Article:

Nguyen, James (2019) It's not a game: accurate representation with toy models. British Journal for the Philosophy of Science, 71 (3). 1013–1041. ISSN 0007-0882

https://doi.org/10.1093/bjps/axz010

It's not a game: accurate representation with toy models

James Nguyen

Forthcoming in *The British Journal for the Philosophy of Science*
https://doi.org/10.1093/bjps/axz010

Abstract

Drawing on 'interpretational' accounts of scientific representation, I argue that the use of so-called 'toy models' provides no particular philosophical puzzle. More specifically; I argue that once one gives up the idea that models are accurate representations of their targets only if they are appropriately similar, then simple and highly idealised models can be accurate in the same way that more complex models can be. Their differences turn on trading precision for generality, but, if they are appropriately interpreted, toy models should nevertheless be considered accurate representations. A corollary of my discussion is a novel way of thinking about idealisation more generally: idealised models may distort features of their targets, but they needn't misrepresent them.

Contents

1 Introduction

'Toy models' seem to pose a puzzle: they are ubiquitous in scientific practice, but are so different from the messy systems out there in the real world that we are ultimately interested in. How are we supposed to learn anything about complex real systems by investigating simple and highly idealised models? In this paper I argue that this only appears problematic if one thinks that accurate representations have to be, in some sense, be similar to their targets. Once this assumption is dropped, and there are good reasons to drop it, the puzzle dissolves. I argue that toy models can be understood as accurate representations in much the same way as more complex models are accurate. I further suggest that the epistemic status of toy models is better understood in terms of a trade-off between precision and generality.

In Section 2 I clarify what I mean by 'toy model'. Following Reutlinger et al. [2017] I characterise them as models that are simple and highly idealised, yet represent actual target systems. I argue that such models should be distinguished from 'substitute', 'targetless', and 'minimal' models (at least in the sense of [Grüne-Yanoff, 2009]). In Section 3 I briefly present the, by now familiar, case studies utilised throughout the paper: the Lotka–Volterra model of predation; Schelling's model of social segregation; and Akerlof's model of information asymmetry. In Section 4 I argue, contra the likes of Giere [2004, 2010]; Mäki [2009, 2011]; and Weisberg [2013], that similarity (structural or otherwise) is not a necessary

condition on accurate representation, and briefly outline an alternative family of accounts of scientific representation that allow for accurate representation without similarity [Frigg, 2010b; Frigg and Nguyen, 2016a, 2018; Hughes, 1997; Suárez, 2004, 2015]. I illustrate this with the models presented previously. I argue that each of them licence truths about their targets, and are therefore accurate, despite their highly idealised nature. In Section 5 I argue that toy models trade precision for generality in the sense of [Levins, 1966]. They needn't be taken to licence any falsehoods about their target systems, although they may fail to represent (but not misrepresent) certain target features and/or details. As such, their utility comes precisely from being able to accurately represent a large number of target systems. Section 6 concludes.

Before moving on, it's worth noting that although my focus is on toy models, my arguments concern the role of idealisation in science more generally. I choose to frame the discussion in terms of toy models because they are exemplars of heavy idealisation. Thus, they provide the concrete examples with which I illustrate my claims. If preferred, this paper can be read as connecting recent debates on scientific representation with what I take to be mistaken assumptions regarding idealisation as misrepresentation. With that in mind, let's begin.

2 Characterising 'toy models'

Following Reutlinger et al. [2017] I take toy models to be characterised by the following three criteria, toy models are: (i) extremely simple; (ii) highly idealised; and (iii) nevertheless represent some target system(s) in the world. The Lotka–Voleterra model is just a coupled pair of first-order differential equations and represent fish in the Adriatic sea (amongst others); Schelling's model can be implemented by hand on a chequerboard, yet it represents a wide variety of segregated social systems; and Akerlof's market for lemons was introduced as a 'finger exercise' [Akerlof, 1970, p. 498] and yet it offers deep insight into markets where the actors have asymmetric information. Each of these models clearly satisfy (i)-(iii).

