Synthese (2023) 201:171
https://doi.org/10.1007/s11229-023-04084-8

ORIGINAL RESEARCH

Toy models, dispositions, and the power to explain

Philippe Verreault-Julien[1]

Received: 21 March 2022 / Accepted: 8 February 2023 / Published online: 5 May 2023

Abstract
Two recent contributions have discussed, and disagreed, over whether so-called toy models that attempt to represent dispositions have the power to explain. In this paper, I argue that neither of these positions is completely correct. Toy models may accurately represent, satisfy the veridicality condition, yet fail to provide how-actually explanations. This is because some dispositions remain unmanifested. Instead, the models provide how-possibly explanations; they *possibly* explain.

1 Introduction

Some models are so idealized and simple that they are occasionally called *toy* models. But far from only playing with them, scientists use these models to represent actual targets and for epistemic purposes like explaining and understanding. Whether and how they can fulfil that representational role has been the subject of substantial philosophical debate. In particular, under what conditions they have the power to explain is an enduring puzzle.

Participants in the discussion sometimes view these models as not performing any representational function (Luczak, 2017), as not satisfying the veridicality condition and providing how-possibly explanations (Reutlinger et al., 2018), as accurately representing their targets and providing how-actually explanations (Nguyen, 2020), as helping to probe the modal features of phenomena (Gelfert, 2019), or as serving epistemic functions *because* of their highly idealized nature (Jebeile, 2020). In this paper, I put forward an underappreciated alternative, namely that some of these models accurately represent their targets, but only provide how-possibly explanations. More specifically, I argue that a toy model may accurately represent an explanandum and an explanans, yet only *possibly* explain.

✉ Philippe Verreault-Julien
p.verreault-julien@tue.nl

[1] Eindhoven University of Technology, Eindhoven, Netherlands

I use the recent debate between Reutlinger et al. (2018) and Nguyen (2020) as a frame for my discussion. Although they study similar models, they disagree on how to interpret them. Reutlinger et al. hold that some toy models do not satisfy what they call the 'veridicality condition' and provide how-possibly explanations. Nguyen contends that they accurately represent their targets and provide how-actually explanations. One noteworthy feature of that debate is that they consider how models may explain by citing broadly construed *dispositions*. In short, the position I defend is that some of these toy models (1) accurately represent dispositional properties of actual targets, yet (2) only provide how-possibly explanations because the dispositional properties remain unmanifested. This is an interesting case of a how-possibly explanation with a true explanandum and explanans, but with a non-actual explanatory relation.

Section 2 introduces the debate between Reutlinger et al. and Nguyen. In Sect. 3, I introduce the theoretical framework that specifies under what conditions dispositions have the power to explain. Section 4 argues that the disposition under consideration satisfies the veridicality condition, but does not explain how-actually. Then, in Sect. 5 I defend the how-possibly interpretation. Section 6 concludes.

2 The debate

Reutlinger et al. (2018) distinguish between two types of toy models: embedded and autonomous. The former stem from a well-confirmed framework theory (e.g. quantum mechanics) while the latter do not; they are independent of a particular theoretical framework and stand on their own. According to them, embedded toy models may afford understanding of phenomena because they satisfy, among other things, what they call the *veridicality condition*.[1] This condition captures the widely shared idea that an explanation's "explanatory assumptions (that is, the explanans of an explanation) are required to be true or, at least, approximately true" (Reutlinger et al., 2018, p. 1082).[2] Since embedded toy models are models of a well-confirmed — i.e., true — theory, whatever justification we have in the framework theory will carry over the embedded toy model. And because their idealizations are justified on the basis of theory, these toy models may satisfy the veridicality condition. For that reason, they conclude that these models provide *how-actually explanations* because they contain "actually true (or approximately true) statements" (p. 1085).

Autonomous toy models, however, are more complicated. They are as idealized and simple as their embedded counterparts, but do not stem from any well-confirmed framework theory. They stand on their own and, in consequence, the epistemic justification for their idealizations cannot come from theory. It thus needs to originate

[1] They also state two other conditions for understanding phenomena, viz. the explanation and epistemic accessibility conditions. For the purpose of the current discussion, only the veridicality condition is important.

[2] The veridicality condition also applies to explananda and correctly explaining also demands a true description of the explanandum (see Reutlinger, 2016, 2018). Here I focus on the truth of the explanans.

from another source. Reutlinger et al. identify three possible strategies for justifying the idealizations of autonomous toy models.

1. **McMullin's (1985) strategy**: Idealizations serve to isolate (approximately) true explanatory factors of interest. Framework theories then allow to de-idealize the models.
2. **Minimalism**: Toy models accurately represent the relevant explanatory factors and the idealizations concern features that are irrelevant to the explanation.
3. **Dispositionalism**: Toy models accurately represent the dispositional behaviour of systems in the absence of disturbing factors.

Using one of these strategies (or a combination thereof) with embedded toy models allows showing how the models satisfy the veridicality condition. For instance, a background theory may allow distinguishing between the explanatory relevant idealizations and those that are not (minimalism) (see also Strevens, 2008, Chap. 8; Weisberg, 2007). Even though the models are highly idealized, they can still truthfully represent explanatory factors of interest and provide how-actually explanations.

However, Reutlinger et al. argue, these strategies do not always work with autonomous toy models. That is, neither McMullins's strategy, minimalism, nor dispositionalism can in every instance justify their idealizations and salvage truth. As a result, they conclude that some autonomous toy models do not provide how-actually explanations.[3] They nevertheless see an epistemic role for these models, namely to provide *how-possibly explanations*. According to them, how-possibly explanations, contrary to how-actually explanations, do not satisfy the veridicality condition because their explanantia "refer to merely possible explanatory factors (for instance, to possible causes and mechanisms bringing about the explanandum phenomenon, if the explanation is causal)" (Reutlinger et al., 2018, p. 1085). How-possibly explanations, they say, are not epistemically inert and may afford 'how-possibly understanding'. That understanding may serve modal (contributing to understanding modal phenomena), heuristic (stepping stone towards how-actually explanation), and pedagogical purposes (helping people grasp ideas).

