individuals could fulfil the fitness requirements to be considered in breeding condition by the end of the simulation. (Senior et al. 2015, 5)

That is the outline of the agent-based model. Now let us go back to my claim that, in the first step of model explanation, a model can be used to make counterfactual statements about itself, i.e., can answer how changes in explanans variables that figure in the model can be systematically associated with changes in explanandum variables. In the agent-based model described above, there are three focal variables: competition for food (c), food composition (r), and nutritional latitude (K). The values of both c and K range continuously from 0 to 1, while r only takes two values in our model: mildly imbalanced foods (r_1) and severely imbalanced foods (r_2).

First, consider the case where the food environment takes the value r_1. In r_1, for example, we may ask ourselves: if we were to change the value of c from c_1 to c_2, what would happen to the value of K. In particular, we may ask if we were to change the value of c from 0.1 to 0.2, what would happen to the value of K; if we were to change the value of c from 0.2 to 0.3, what would happen to the value of K; and so on. There are two points that may particularly interest us, i.e., when $c = 0.7$ and $c = 0.9$, because these are points at which K changes drastically. The corresponding counterfactual questions are (a) if we were to change the value of c from 0.6 to 0.7, what would happen to the value of K, and (b) if we were to change the value of c from 0.8 to 0.9, what would happen to the value of K. Based on these counterfactuals and their corresponding answers in r_1, a systematic counterfactual dependence relationship between c and K can be built, which is summarized as follows:

$c = 0, K\ (mean\ value) = 0.27$
$0 < c < 0.7, K = 0.271$
$c = 0.7, K = 0.91$
$0.7 < c < 0.9, K = 0.95$
$c \geq 0.9, K = 0$

The similar situation holds for the case where the food environment takes the value r_2. Therefore, we have built a whole range of CSs based on the relationships between the three focal variables of the model.

Having shown how to build CSs based on a model, now consider how to make HSs that transfer over claims about the model onto claims about the target. We know how variables c, r and K are involved in producing the pattern of behavior.

That is, we know that, for example, “had c taken the value c_i and r taken r_n, the spider population would have instantiated the pattern of nutritional strategy $K = k_j$”. Given these, we make the following HS:

> (**HS***):
>
> (i) In M, if c takes the value c_i and r takes r_n, the spider population instantiates the pattern of nutritional strategy $K = k_j$; and
>
> (ii) in T, if c takes c_i and r takes r_n; then
>
> (iii) *hypothetically*, the spider population in T would also instantiate the pattern of nutritional strategy $K = k_j$.

As in the first step, in this step a whole range of HSs linking the model to its target can be systematically constructed. For example, we can not only make the HS that (i) in M, if c takes the value 0.2 and r takes r_1, the spider population would instantiate the pattern of nutritional strategy $K = 0.271$, and (ii) in T, if c takes the value 0.2 and r takes r_1, then, (iii) hypothetically, the spider population in T would also instantiate the pattern of nutritional strategy $K = 0.271$, but also the HS that (i) in M, if c takes the value 0.7 and r takes r_1, the spider population would instantiate the pattern of nutritional strategy $K = 0.91$, and (ii) in T, if c takes the value 0.7 and r takes r_1, then, (iii) hypothetically, the spider population in T would also instantiate the pattern of nutritional strategy $K = 0.91$, and so on.

Making these whole range of HSs matters to both model prediction and explanation. For instance, given the actually observed or detected values of the variables c and r in a spider population, we are able to predict what kind of nutritional strategy K that spider population would instantiate. Likewise, we are able to explain why a real spider population instantiates a specific nutritional strategy K by pointing to the two variables c and r in the population that are believed to be causally responsible for bringing about that nutritional strategy. Recall that the counterfactual dependence relationships (causal relations in this case) among these focal variables are first established within the agent-based simulation model, and then hypothetically extrapolated to the reality in order to see whether the established counterfactual dependence relationships can shed some light on the phenomenon observed in reality. The thrust of model explanation, therefore, relies on how the modeler could appropriately extrapolate claims based on the counterfactual structure of the model to the target system. And there are many ways that, as discussed in the last section, a modeler may fail to do so: (a) she has an explanatorily bad model, or (b) she is an incompetent modeler with an explanatorily good model, or (c) she is an incompetent modeler with an explanatorily bad model.

It is time to take stock. So far we have shown that the agent-based simulation model is explanatory because (i) it allows the modeler to entertain the counterfactual structure of the model in various ways, and (ii) it helps the modeler to build hypothetical statements transferring over claims about the model onto claims about the target based on the counterfactual statements constructed in (i).