I do not offer these conditions as sharply delineating toy models from other models used in the natural and social sciences; such a delineation would be neither accurate nor useful. Some models appear more 'playful' or 'toylike' than others, but this comes in degrees. Moreover, I take it that models can cross the boundary: they can be introduced as 'serious' models, and then as the science develops and they are superseded by more complex successors, reach the status of toy models introduced in university classrooms and lecture theatres. As such, criteria (i) and (ii) are supposed to capture aspects of models that come in degrees, and moreover the extent to which the conditions are met is sensitive to the context in which the models are used. From one perspective a model may appear simple and idealised, from another it may not. However, any analysis of toy models should start with paradigmatic examples, like those discussed here, since understanding how they work should shed light on the concept, even if the concept has no sharp boundary.

With this in mind we can distinguish toy models in the preceding sense from related kinds of models prevalent the philosophical literature. Models that exhibit features (i)–(iii) can be distinguished from both 'targetless' (or 'substitute') models and Grüne-Yanoff's 'minimal models'.

Condition (iii) clearly distinguishes toy models from 'targetless' [Frigg and Nguyen, 2017, p. 54] or 'substitute' [Mäki, 2009, p. 36] models. A model may lack a target in one of two pertinent ways. First, a model may be *intended* to represent some actual system, but no such system may exist. Such examples can be drawn from the history of science—for example, models of phlogiston or ether—or one might consider architectural plans of buildings that are never built as this sort of model. Alternatively, a model may not even be offered with the intention of representing any actual target system.[1] Examples include the likes of Fisher's n-sex models in population biology, for $n > 2$ [discussed in Weisberg, 2013, Chapter 7], or Norton's Dome in Newtonian mechanics [Norton, 2003].[2] Such models are explicitly not directed at any (actual) target system, and thus their epistemic role should be distinguished from models involving target-directed reasoning. Models of this kind are better understood as exploring the implications and commitments of the theories in which they are embedded (which is not to say that models with targets don't also do this). As such, whilst there may be simple highly idealised models which lack target systems—although one might ask how a model can be heavily idealised without reference to a target system to compare it too—they do not qualify as toy models for my current purposes.

[1] A complication arises here: one might think that these models do have targets, but they are non-actual [Suárez, 2004, 2015]. This is a subtle issue that I set aside here.

[2] Mäki [2009] also considers criticisms aimed at economic theorists who construct models without regard for anything they tell us about actual economic systems.

It's important to note that in characterising 'toy models' in this way I use the phrase differently to Frigg and Hartmann [2018, Sec. 4.2] who explicitly use it to refer to models that 'do not perform a representational function' but rather those whose purpose is to 'test new theoretical tools', and Luczak [2017] who also characterises 'toy models' in this way. I take it there is no important disagreement here, these authors just use the phrase 'toy model' in a different sense to how it is used by those investigating highly idealised simply models of actual targets (such as Reutlinger et al. [2017] and myself), and as such are simply talking about different kinds of models, whilst also using different examples. However, it is worth bearing in mind that one reason to think that a model isn't supposed to perform a representational function is because it is so simple, so idealised, that it couldn't possibly hope to succeed in representing its target accurately. My argument in this paper provides the resources to avoid this way of thinking (although of course I allow that there are some models that are better understood as playing a purely theoretical, non-target-directed, role, I just don't consider them 'toy models' as I use the phrase here).