Examining similar cases as Reutlinger et al., Nguyen (2020) argues that (some) autonomous toy models accurately represent their target systems by identifying 'susceptibilities'. Susceptibilities are broadly construed properties that make a system disposed to manifest a behaviour. Even though toy models are simple and highly idealized, when properly interpreted they may nevertheless generate true claims about their targets. But not only does Nguyen dispute that toy models misrepresent, he also submits that "it's plausible that all of the models I discuss in this article go beyond how-possibly to how-actually explanations of the behaviours of target systems" (2020, p. 1017). In fact, he explicitly contrasts his approach to Reutlinger et al's, which prompts the questions of why they reach different conclusions and what is the correct interpretation of these toy models.

[3] The 'some' is important here. They do not claim that *all* autonomous toy models fail to provide how-actually explanations. My own argument does not hinge on the scope of their claim.

Let us examine more closely the crux of the debate. To illustrate their arguments, Reutlinger et al. use a collision model in econophysics (Dragulescu & Yakovenko, 2000) and Schelling's checkerboard model of residential segregation (Schelling, 1971, 1978). Nguyen discusses Akerlof's (1970) market for lemons, the Lotka-Volterra (1927, 1928) predation model, and also the checkerboard model of residential segregation.[4] In order to stay as faithful as possible to the debate and better expose the issue at stake, in the following I will mainly focus on the checkerboard model. However, I believe the following discussion generalizes to other examples of autonomous toy models.

The checkboard model has been widely discussed in the literature (e.g. Aydinonat, 2007; Grüne-Yanoff, 2009; Sugden, 2000, 2009; Ylikoski & Aydinonat, 2014) and I will thus only very briefly introduce it. It is an agent-based model that represents two types of agents spatially located on a checkerboard. Each type of agent prefers to have at least 30% of like neighbours and becomes discontent whenever the neighbours ratio goes below the threshold. Since each agent can have eight neighbours, this implies that agents want to have at least three neighbours of the same type. If an agent becomes dissatisfied, it moves to another location where the minimum ratio holds. Agents assess in turn whether they are satisfied or not and either stay in place or move to a different location. For instance, an agent leaving or moving in may make other agents dissatisfied and cause them to relocate. The model is simple and highly idealized: it assumes only two types of agents, they are 'living' on a grid, agents can and do move when dissatisfied, it abstracts from other plausible factors, etc. The key result of the model is that patterns of segregation will emerge as the unintended consequence of a moderate preference for like neighbours. Importantly, this is viewed as contradicting the claim that strong discriminatory preferences were necessary for bringing about segregation (see e.g. Grüne-Yanoff, 2009). This result is robust under a wide variety of changes of assumptions (e.g. Muldoon et al., 2012), although the model is also sensitive to other assumptions such as modelling agents' preferences as a threshold function (versus continuous) (e.g. Bruch & Mare, 2006).

Reutlinger et al. argue that the dispositionalist strategy is not available for justifying the model's idealizations. Recall that the strategy involves identifying the dispositional behaviour of a system in the absence of disturbing factors. For instance, suppose a model representing objects falling at a rate of 9.80665 m/s^2 when friction or other disturbing factors are absent. We could say that this model represents the *disposition* (earthly) objects have to fall when disturbing factors like air resistance are absent. However, not all objects will actually fall at that rate. In reality, some objects are non-trivially subject to air resistance, like leaves or feathers. Thus, we need to know the 'laws of interaction and composition' of the dispositions the model represents. These laws state how different dispositions (or causes) interact and aggregate with one another to produce the actual behaviour of the system. As it happens, we know how air resistance and gravitational attraction interact with one another, i.e. we know the laws of interaction and composition. Thus, that the model assumes away disturbing factors is not necessarily a problem. What it claims about falling objects is

[4] Reutlinger et al. mention that the Lotka-Volterra model is an autonomous toy model, but do not discuss it further.

still true, viz. that they are actually subject to that force even if their actual acceleration might depend on other factors.

The dispositionalist strategy consists in applying a similar interpretation to the checkerboard model. The model represents the *disposition* of cities or neighbourhoods to become residentially segregated when people have a moderate preference for like neighbours. This holds *in the absence* of other possible causes of segregation like socioeconomic status, preferences for local public goods, policies, institutional discrimination, etc. In reality, many causes may interact together to bring about segregation. Reutlinger et al's main argument against applying the dispositional strategy on some autonomous toy models such as the checkerboard is that since we do not know the laws of interaction and composition of the dispositions, then they cannot satisfy the veridicality condition. We do not know the laws precisely because these are not *embedded* toy models, but *autonomous* ones. And without the laws, we cannot evaluate whether and how the disposition relates to the phenomena to be explained. Therefore, the dispositionalist strategy cannot be put to use to satisfy the veridicality condition.

As already noted, Nguyen (2020) explicitly proposes to interpret the checkerboard model in dispositional terms. First, he says that we can view the model as establishing the following claim.

> r_2: Even for low 'content' thresholds, almost irrespective of the initial set up of the board and the order of movement, the board results in a segregated state (in fact, for particular model runs and choice of movement order, and rule, an initial distribution is associated with a unique segregated outcome). (p. 1030)

However, Nguyen argues, we do not and should not directly export this claim to target systems of interest. Rather, we should reinterpret the above claim as follows.