6. Conclusion

Bokulich's account of model explanation ties too closely to a substantive representational relationship between the model and its target system, because it holds that to explain a model should correctly represent the pattern of the counterfactual structure of its target system. I claimed that to properly account for how a model explains, we should divert our attention from postulating any substantial representational relationship to something else. That is the essential role played by the modeler in modeling practice, rather than any substantive model-world relationship between the model and its target system. Putting the modeler on center stage, we have seen that model explanation essentially proceeds in a two-step way: (i) the modeler first entertains the counterfactual structure of the model in various but also constrained ways, and (ii) she then infers from the model to the target by making a whole range of hypothetical statements that transfer over claims derived from the model onto claims about the target. This conception of model explanation does not deny the existence of some sort of relationship between the model and its target, but rather suggesting that the relationship itself might be not the right place to look at in understanding model explanation.

References

Ariew A, Rice C, Rohwer Y (2015) Autonomous-Statistical Explanations and Natural Selection. British Journal for the Philosophy of Science 66:635–658.

Baron, Sam, Mark Colyvan, and David Ripley. 2017. "How Mathematics Can Make a Difference." *Philosophers' Imprint* 17 (3):1–29.

Batterman, Robert W. 2002a. "Asymptotics and the Role of Minimal Models." *The British Journal for the Philosophy of Science* 53 (1):21–38.

———. 2002b. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press.

Batterman, Robert W., and Collin C. Rice. 2014. "Minimal Model Explanations." *Philosophy of Science* 81 (3):349–376.

Bokulich, Alisa. 2008. "Can Classical Structures Explain Quantum Phenomena?" *The British Journal for the Philosophy of Science* 59 (2):217–35.

———. 2011. "How Scientific Models Can Explain." *Synthese* 180 (1):33–45.

———. 2012. "Distinguishing Explanatory from Nonexplanatory Fictions." *Philosophy of Science* 79 (5):725–37.

Bueno, Otávio. 1997. "Empirical Adequacy: A Partial Structures Approach." *Studies in History and Philosophy of Science Part A* 28 (4):585–610.

Bueno, Otávio, Steven French, and James Ladyman. 2002. "On Representing the Relationship between the Mathematical and the Empirical*." *Philosophy of Science* 69 (3):452–73.

———. 2012. "Empirical Factors and Structure Transference: Returning to the London Account." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 43 (2):95–104.

Contessa, Gabriele. 2007. "Scientific Representation, Interpretation, and Surrogative Reasoning." *Philosophy of Science* 74 (1):48–68.

Craver, Carl F. 2006. "When Mechanistic Models Explain." *Synthese* 153 (3):355–76.

Craver CF (2007) Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Oxford University Press, Oxford.

Da Costa, Newton CA, and Steven French. 2003. *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*. Oxford University Press.

Downes, Stephen M. (1992) "The Importance of Models in Theorizing: A Deflationary Semantic View". P*SA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992:142–153.

French, Steven. 1997. "Partiality, Pursuit and Practice." In *Structures and Norms in Science*, by M. L. Dalla Chiara, K. Doets, D. Mundici, and J. van Bentham, 35–52. Dordrecht: Kluwer Academic Publishers.

———. 2003. “A Model-Theoretic Account of Representation (Or, I Don’t Know Much about Art…but I Know It Involves Isomorphism).” *Philosophy of Science* 70 (5):1472–83.

French, Steven, and James Ladyman. 1998. “Semantic Perspective on Idealization in Quantum Mechanics.” In *Poznan Studies in the Philosophy of the Sciences and the Humanities*, by N. Shanks, 63:51–74. Amsterdam: Rodopi.

Frigg, R. (2006). “Scientific Representation and the Semantic View of Theories.” *Theoria*, 21 (1): 49-65.

Frisch, Mathias Florian. 1998. “Theories, Models, and Explanation.” University of California, Berkeley.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

———. 2008. “Why Scientific Models Are Not Works of Fiction.”

Giere, Ronald N. 2004. “How Models Are Used to Represent Reality.” *Philosophy of Science* 71 (5):742–52.

Godfrey-Smith, Peter. 2006. “The Strategy of Model-Based Science.” *Biology and Philosophy* 21 (5):725–40.

Hempel C (1965) Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. The Free Press

Hughes, Richard IG. 1997. “Models and Representation.” *Philosophy of Science* 64:S325–S336.

Kennedy, Ashley Graham. 2012. “A Non Representationalist View of Model Explanation.” *Studies in History and Philosophy of Science Part A* 43 (2):326–332.

King, Martin. 2016. “On Structural Accounts of Model-Explanations.” *Synthese* 193 (9):2761–78.