Secondly, toy models should be distinguished from 'minimal models', at least as characterised by Grüne-Yanoff [2009].[3] His central focus is how we can learn from scientific models, predominantly from economics, which are offered without concern as to whether they satisfy 'world-linking' properties such as being similar, or appropriately morphic, to their targets, or adhering to regularities about the world [Grüne-Yanoff, 2009, Section 4]. He argues that '[i]f we are to learn from [such] a model [...] it must (1) present a relevant possibility that (2) contradicts an impossibly hypothesis that is held with sufficiently high confidence by the potential learners' (p. 97). Minimal models are those that are offered with the intention of meeting these conditions, despite not relating to their targets in any of the preceding world-linking ways. A central point I argue for in this paper is that toy models also allow for learning without world-linking properties (or at least world-linking properties such as similarity, structural or otherwise). And I agree that that by providing a how-possibly explanation, which is in some sense surprising in light of background beliefs, is one way that toy models inform us about their target systems. But I don't think it's the only way. As discussed later, the Lotka–Volterra model doesn't contradict any impossibility thesis that is, or was, held with sufficiently high confidence by potential learners, and as such it doesn't meet Grüne-Yanoff's condition (2).[4] In fact, it wasn't even offered with the intention of meeting such a condition (given the data that were already available to Volterra when he constructed it). But we still learn from the model. Moreover, although Grüne-Yanoff doesn't utilise this terminology, it's plausible that all of the models I discuss in this paper go beyond how-possibly to how-actually explanations of the behaviours of target systems (more on this in subsection 4.2). As such, I think it's beneficial to think about 'toy models' in a broad enough sense to capture models that do more than provide surprising how-possible explanations.[5]

In sum. By 'toy models' in this paper I mean models that are (i) highly simple; (ii) highly idealised when compared to; (iii) the (actual) target systems that they represent. Condition (iii) immediately distinguishes them from 'targetless' or 'substitute' models. And although the sorts of models discussed by Grüne-Yanoff might satisfy these conditions, I have a broader notion in mind that his 'minimal models', which provide how-possibly explanations contradicting a previously held impossibility hypothesis.

3 Cases

In this section I briefly outline the Lotka–Volerra model of predation [Volterra, 1926, 1928][6]; the Schelling model of social segregation [Schelling, 1971, 1978]; and Akerlof's 'market for lemons' model of the impact asymmetric information can have on markets [Akerlof, 1970]. I should stress that these examples are well-trodden (possibly over-trodden) in the philosophical literature.[7] I choose them to illustrate my claims precisely because of their familiarity, rather than to shed greater insight on the particular models themselves. Having said that, if what I argue in the Section 4 is true, I hope to provide a lens through which the beauty, and utility, of these models can be better appreciated.

[3] Batterman and Rice [2014] also use the phrase 'minimal model' but in a different sense to Grüne-Yanoff. I briefly discuss their account in footnote 25.

[4] Fumagalli [2016] makes a similar point about Grüne-Yanoff's use of Schelling's model to illustrate his claim.

[5] It should be clear that my account here isn't that toy models cannot exhibit some of the features associated with Grüne-Yanoff's account. Rather I'm claiming that the concept TOY MODEL shouldn't be defined by the conditions he offers.

[6] It's worth noting that Lotka [1925] developed the model independently. For my current purposes I focus on Volterra's presentation, but see [Knuuttila and Loettgers, 2016] for a philosophical discussion comparing their approaches.

[7] For philosophical discussions of Lotka–Volterra see [Knuuttila and Loettgers, 2016; Weisberg, 2007b, 2013]; Schelling see [Mäki, 2009; Sugden, 2000]; and Akerlof see [Sugden, 2000].

3.1 Lotka–Volterra

The Lotka–Volterra model consists of the following pair of coupled differential equations:

$$\frac{dN_1}{dt} = (\epsilon_1 - \gamma_1 N_2)N_1 \tag{1}$$

$$\frac{dN_2}{dt} = (\gamma_2\gamma_1 N_1 - \epsilon_2)N_2 \tag{2}$$

where t denotes time; N_1 denotes the size of the prey population; N_2 denotes the size of the predator population; ϵ_1 is a coefficient measuring the intrinsic growth rate of the prey; ϵ_2 the intrinsic death rate of the predators; γ_1 measures how efficient the predators are at capturing prey; and γ_2 how efficient the predators are at converting eaten prey into new predators.

Although the model contains a number of interesting features, Volterra took the following to be 'the most interesting of all' [1926, p. 559]:

> **Law of the disturbance of the averages**: 'If an attempt is made to destroy the individuals of the two species uniformly and in proportion to their number, the average of the number of individuals of the species that is eaten increases and that of the individuals of the species feeding upon the other diminishes' [1928, p. 20].