> r_2': A city whose residents have weak preferences regarding the skin colour of their neighbours has a susceptibility towards global segregation. (p. 1030)

While the model claim suggests that segregation almost necessarily occurs and is quite specific about how this happens (because of the rules agents follow), the features we impute on the target are instead dispositional and less specific. Suitably interpreted, Nguyen argues, these dispositional claims about target systems *are* true: "r_2' tells us that in a particular city, Chicago for example, if residents only require a small proportion of their neighbours to share their skin colour, this still makes the city susceptible to global segregation" (p. 1031). And they are true even if the disposition interacts with disturbing factors or is fully overridden.

That autonomous toy models can serve to establish true dispositional claims about target systems leads Nguyen to hold that they do not only provide how-possibly explanations, but how-actually ones. If we were to directly export the claims from the toy models to the target, then maybe we would need the laws of interaction and composition. But, properly qualified, the dispositions imputed to target systems remain true even if we do not know how they interact or if they are overridden.

Who is right, then? Reutlinger et al., who argue that the dispositionalist strategy fails and that the models thus does not satisfy the veridicality condition, or Nguyen who says that the same models accurately represent and provide how-actually explanation? I contend that they are both partly right and wrong. With Nguyen (2020), I hold that some of these models indeed accurately represent their targets by identifying dispositional properties of actual targets. However, with Reutlinger et al. (2018), I agree that these models should not be interpreted in a how-actually sense. My view is that (some) autonomous toy models:

1. Accurately represent dispositions of actual targets.
2. Only provide how-possibly explanations because the dispositions are unmanifested.

Importantly, I do not claim that all or even most autonomous toy models have these features. My claim is simply that at least one very important case discussed in the literature, namely the checkerboard model, has them. I believe it is plausible that this novel view applies to other cases, but I will not discuss it. In the following sections, I defend my view. But first, I introduce the required theoretical framework.

3 Dispositions and the power to explain

What are dispositions?[5] For the current purposes, what is often called the 'simple conditional analysis' (SCA) (Choi & Fara, 2021) will do. It is an open question how best to characterize dispositions and I want to remain non-committal with respect to which account of dispositions is the correct one. What matters is that dispositions have manifestations in a range of circumstances and thus I believe my point generalizes to other types of accounts (e.g. Vetter, 2015). I am using the SCA to help highlight one problem I want to discuss, viz. that dispositions may remain unmanifested. According to the SCA,

Simple conditional analysis (SCA)
x is disposed to *M* when *C* iff *x* would *M* if it were the case that *C*

where *x* is typically an object, *M* is the disposition's manifestation, and *C* denotes a range of circumstances under which *x* would do *M*. To take a stock example, to say that glass is breakable — that it is disposed to break when struck — means that glass would break if it were struck. There are many counterexamples to the SCA due to "finks", "masks", or other "Achilles' heels" (see e.g. Manley & Wasserman, 2008; Martin, 1994). For example, we may want to say that glass is breakable even if it is in a packaging that prevents it from breaking (a mask). But what matters here is that dispositions have manifestations *M* that occur in a range of circumstances *C*.

[5] I will talk of dispositions, but they are sometimes discussed as being 'potentialities' (Vetter, 2015), 'powers' (Mumford & Anjum, 2011), or 'capacities' (Cartwright, 1989).

Although how and under what conditions dispositions explain is controversial, there is substantial support that they can do so (Bird, 2007; Cartwright, 1989; Mckitrick, 2005; Mumford & Anjum, 2011).[6] The debate of interest in the current paper concerns whether the dispositionalist strategy allows to interpret some toy models as providing either how-actually or how-possibly explanations. As we have seen above, Reutlinger et al. (2018) spell out the distinction in terms of how-possibly explanations not satisfying the veridicality condition and referring to merely possible explanantia. But the veridicality condition and referring to merely possible explanantia seem to be two different sets of issues. A first important step consists in reassessing the role of the veridicality condition in the distinction between how-actually and how-possibly explanations. According to a recent account (Verreault-Julien, 2019), how-actually explanations have the following form,

How-actually explanation (HAE)
p because *q*

where *p* denotes the explanandum and *q* the explanans (e.g. a causal generalization plus initial conditions). Crucially, '*p* because *q*' is an HAE because *p* and *q* are *actual* and the actual explanandum *p* is actually because the actual explanans *q*. Following Verreault-Julien, how-possibly explanations have the following form,

How-possibly explanation (HPE)
$\Diamond$(*p* because *q*)

where the explanation is now under the scope of a modal possibility operator $\Diamond$. In a nutshell, HPEs are *possible* explanations. Often, an HPE will have an actual explanandum, but a merely possible explanans. For example, there are many possible explanations for the disappearance of Neanderthals (Vaesen et al., 2021). Some cite competition between modern humans and Neanderthals while others emphasize environmental factors. These are all HPEs of the disappearance. While there might be multiple HPEs of the disappearance, there is only one HAE (which may combine explanatory factors). Possible explanations play an important role in explanatory reasoning such as inference to the best explanation (Lipton, 2004).

One key feature of that characterization is that both HAEs and HPEs have a truth value. '*p* because *q*' is true iff *p* and *q* are both actual and *p* is actually because *q*, where 'because' refers to one's favourite relation of explanatory entailment (e.g. causal or deductive-nomological). Likewise, that an explanation is possible also has a truth value. '$\Diamond$(*p* because *q*)' is true iff it is the case that '*p* because *q*' is possible.