Knuuttila, Tarja. 2005. “Models, Representation, and Mediation.” *Philosophy of Science* 72 (5):1260–71.

———. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Model-Based Representation in Scientific Practice* 42 (2):262–71.

Lange M (2013) Really Statistical Explanations and Genetic Drift. Philosophy of Science 80:169–188

Lihoreau, Mathieu, Jerome Buhl, Michael A. Charleston, Gregory A. Sword, David Raubenheimer, and Stephen J. Simpson. 2014. “Modelling Nutrition across Organizational Levels: From Individuals to Superorganisms.” *Journal of Insect Physiology* 69:2–11.

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

———. 1984. *A Case for Scientific Realism*. na.

———. 1985. “Galilean Idealization.” *Studies in History and Philosophy of Science Part A* 16 (3):247–73.

Morgan, Mary S., and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Vol. 52. Cambridge University Press.

Morrison, Margaret. 1999. “Models as Autonomous Agents.” In *Models as Mediators: Perspectives on Natural and Social Science*, 52:38–65. Cambridge: Cambridge University Press.

Odenbaugh, J. (2008). “Models.” In S. Sarkar, and A. Plutynski (Eds.), A Companion to the Philosophy of Biology, 506-524. Blackwell Publishing.

Parker, Wendy S. 2003. “Computer Modeling in Climate Science: Experiment, Explanation, Pluralism.” University of Pittsburgh.

Pincock, Christopher. 2012. *Mathematics and Scientific Representation*. Oxford: Oxford University Press.

———. 2015. “Abstract Explanations in Science.” *The British Journal for the Philosophy of Science* 66 (4):857–882.

Rice C (2015) Moving beyond causes: Optimality models and scientific explanation. *Noûs* 49:589–615.

Salmon W (1984) Scientific explanation and the causal structure of the world. NJ: Princeton University Press, Princeton.

Senior, Alistair M., Michael A. Charleston, Mathieu Lihoreau, Jerome Buhl, David Raubenheimer, and Stephen J. Simpson. 2015. “Evolving Nutritional Strategies in the Presence of Competition: A Geometric Agent-Based Model.” *PLoS Comput Biol* 11 (3):e1004111.

Sneed, Joseph D. 1971. *The Logical Structure of Mathematical Physics*. Vol. 35. Springer Science & Business Media.

Stegmüller, Wolfgang. 1976. *The Structure and Dynamics of Theories*. Springer.

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

Suárez, Mauricio. (2003). “Scientific Representation: against Similarity and Isomorphism.” *International Studies in the Philosophy of Science* 17 (3): 225-244.

———. 2004. “An Inferential Conception of Scientific Representation.” *Philosophy of Science* 71 (5):767–779.

———. 2015. “Deflationary Representation, Inference, and Practice.” *Studies in History and Philosophy of Science Part A* 49:36–47.

———. 2016. “Representation in Science.” In *The Oxford Handbook of Philosophy of Science*, by Paul Humphreys, 441–60. Oxford: Oxford University Press.

Suppe, Frederick. 1977. *The Structure of Scientific Theories*. Urbana, Illinois: University of Illinois Press.

Suppes, Patrick. 1962. “Models of Data.” In *Logic, Methodology, and the Philosophy of Science*, by Ernest Nagel, Patrick Suppes, and Alfred Tarski, 24–35. Stanford University Press.

———. 1967. “What Is a Scientific Theory?” In *Philosophy of Science Today,* by Sidney Morgenbesser. New York: Meridian Books.

Van Fraassen, Bas C. 1970. “On the Extension of Beth’s Semantics of Physical Theories.” *Philosophy of Science* 37 (3):325–339.

———. 1972. *A Formal Approach to the Philosophy of Science*. University of Pittsburgh Press.

———. 1980. *The Scientific Image*. Oxford: Oxford University Press.

———. 2008. *Scientific Representation*. Oxford: Oxford University Press.

Walsh DM, Ariew A, Lewens T (2002) The trials of life: Natural selection and random drift. Philosophy of Science 69:452–473

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

Wilensky, Uri, and William Rand. 2015. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. MIT Press.

Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press.

Woodward, James. 1997. “Explanation, Invariance, and Intervention.” *Philosophy of Science* 64:S26–S41.

———. 2000. “Explanation and Invariance in the Special Sciences.” *The British Journal for the Philosophy of Science* 51 (2):197–254.

———. 2001. “Law and Explanation in Biology: Invariance Is the Kind of Stability That Matters.” *Philosophy of Science* 68 (1):1–20.

———. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

———. 2010. “Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation.” *Biology & Philosophy* 25 (3):287–318.