This can be derived as follows. First we note that equations 1 and 2 have no stable solutions: they oscillate indefinitely. However, for fixed ϵ_1, ϵ_2, γ_1, and γ_2 the solution where $\frac{dN_1}{dt} = \frac{dN_2}{dt} = 0$ and where $N_1 \neq 0$ and $N_2 \neq 0$, corresponds to the time averaged size of N_1 and N_2, denoted $\hat{N_1}$ and $\hat{N_2}$, with the fixed parameter values. From this we can derive:

$$\hat{N_2} = \frac{\epsilon_1}{\gamma_1} \tag{3}$$

$$\hat{N_1} = \frac{\epsilon_2}{\gamma_2\gamma_1} \tag{4}$$

and, comparing the ratio of $\hat{N_1}$ to $\hat{N_2}$, we note that a biocide—an 'attempt to destroy the individuals of the two species uniformly in proportion to their number'—corresponds to a *decrease* in ϵ_1 (the growth rate of the prey), and an *increase* in ϵ_2 (the death rate of the predators). Which corresponds to an increase in the ratio of $\hat{N_1}$ (the time averaged size of the prey population) with respect to $\hat{N_2}$ (the time averaged size of the predator population), as stated by the law of the disturbance of averages.

The motivation for the model was as follows. D'Ancona, Volterra's son-in-law, had been gathering data concerning the number of selachians (sharks), and 'food-fish' (shark prey) in the Italian fish markets, and asked Volterra to explain the patterns he had discovered. So Volterra's model had a specific target: fish in the Adriatic sea in those time periods. What D'Ancona's data showed, and Volterra's model predicted, was that the decreased fishing during the First World War acted as a form of 'protection', and thus outside of the war period, where fishing acted as a biocide, there was a higher proportion of prey-fish in the fish markets [Volterra, 1926, p. 559].

So fishing, as a form of biocide, favoured the prey. This demonstrates that the Lotka–Volterra model, at least as Volterra used it, had a target system: the fish in the Adriatic sea before, during, and after, the war.[8] Moreover, Volterra was aware of the statistics from the fisheries before constructing his model. As such, although a toy model, the Lotka–Volterra model does not provide a novel possibility result in the sense of [Grüne-Yanoff, 2009].[9]

It should also be clear that this model (and this use in particular) is simple. The derivation of the law of the disturbance of averages does not take too much mathematical sophistication. Moreover, the model is highly idealised. As noted by Volterra the size of the model populations are measured by real, rather than integer, values; births take place continuously; and each species is taken to be homogeneous, for example ignoring variations of age and size [1928, p. 6]. Additionally, the model is stated in purely aggregate terms: no reference is made to the individual make-up of the populations, nor to the details of the physical theatre in which the predation plays out. So the Lotka–Volterra model meets conditions (i)-(iii) above.

[8] As I discuss in Section 5, the model, even as introduced by Volterra, has a number of target systems, but this can be seen as the most pertinent one for my purposes in this section.

[9] In a sense the law of the disturbance of averages is surprising, but this surprise isn't generated by a model result. Rather, the model accounts for an already known fact.

3.2 Schelling

Schelling's model, originally introduced in [Schelling, 1971] and further discussed in his *Micromotives and Macrobehaviour* [1978] consists of the following: an eight by eight lattice; two types elements (in Schelling's presentation, dimes/#s, and pennies/Os); and a way of selecting squares at random. On a lattice containing #s and Os we define a dynamics: an element is 'content' if $\frac{1}{3}$ of the other elements in its Moore neighbourhood are of the same kind, and discontent otherwise. If an element is content then it stays where it is, otherwise it moves to one of the nearest squares that makes it content. The order of moves is not too important for the dynamics, but we can suppose that we start in the upper left of the grid, and sweep to the bottom right. Having defined the above, Schelling [1978, p. 149-150] distributes some elements across the board, leaving some squares empty. He identifies which elements are discontent and moves them. After every discontent element has been moved the process starts again. The result is that, almost irregardless of the initial distribution of #s and Os on the board and even weakening the 'content' requirement below $\frac{1}{3}$, once the system reaches a static state that state is highly segregated with islands of #s and Os clustered together.[10]