[6] A worry one might have with citing dispositions is that they do not provide the sort of explanatory information that allows answering explanation-seeking why-questions. Explaining why opium induces sleep by citing its 'dormitive virtue', as in Molière's *Le Malade Imaginaire*, is arguably not very informative. However, other usages do not seem as patently inadequate. If one asks me why a given glass-made object broke, citing the fragility of glass may be a valuable piece of information; some objects are fragile and, as a result, can break. Dispositions also play an important role in scientific practice (see Illari & Russo, 2014, Chap. 15). At any rate, here I want to bracket as much as possible these concerns and assume that dispositions may explain.

Verreault-Julien argues that the truth value depends on the particular interpretation of the modal operator. For instance, claims of nomic possibility do not have the same truth conditions as claims of more restricted causal possibility. And it is also possible to distinguish between other general forms of possibility such as epistemic and objective possibility (Sjölin Wirling & Grüne-Yanoff, 2021; see also Grüne-Yanoff & Verreault-Julien, 2021). The veridicality condition, we have seen, requires the truth of the explanans. There is a sense in which the explanans of an HPE *is* true; it is truly possible.[7] Truth of possibility is still truth. Thus, we can disentangle the veridicality condition from reference to possible explanatory factors. HAEs and HPEs alike can satisfy the veridicality condition, but only the former involve citing actual explanatory factors.

Now that we have properly distinguished between HAEs and HPEs, let us apply the framework to *dispositional* HAEs and HPEs. Combining the SCA and HAE, we arrive at the following formulation,

Dispositional HAE
p because [*q* is disposed to *p* when *C*]

where the explanans *q* is now a disposition.[8] To explain how-actually with a disposition consists in explaining the explanandum, typically the disposition's manifestation, by appealing to the disposition. To understand the conditions under which dispositions may provide an HAE, it is useful to examine two ways dispositions can *fail* to explain in that sense.

First, the truth value of the disposition may be false. Disposition ascriptions such as '*q* would *p* if it were the case that *C*' may be true or false, objects or states of affairs may or may not have the properties we ascribe to them. For instance, the claim that 'maraging steel would break if it were struck' is false; maraging steel is among the toughest materials there is. To provide a dispositional HAE requires citing the true dispositional properties of an object or states of affairs. Second, even if the disposition is true, the explanandum may not be *actually because* of the disposition. In other words, the explanandum may not be a manifestation of the disposition. So the failure is not with the explanandum or the explanans proper, but with the explanatory relation between them.

If, as Nguyen (2020) holds, the checkerboard model provides an HAE, then the model does not fail for any of these two reasons; there is an explanandum that is actually because of a true disposition. But if, as Reutlinger et al. (2018) hold, the model does not provide an HAE, the model must fail for one of the two reasons. Clearly, Reutlinger et al. would at least subscribe to the second failure, i.e. that the explanandum is not actually because of the disposition. Since we do not know the laws of interaction and composition, we cannot determine whether the explanandum is due

[7] I simplify by considering only the case of an actual explanandum with a possible explanans, but there are other possibilities that will become relevant later.

[8] I remain agnostic between the views that dispositions cause their manifestations or that it is a disposition's causal basis that causes its manifestation. There might also be non-causal dispositions (Nolan, 2015).

to the disposition or not. Perhaps there are other disturbing factors that bring about a similar manifestation, but prevent the disposition from contributing to the manifestation. And since autonomous toy models do not rely on a theoretical framework, we do not know how the dispositions interact with each other. But whether they would also hold that the dispositional HAE fails in the first sense is not as evident. They repeatedly suggest that if we cannot use any of the strategies to interpret and justify the idealizations, then the toy model's idealizations (and explanantia) are not compatible with the veridicality condition. Reutlinger (2018) also suggests in other writings that HPEs do not satisfy the veridicality condition. One plausible interpretation is that they consider the HAE to fail in both ways.

In the remainder of this paper, I will argue with Nguyen that the HAE does not fail for the first reason: the model accurately represents a true disposition. However, pace Nguyen and with Reutlinger et al., I will argue that it fails for the second reason.

4 True disposition, but not a how-actually explanation

We have seen that HAEs take the form of '*p* because *q*' statements. In the case of the checkerboard model, what could be the explanandum and the explanans? The explanandum is the (actual) phenomenon of residential segregation. Many cities exhibit patterns of highly segregated areas and this is in need of an explanation. For the explanans, there are two main possibilities, one identifying cities and the other preferences as the bearers of the disposition. Let us look at the first one.

HAE 1-checkerboard
There is residential segregation because cities are disposed to be segregated when people have moderate preferences for like neighbours.

For the sake of the argument, suppose that cities have that disposition and that people have these preferences. In other words, the disposition is in the right sort of circumstances to bring about its manifestation. There is a weak and a strong objection to the HAE interpretation. The weak objection consists in pointing out that, in fact, we do not know how the disposition interacts with other known dispositions for segregation (e.g. prejudice or institutional discrimination; see, e.g., Fossett [2006]). This is basically Reutlinger et al.'s argument. Without that sort of information, we cannot determine to what extent, if at all, segregation is actually due to this disposition or other ones. Other dispositions may, for instance, mask the manifestation of cities' disposition to become segregated when people have moderate preferences for like neighbours.

The strong objection asks a different, but crucial, question. What if cities did not have that disposition or people did not have those preferences? Would there still be segregation? Whether an explanandum counterfactually depends on an explanans is a typical test for explanation (e.g. Reutlinger, 2018; Strevens, 2008; Woodward, 2003). For one, Reutlinger holds that one condition on what makes explanations explanatory (in the HAE sense) is that they satisfy what he calls the 'dependency condition'. This condition requires that the explanatory generalization supports at least one coun-

terfactual such that had the initial conditions been different, then the explanandum would also have been different. The idea behind the condition is that an explanandum that does not depend on an explanans is not actually explained by it. Does HAE 1-checkerboard satisfy the dependency condition?

We have good reasons to believe it does not. In short, the empirical and theoretical evidence about residential segregation indicates that, yes, there would still be segregation *absent* that disposition. Empirically, some have noted that many people do not have mere moderate preferences for like neighbours (e.g. Clark, 1992; Farley et al., 1997; see also Bobo & Zubrinsky, 1996; Charles, 2003; Krysan & Farley, 2002). Notably, people tend to prefer to live with a small or large *majority* of people of their own group. Clark (1992, p. 464) concludes that "[t]he unwillingness of groups to prefer and/or to choose combinations that do not include large proportions of their own race is a significant force in creating separate racial and ethnic areas in this metropolitan area". Theoretically, many other candidate explanations of residential segregation do not rely on the disposition. Some hold that preferences simpliciter cannot explain the phenomenon and that discrimination in housing is instead responsible (e.g. Massey & Denton, 1993). And while it is clear according to Fossett (2006) that preferences can bring about segregation, he also notes that "[t]he data needed to generate empirically grounded estimates of the separate effects of discrimination and social distance and preference dynamics *in real urban systems* simply do not exist" (195, emphasis in original). He also acknowledges that segregation may be overdetermined, e.g. the result of discrimination dynamics. Since neighbourhoods *are* segregated, the overdetermination problem implies that preferences may not actually have the opportunity to manifest themselves. All this indicates that if HAE 1-checkerboard is not simply false, at best we are not in the position to establish its truth.

Nguyen (2020, p. 1033) asks, "[w]hat about cases where the susceptibilities are overridden entirely"? According to him, the model is still accurate; that a disposition is overridden still implies that the disposition *is* in the target system. Otherwise, it could not be overridden. To use his smoking example, it is true that a smoker is disposed to develop lung cancer even if that person does not actually develop it. Dispositions may remain unmanifested. But if we push the 'overridden' logic further, what could the disposition possibly explain? Granted, smoking increases one's susceptibility to lung cancer. Imagine, however, a smoker who receives a lung cancer 'vaccine' that effectively prevents cancer from developing. Since it is precisely this susceptibility that the vaccine is preventing, would that person still have a susceptibility to developing lung cancer? An overridden disposition might exist, yet never be in the position to manifest itself. We might still want to say that, e.g., *smoking* has dispositional properties, but it is difficult to see how a disposition that does not manifest itself could actually explain its manifestation qua explanandum. It should be noted that this problem occurs *even if* we would know the laws of interaction and composition. In fact, knowing them would help in establishing that a disposition does not explain.

But one could also object that the disposition identified above is not the correct one. The disposition is not a property of cities, but of *preferences*. In particular, we are learning that moderate preference for like neighbours have the 'power' to bring about segregation.

HAE 2-checkerboard
There is residential segregation because preferences are disposed to bring about segregation when they are moderately for like neighbours.

Although I believe this in fact better represents the contribution of the checkerboard model, this does not make a difference to the HAE interpretation. The counterfactual dependence test yields the same result: were people to not have moderate preferences for like neighbours, there would still be segregation. One might here object that it is because of the result's robustness. Regardless of what the preferences are, they are disposed to bring about segregation. This is certainly an interesting result, but one that goes way beyond the checkerboard model qua toy model. The model is taken to show that moderate preferences for like neighbours can bring about segregation, not that a wide range of preferences have a similar disposition. In any case, the disposition the model would propose as an explanation for segregation would be quite different than the one I and other participants in the debate have analysed. Moreover, as Muldoon et al. (2012) point out, 'extreme robustness' in fact indicates that the model may fail to capture the actual dynamics of segregation.

Let us take stock. Here, I have argued that the disposition fails to explain in the HAE sense. This is because the explanandum is not actually *because* of the disposition. Although related to Reutlinger et al.'s point that we need to know the laws of interaction and composition, the issue I am raising is slightly different. Even without having detailed knowledge of these laws, knowing that the phenomenon does not counterfactually depend on the disposition is sufficient to reject the HAE interpretation. And we can know that the disposition fails the counterfactual dependence test without knowing the laws of interaction and composition. These may be useful in establishing whether or not a disposition manifests itself and thus make a difference to the explanandum, but that information is not always necessary. We have a true disposition that does not provide an HAE. In the next section, I argue that it provides an HPE.

5 True disposition, but a how-possibly explanation

According to Reutlinger et al's (2018) framework, rejecting the HAE interpretation implies that the veridicality condition is not satisfied. Recall that the veridicality condition requires the truth of the explanans. As the preceding discussion already indicates, I believe Nguyen (2020) is right to say that some autonomous toy models allow us to make true inferences about their target systems and thus accurately represent them. The reason why some of these do not provide HAEs is not, I contend, because the veridicality condition is not satisfied. We have good reasons to believe that the disposition ascription is true. Rather, the problem is with the explanatory relation: it is not *actual*. And although the explanatory relation is not actual, it is *possible*. Accordingly, I side with Reutlinger et al.'s interpretation that the checkerboard model provides not an HAE, but an HPE. So we have a particular case of a model that provides an HPE *and* which satisfies the veridicality condition. Using the characterization of HPEs in Sect. 3, a dispositional HPE has the following form.

Dispositional HPE
◊ (*p* because [*q* is disposed to *p* when *C*])

A dispositional HPE is a possible explanation that cites a disposition. There are two key features to note. As we have seen above, HPEs have a truth value. That an explanation is indeed possible or not can be true or false. The other is that while the possibility operator scopes over the whole explanation, in fact only some parts might be possible. As Verreault-Julien (2019) argues, sometimes HPEs will have an actual explanandum, but a merely possible explanans, but other times it may have an actual explanans, but a possible explanandum, or an actual generalization, but only possible initial conditions.[9] The checkerboard model showcases one underappreciated possibility, viz. a possible explanation with an actual explanandum and an actual explanans, but a merely possible 'because' relation. What would be the checkerboard HPE?[10]

HPE checkerboard
◊ (There is residential segregation because preferences are disposed to bring about segregation when they are moderately for like neighbours.)