Again the model is very simple. It can be run manually 'by any reader with a half-hour to spare, a roll of pennies and a roll of dimes, a tabletop, a large sheet of paper, a spirit of scientific inquiry, or lacking that spirit, a fondness for games' [Schelling, 1978, p. 147]. It also has a clear target: residential segregation according to colour in the United States.[11] And with respect to that target system, the model is highly idealised: it doesn't take into account the costs of moving; cities do not have 64 houses; whether or not an element in the model is content is defined solely in terms of the make up of its immediate neighbourhood; when moving the element moves to an empty square at random; and so on. So again, the model meets (i)-(iii) above, and is a paradigmatic example of a toy model.

3.3 Akerlof

Akerlof's 'market for lemons' model introduced the concept of 'asymmetric information' into economic theory. The model involves two groups of traders with the following utility functions:

$$U_1 = M + \sum_{i=1}^{n} x_i, \tag{5}$$

$$U_2 = M + \sum_{i=1}^{n} \frac{3}{2} x_i, \tag{6}$$

where M is the consumption of goods other than cars (whose price is assumed to be unity), x_i is the quality of the ith car, and n is the number of cars in the market. So, for any car, its monetary price is higher for members of group 2 compared to members of group 1; members of the second group would gain more utility from the car than members of the first group. These functions already introduce significant idealisations into the model. There are only two types of traders and, as Akerlof notes, that the utility functions are linear 'allows a focus on the effects of asymmetry of information' [1970, p. 491] without getting us 'needlessly mired in algebraic complication' (p. 490), and both functions are such that 'the addition of a second car, or indeed a kth car adds the same amount of utility as the first' (p. 491). Such assumptions are clearly distortions of any actual used-car market.

We assume that both groups of traders are utility maximisers, and that group 1 has n cars with quality x distributed uniformly between $0 \leq x \leq 2$. Akerlof assumes, given the lack of information available to the buyers, that there is a single market price p for cars and that this is a function of μ, the average quality of the cars (this partly encodes the asymmetric information; it's 'the best' the buyer can offer given their lack of information about the quality of an individual car). From this he derives the demand and supply functions of groups 1 and 2, which can be summed to deliver the following total demand (where Y_i is the income of group $i \in \{1, 2\}$):

$$D(p, \mu) = \begin{cases} (Y_1 + Y_2)p & \text{if } p < \mu, \\ Y_2/p & \text{if } \mu < p < 3\mu/2, \\ 0 & \text{if } p > 3\mu/2, \end{cases} \tag{7}$$

[10] The model has been implemented in NetLogo allowing variations in the dynamics and initial conditions. See `http://ccl.northwestern.edu/netlogo/models/Segregation`.

[11] As I discuss in Section 5, the model has a number of other target systems as well.

where, crucially, for every price p, $\mu = p/2$ (this follows from the fact that the supply function for traders of the first group is given by $S_1 = pN/2$ where $p \leq 2$, with average quality $\mu = p/2$, which encodes the fact that buyers won't supply their cars if they are higher than average quality thus reflecting that they have information unavailable to the buyers). But this means that total demand $D(p, \mu) = 0$, which means that no trade takes place, even though at any given price (within certain limits), there are traders from group 1 who are willing to sell their car at a price which traders from group 2 are willing to pay. Akerlof goes on to show what would happen if information is symmetric where trades do take place and a Pareto-efficient equilibrium is reached.

Again, the model is simple (recall that Akerlof introduces it as a 'finger exercise'), and moreover has a target system. Akerlof motivates his model as follows:

> 'The example of used cars captures the essence of the problem [asymmetric information]. From time to time one hears either mention of or surprise at the large price difference between new cars and those which have just left the showroom. The usual lunch table justification for this phenomenon is the pure joy of owning a "new" car. We offer a different explanation' [1970, p. 489].