This HPE says that it is possible that there is residential segregation because of the dispositional property of preferences of bringing about segregation when they are moderately for like neighbours. The purported HAE fails the counterfactual dependence test; had the preferences been different, it would not have made a difference to residential segregation. Hence, segregation is not actually because of the preferences. But could it be possibly because? In any case, that the checkerboard model provides an HPE is a widespread interpretation (e.g. Aydinonat, 2007; Grüne-Yanoff, 2013; Kuorikoski & Ylikoski, 2015; Rohwer & Rice, 2013; Weisberg, 2013; Ylikoski & Aydinonat, 2014).[11] One common view is that it identifies a 'possible cause' or a 'possible mechanism' for segregation. But this view is ambiguous between possible cause qua possible explanatory relation or possible cause qua possible disposition. For instance, Aydinonat (2007, p. 440) says the following: "We know that these mechanisms exist. We know that individuals have a tendency to avoid an extreme minority status". He continues and claims that the model "alerts us to a possible 'aggregate' tendency: a possible way in which those individual mechanisms may interact in bringing about residential segregation" (p. 440). According to him, at least some tendencies (dispositions) actually exist. One way to understand his point is that the model proposes a novel, possible, tendency (disposition). Another way is that we are learning about hitherto unknown (actual) dispositional properties of preferences which may nevertheless fail to actually explain segregation. Or, consider Weisberg's view that the checkerboard model has a generalized (actual) target and does not aim at "explaining how any specific system actually works, but rather on how they might

[9] See Grüne-Yanoff (2013) for examples of HPEs that differ in where the modality is located.

[10] I follow the formulation of HAE 2-checkerboard, but nothing hinges on that choice vs. HAE 1-checkerboard.

[11] One notable exception is Sugden (2009, 2013) who defends a how-actually interpretation.

work" (2013, p. 118). The model shows that segregation "is possible when every individual has a small preference for similar neighbours and tries to satisfy this preference" (p. 119). Although preferences possibly explain segregation, it is unclear whether for Weisberg it would be the disposition itself that is possible.

The 'possibly because' clause can hold either epistemically or objectively (Sjölin Wirling & Grüne-Yanoff, 2021). Epistemically, we might be in a position where our knowledge does not rule out that segregation may actually depend on moderate preferences for like neighbours. For instance, if someone asks me why Robin developed lung cancer and I know that she was a smoker, I could answer that it may be because of her smoking. This answer is consistent with my evidence and incorporates an actual disposition to developing lung cancer, viz. smoking. But, unbeknownst to me, the cancer's actual cause might be different. In the case of the checkerboard model, the epistemic interpretation is at least doubtful. As we have seen in the previous section, there is evidence that the actual explanation of segregation has little, if nothing, to do with moderate preferences for like neighbours. Objective possibility, on the other hand, concerns the — objective — properties objects or states of affairs have independently of one's epistemic position. When I say that my mug is breakable, I am ascribing a dispositional property to my mug, not to states of my knowledge. I am saying that the mug *is* such and such. Applying this idea to preferences means that they *have* the property of being disposed to bringing about segregation when they are moderately for like neighbours. That they indeed have that property is widely accepted. In particular, the segregation result is robust to many changes of assumptions (e.g. Muldoon et al., 2012).

Saying that moderate preferences for like neighbours have that property, however, does not by itself account for the fact that segregation is *possibly because* of that disposition. One way of spelling out this claim is that in some possible world, nearby or not, preferences *would* bring about segregation when they are moderately for like neighbours. The checkerboard model arguably provides us with reasons to believe there is such a possible world where the disposition explains segregation. Commentators often interpret the model's contribution as establishing what-if counterfactuals (e.g. Kuorikoski & Ylikoski, 2015; Ylikoski & Aydinonat, 2014). Under certain circumstances, moderate preferences for like neighbours bring about segregation. If the counterfactual is true, this implies that segregation is *possibly because* of the disposition; it explains segregation in that possible world. What we export to the world is this actual property of preferences, that they have the possibility, the disposition, of bringing about segregation and thus explaining it. But whether that property actually explains it is a different, separate, question.

The crucial point to note is that accurately representing a system as having certain dispositional properties does not require that the system actually manifests them. I can accurately represent my mug as being disposed to break even though it never manifested its breakability. Likewise, we can accurately represent preferences as having dispositional properties without them typically (or ever) manifesting them. What is peculiar in the case of the checkerboard model is that the explanandum is also actual, unlike my mug which sits in one piece on my desk. An analogous case would be to explain why my mug broke even though it is not actually because of its breakability. For instance, suppose my mug is well packaged, but someone massively

smashes it. As a result, my beloved mug breaks. In this case, we would not want to explain the breakage by citing the mug's fragility. The object broke, but it did outside the usual circumstances where we consider fragility to be responsible for breakages. Instead, it is because of the massive smashing. Yet, even if the mug broke for other reasons it still had the — actual — disposition to break.

One might here object that merely referring to a possible world where the explanatory relation holds runs the risk of trivializing how dispositions may provide HPEs.[12] After all, unless some *impossibility* is involved, there will at least be one possible world where a disposition may explain its manifestation. First, while relevance is important to assess the overall epistemic contribution of an HPE, I do not think that all (true) HPEs are necessarily relevant. Some truths are irrelevant and the same goes for HPEs. Relevance judgements are in principle separate from judgements of truth. However, an account of HPEs should also have the resources to constrain the set of possible worlds in which an HPE holds. This is the role played by the possibility operator $\Diamond$. Different interpretations of the operator will put different constraints on what is, or is not, possible (Kment, 2021). Typically, scientists are interested in varieties of modality that are more restricted than metaphysical possibility, for instance nomic possibility. Possibilities can also be constrained relative to a given scientific field, for example the sets of biological and economic possibilities contain different possibilities. One may also interpret the possibility operator as being restricted to only the possible worlds that are the closest to the actual world. More generally, possibility can be understood as a graded notion and, as such, possibility claims can have differing modal force. In the case of the checkerboard model, commentators and practitioners clearly consider that the possible world is relatively close to the actual world. For one, Sugden (2000, 2011) argues that the model and the actual world are similar in relevant respects. And scientists would not conduct theoretical and empirical explorations of the explanatory relation if its possibility was very remote.

Another potential objection is that, in fact, we do not really know whether preferences really have the disposition we ascribe to them. My reply is two-fold. First, this would not be an objection against the HPE interpretation, but only against the disposition being actual. HPEs can have merely possible explanantia. Second, even if we grant that the disposition is not actual, it does not imply it is not possible. There is thus a weaker reading available; the *disposition* is possibly true. Although current accounts of representation do not explicitly accommodate the notion of accurately representing possibility, there is also no reason to believe this is not in principle possible. For instance, one way this could occur is when a disposition is epistemically possible, i.e. the disposition is not ruled out by our knowledge and evidence. In that sense, we could say that it is a true epistemic possibility which thus satisfies the veridicality condition.

We here have a case where the explanandum and the explanans are actually true and thus satisfy the veridicality condition. The model accurately represents the (actual) dispositional property of preferences for bringing about segregation when they are moderately for like neighbours. However, the explanatory relation is only

[12] I thank a reviewer for raising this point.

possibly true. Segregation is not *actually* because of the preferences, only *possibly* because of them.

6 Conclusion

A recent debate between Reutlinger et al. (2018) and Nguyen (2020) served as a frame for discussing the explanatoriness of toy models. While the former hold that some toy models do not satisfy the veridicality condition and provide how-possibly explanations, the latter holds that they accurately represent and provide how-actually explanations. Using an overlapping case in their discussion, I have argued that some toy models may accurately represent, satisfy the veridicality condition, yet only provide how-possibly explanations. More specifically, I have defended the view that a model may (1) accurately represent dispositional properties of actual targets yet (2) only provide how-possibly explanations because the dispositional properties remain unmanifested. Therefore, the explanatory relation is only *possibly* realized.

The preceding discussion has three important upshots. Firstly, insofar as both how-actually and how-possibly explanations can be true (e.g. Verreault-Julien, 2019), the veridicality condition does not demarcate them. What matters is whether the explanation itself is actual or possible. Second, and this is an underappreciated point, accurate representation is insufficient for how-actually explanation. Toy models can accurately represent actual and possible explanantia, but the explanatory relation also needs to be actual. Finally, while the current analysis remains agnostic on whether how-possibly explanations afford understanding (cf. Reutlinger et al., 2018), it suggests new areas of inquiry to the extent that how-possibly explanations may be true and accurately represent.

Acknowledgements I would like to thank audiences at the 4th SURe Workshop and CSHPS 2022 for constructive feedback on the manuscript. Special thanks also to Ylwa Sjölin Wirling and Till Grüne-Yanoff for their support as well as helpful comments.

Declarations

Conflict of interest There is no conflict of interest to report.

References

Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, *84*(3), 488–500.

Aydinonat, N. E. (2007). Models, conjectures and exploration: An analysis of Schelling's checkerboard model of residential segregation. *Journal of Economic Methodology*, *14*(4), 429–454.

Bird, A. (2007). *Nature's metaphysics: Laws and properties*. Oxford University Press.

Bobo, L., & Zubrinsky, C. L. (1996). Attitudes on residential integration: Perceived status differences, mere in-group preference, or racial prejudice?*. *Social Forces*, *74*(3), 883–909. https://doi.org/10.1093/sf/74.3.883.

Bruch, E. E., & Mare, R. D. (2006). Neighborhood choice and neighborhood change. *American Journal of Sociology*, *112*(3), 667–709. https://doi.org/10.1086/507856.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford University Press.

Charles, C. Z. (2003). The dynamics of racial residential segregation. *Annual Review of Sociology*, *29*, 167–207.

Choi, S., & Fara, M. (2021). Dispositions. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University.

Clark, W. A. V. (1992). Residential preferences and residential choices in a multiethnic context. *Demography*, *29*(3), 451–466. https://doi.org/10.2307/2061828.

Dragulescu, A., & Yakovenko, V. M. (2000). Statistical mechanics of money. *The European Physical Journal B - Condensed Matter and Complex Systems*, *17*(4), 723–729. https://doi.org/10.1007/s100510070114.

Farley, R., Fielding, E. L., & Krysan, M. (1997). The residential preferences of blacks and whites: a four-metropolis analysis. *Housing Policy Debate*, *8*(4), 763–800. https://doi.org/10.1080/10511482.1997.9521278.

Fossett, M. (2006). Ethnic preferences, social distance dynamics, and residential segregation: Theoretical explorations using simulation analysis. *The Journal of Mathematical Sociology*, *30*(3–4), 185–273. https://doi.org/10.1080/00222500500544052.

Gelfert, A. (2019). Probing possibilities: Toy Mmodels, minimal models, and exploratory models. In Á. Nepomuceno-Fernández, L. Magnani, F. J. Salguero-Lamillar, C. Barés-Gómez, & M. Fontaine (Eds.), *Model-based reasoning in Science and Technology* (pp. 3–19). Springer International Publishing.