Although he immediately goes onto to introducing the model, this demonstrates that he explicitly takes his model to be directed towards actual car markets. So again, the model satisfies our conditions.

4 Accurate representation without similarity

The crucial claim of this section is that toy models, like the ones just outlined, can be accurate representations despite, or even in virtue of, their simple and highly idealised nature. What underpins this view is an approach to scientific representation that emphasises the way in which scientific models are interpreted as licensing inferences about their target systems. Although (proposed) similarity relations are one way that scientific models are interpreted, they are not the only way. As such, a model can licence truths, despite failing to be similar to its target in crucial respects. I first briefly recap the relevant literature on scientific representation with a particular focus on how it allows for the conceptual possibility of accurate representation without similarity. I then apply these insights to the models outlined above and preempt some objections to my account.

It's commonplace in the literature on scientific representation to distinguish between representation simpliciter and accurate representation.[12] The reason for this is clear: models can represent their target systems, but inaccurately. Presumably an analysis of scientific representation is supposed to cover such models, hence, representation and accurate representation should be distinguished. The idea that similarity, including at the level of shared structure, plays a constitutive role in establishing representation has faced sharp criticism [Frigg, 2006; Suárez, 2003]. For my current purposes what's important is that in order to accommodate cases of misrepresentation in certain respects, one shouldn't require that the model and its targets are actually similar in these respects at pain of conflating representation simpliciter with accurate representation.

At least partly in response to these criticisms, proponents of accounts involving similarity have retreated to the idea that proposed similarity relations establish scientific representation, and if those relations hold, then the representational relationship is accurate in those respects (see, for instance [Mäki, 2009, pp. 32–33, p. 41], [Giere, 2004, 2010], and [Weisberg, 2013, Chapter 8]).[13] The underlying idea driving such accounts is that similarity, albeit caveated in certain ways by invoking pragmatic constraints and a restriction to only certain features, is, everything else equal, the way in which models accurately represent those features of their target systems.

The relevant question now, is what it means for a model and a target to be 'similar' with respect to some feature r (where by 'feature' I mean the instantiation of a particular property or relation, or structural feature more generally)? Khosrowi [2018], in discussing [Weisberg, 2013] provides three possible interpretations: either the model and the target share the exact same feature; they have features that are 'quantitatively close' to one another; or they have features which are themselves 'sufficiently similar' to one another (see also [Frigg and Nguyen, 2017, p. 64]). So if a model M has some feature r, then according to the first way of explicating similarity, in order for M to accurately represent a target T with respect to r, T must itself have r. According to the second, T must have some feature r' such that

[12] For a short introduction see [Frigg and Nguyen, 2016b]. For a longer one see [Frigg and Nguyen, 2017].

[13] I take it that the view is also implicitly assumed in many discussions of highly idealised models, including Reutlinger et al. [2017]'s discussion of whether toy models satisfy what they call the 'veridicality condition'.

r and r' are quantitative features (for example, they might both correspond to a parameter taking values $r, r' \in \mathbb{R}$ in the model and target respectively), and $|r - r'| < \epsilon$ for some small ϵ, presumably delivered by the context in which the model is used. According to the third, T must have some feature r' such that r and r' are themselves 'sufficiently similar' (again, presumably where what counts as 'sufficient' is provided by context, but I am not aware of an existing explication of what it means for features to be 'similar' to one another). So, according to similarity-based accounts, if M has a feature r, which represents some features r' of T (in virtue of a model user proposing that M and T are similar with respect to it), then for M to accurately represent T with respect to that feature, it must be the case either that $r = r', |r - r'| < \epsilon$, or r and r' are 'sufficiently similar'. This accounts for how we reason using models, according to the similarity-based understanding of how they represent. We observe some relevant feature of the model, and export either it, or something sufficiently like it, to the target. If they are so similar then representation is accurate in this respect and the model user has succeeded in learning about that specific feature of the target.