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

Grüne-Yanoff, T. (2013). Appraising models nonrepresentationally. *Philosophy of Science*, *80*(5), 850–861. https://doi.org/10.1086/673893.

Grüne-Yanoff, T., & Verreault-Julien, P. (2021). How-possibly explanations in economics: Anything goes? *Journal of Economic Methodology*, *28*(1), 114–123. https://doi.org/10.1080/1350178X.2020.1868779.

Illari, P., & Russo, F. (2014). *Causality: Philosophical theory meets scientific practice*. Oxford University Press.

Jebeile, J. (2020). The Kac Ring or the art of making idealisations. *Foundations of Physics*, *50*(10), 1152–1170. https://doi.org/10.1007/s10701-020-00373-1.

Kment, B. (2021). Varieties of modality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University.

Krysan, M., & Farley, R. (2002). The residential preferences of Blacks: do they explain persistent segregation? *Social Forces*, *80*(3), 937–980. https://doi.org/10.1353/sof.2002.0011.

Kuorikoski, J., & Ylikoski, P. (2015). External representations and scientific understanding. *Synthese*, *192*(12), 3817–3837.

Lipton, P. (2004). *Inference to the best explanation*. Second edition. Routledge.

Luczak, J. (2017). Talk about toy models. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *57*, 1–7. https://doi.org/10.1016/j.shpsb.2016.11.002

Manley, D., & Wasserman, R. (2008). On linking dispositions and conditionals. *Mind*, *117*(465), 59–84. https://doi.org/10.1093/mind/fzn003.

Martin, C. B. (1994). Dispositions and conditionals. *The Philosophical Quarterly*, *44*(174), 1–8. https://doi.org/10.2307/2220143.

Massey, D. S., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Harvard University Press.

Mckitrick, J. (2005). Are dispositions causally relevant? *Synthese*, *144*(3), 357–371. https://doi.org/10.1007/s11229-005-5868-z.
McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A*, *16*(3), 247–273. https://doi.org/10.1016/0039-3681(85)90003-2.
Muldoon, R., Smith, T., & Weisberg, M. (2012). Segregation that no one seeks. *Philosophy of Science*, *79*(1), 38–62. https://doi.org/10.1086/663236.
Mumford, S., & Anjum, R. L. (2011). *Getting causes from powers*. Oxford University Press.
Nguyen, J. (2020). It's not a game: Accurate representation with toy models. *The British Journal for the Philosophy of Science*, *71*(3), 1013–1041. https://doi.org/10.1093/bjps/axz010.
Nolan, D. (2015). Noncausal dispositions. Noûs, 49(3),425–439.
Reutlinger, A. (2016). Is there a monist theory of causal and noncausal explanations? The counterfactual theory of scientific explanation. *Philosophy of Science*, *83*(5), 733–745. https://doi.org/10.1086/687859.
Reutlinger, A. (2018). Extending the counterfactual theory of explanation. In A. Reutlinger, & J. Saatsi (Eds.), *Explanation beyond causation: philosophical perspectives on non-causal explanations* (pp. 74–95). Oxford University Press.
Reutlinger, A., Hangleiter, D., & Hartmann, S. (2018). Understanding (with) toy models. *The British Journal for the Philosophy of Science*, *69*(4), 1069–1099. https://doi.org/10.1093/bjps/axx005.
Rohwer, Y., & Rice, C. (2013). Hypothetical pattern idealization and explanatory models. *Philosophy of Science*, *80*(3), 334–355. https://doi.org/10.1086/671399.
Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*, 143–186.
Schelling, T. C. (1978). *Micromotives and macrobehavior*. W. W. Norton & Company.
Sjölin Wirling, Y., & Grüne-Yanoff, T. (2021). Epistemic and objective possibility in science. *The British Journal for the Philosophy of Science*, 1–31. https://doi.org/10.1086/716925.
Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.
Sugden, R. (2000). Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology*, *7*(1), 1–31.
Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis*, *70*(1), 3–27. https://doi.org/10.1007/s10670-008-9134-x.
Sugden, R. (2011). Explanations in search of observations. *Biology and Philosophy*, *26*(5), 717–736.
Sugden, R. (2013). How fictional accounts can explain. *Journal of Economic Methodology*, *20*(3), 237–243. https://doi.org/10.1080/1350178X.2013.828872.
Vaesen, K., Dusseldorp, G. L., & Brandt, M. J. (2021). An emerging consensus in palaeoanthropology: demography was the main factor responsible for the disappearance of Neanderthals. *Scientific Reports*, *11*(1), 4925. https://doi.org/10.1038/s41598-021-84410-7.
Verreault-Julien, P. (2019). How could models possibly provide how-possibly explanations? *Studies in History and Philosophy of Science Part A*, *73*, 22–33. https://doi.org/10.1016/j.shpsa.2018.06.008.
Vetter, B. (2015). *Potentiality: From dispositions to modality*. Oxford University Press.
Volterra, V. (1927). Fluctuations in the abundance of a species considered mathematically. *Nature*, *119*(2983), 12–13. https://doi.org/10.1038/119012b0.
Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science*, *3*(1), 3–51. https://doi.org/10.1093/icesjms/3.1.3.
Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy*, *104*(12), 639–659.
Weisberg, M. (2013). *Simulation and similarity. Using models to understand the world*. Oxford University Press.
Woodward, J. (2003). *Making things happen. A theory of causal explanation*. Oxford University Press.
Ylikoski, P., & Aydinonat, N. E. (2014). Understanding with theoretical models. *Journal of Economic Methodology*, *21*(1), 19–36. https://doi.org/10.1080/1350178X.2014.886470.