But from this perspective, toy models appear puzzling: given their simple and idealised nature, they are not very similar to their target systems in the sense that they do not share very many (similar) features with them. As such, they cannot be very accurate representations of those systems. But then why are they so prevalent across the natural and social sciences?

Giving up on the idea that accurate representation must involve similarity dissolves this puzzle. For the purposes of this paper we can call accounts of scientific representation that allow us to do this 'interpretational' accounts [*cf.* Nguyen, 2017, pp. 983–984]. Suárez [2004, 2015] suggests that a model represents its target only if the 'representational force' of the model points towards the target, and the former allows 'competent and informed agents' to draw inferences about the latter. Hughes [1997] argues that models denote their targets, are used to perform demonstrations, the results of which can be interpreted in terms of the denoted target system. Frigg [2010a] agrees that models denote, but argues that they represent them in virtue of the existence of a 'key' that translates model-facts to claims to be exported to a target system. Frigg and Nguyen [2016a, 2018] build on this by arguing that the model facts to which the key applies are those that are 'exemplified' in the sense of [Goodman, 1976] and [Elgin, 2010, 2017]. Although there are significant differences between these accounts, for my current purposes what's important is that they all agree that the primary purpose of scientific models—at least as they are used representationally—is to licence inferences about their targets, and the way they do this is a function on model-facts (Hughes' demonstration, the premises of Suárez's inferences, and the arguments of Frigg and Nguyen's keys) combined with intentional acts of model users interpreting these facts in terms of their target systems (Hughes' interpretations, Suárez's inferential schema used, and Frigg and Nguyen's keys). In the abstract such accounts propose that there are 'interpretation functions' associated with models, and these functions map model-facts to claims—which may be true or false, in order to accommodate misrepresentation—to be exported to their targets.

The way that the interpretational accounts go beyond those that invoke similarity should be clear: although the functions associated with models may map some feature r of a model to the same feature r, or a (possibly quantitatively) 'similar feature' r', to be exported to the target, this isn't required in order to set up the representational relationship, the relationship between model features and features to be exported to the target can be conventional instead. As such, the model can generate true claims about a target system, and thereby accurately represent said system, despite failing to share any relevant feature with its target, and these claims are part of the very 'representational content' of the model.[14]

4.1 Pictorial and cartographic representation

Now although my aim in this paper is not to defend interpretational accounts of representation—such defences can be found in the work cited previously and my task here is rather to highlight a downstream benefit of adopting them—it is worth seeing them in action in some familiar cases before applying them to toy models. To do this I follow the lead of the likes of Elgin [2010] and French [2003] by turning to an example of pictorial representation. Consider the two pictures of Barack Obama in Figures 1 and 2 [*cf.* Goodman, 1976, pp. 35–36]. Figure 1 is a standard monochrome picture of the former POTUS. Figure

[14]The above discussion has been framed in terms of what it means for a model, M, to accurately represent its target, T, with respect to a *particular* feature. In general, we can say that M accurately represents T to the degree that it accurately represents its relevant features $r_1, ..., r_n$. Here which features count as relevant, and how accurate representation of one feature should be weighted against accurate representation of another, will be sensitive to the purposes and context in which a model is used. But this pragmatic element should be expected: part of what is interesting about model-based science is its contextual nature, a model which is accurate with respect to one purpose may not be accurate with respect to another [*cf.* Teller, 2001].

2 has been constructed by taking Figure 1 and inverting the colours: white gets mapped to black, and visa versa, dark grey is mapped to light grey, and so on.

[Insert Figures 1,2 about here]

I take it for granted that Figure 1 is more similar to Obama himself (in the sense that it has strictly more features in common with Obama).[15] The question then, is which image is a more accurate representation? Neither! Both images contain the exact same representational content. The interpretation function associated with Figure 1 utilises similarity relations to generate claims about Obama. For example, that a certain area of the picture is dark and a certain area light allows us to infer that Obama was wearing a dark suit and light shirt when sitting for the photograph. But the exact same claim can be generated by Figure 2, by starting with the picture-fact that an area of the photograph is light and mapping this to the claim that Obama was wearing a dark suit, and an area of the photograph is dark and mapping that to the claim that Obama was wearing a light shirt.

The sort of interpretation that we apply to the former (where colours are mapped to themselves) is more entrenched than the sort we apply to the latter once we were aware of the colour inversion. But increased entrenchment doesn't make the representation any more accurate. As long as a viewer knows how the picture was created, and therefore utilises an inverting interpretation function when transforming picture colours to colours to be exported onto Obama, Figures 1 and 2 generate the exact same claims about the former POTUS, despite their differences with respect to how similar they are to him.

Now one might be tempted to object and say that what underpins Figure 2's accuracy is still similarity in at least two possible senses.[16] First, one might argue that it's similarity with respect to some more abstract feature, a structural one for instance. The problem with this approach is that it doesn't accommodate the particular claims that we are able to generate about the target. We take into account a particular feature of the figure—r_{light} : that a certain area of the figure is lightly shaded—and interpret that in terms of a different feature—r_{dark}: that a particular area of this clothing is dark—to be exported to Obama. We explicitly exploit a particular, non-structural, dissimilarity between the two, r_{light} and r_{dark}, to generate the accurate, particular, and non-structural, claim. Recasting this as the exploitation of structural similarities fails to account for why both figures, suitably interpreted, represent Obama as wearing a dark suit and light shirt, rather than a light suit and dark shirt, given that there is no structural difference between the cases.[17]

An alternative option would be to claim that Figure 2 is just as similar to Obama as Figure 1, even with respect to particular colour features, it's just that what we compare with respect to similarity isn't Obama and the figure itself, but Obama and the figure-under-the-interpretation-that-inverts-colour. When this is the relevant comparison the figures are indeed equally similar to Obama, precisely because the same comparison is being made in both instances. However, this sense of similarity is not the sense that is assumed by similarity-based accounts of scientific representation. The representation and the target are no longer required to share features (or have features which are sufficiently similar/close to one another). The representation is required to have a feature r, such that there is an interpretation function f that takes r to a feature $f(r)$, and it is $f(r)$ that is compared to the features of the target with respect to similarity. But then the crucial question becomes how f works. Given that it is no longer required to involve any notion of similarity (recall that in the case in question it would map white-to-black and black-to-white), we return to a version of the interpretational accounts where the interpretation functions needn't exploit similarities.[18]

To take another example consider a map of the Earth's surface constructed using the Mercator projection. It's well-known that, given mathematical limitations of projections from S^2 to $\mathbb{R}^2$, there is no way to construct a map that preserves all metric properties of the Earth's surface. The Mercator projection is conformal (preserves angles) but it distorts the relative size of regions (in particular, as one moves further from the equator on the North-South axis, larger and larger areas of the map represent the same sized area on the Earth's surface). One way of thinking about how such a map represents the Earth's surface is as follows: the map and the Earth's surface are similar with respect to angles, but dissimilar

[15] This point could be made using coloured images, but I take it that it's clear enough in the current context.

[16] I'm grateful to two anonymous referees for encouraging me to be explicit about this.

[17] To be even more explicit about this, if one were to interpret the inverted image using a similarity-based interpretation, then the structural content of the representation would remain the same, but the result would be an inaccurate representation with respect to the particular colours of Obama's clothing.

[18] This approach does differ slightly from the interpretation-based accounts as outlined above. For them the question of representational accuracy turned on whether the interpretation function delivered a feature the target has, under this approach it could turn on whether it delivered a feature sufficiently similar/close to what the target has. Here is not the place to explore this further since for my current purposes what's important is that it would still rely on non-similarity-based interpretation functions.

with respect to their area properties. On such a reading the map may be taken to accurately represent bearings, but to misrepresent the relative area comparisons between, say, Greenland and continental Africa.

But the map does not have to be read this way. Features like 'being of equal area' on the map, don't have to be interpreted as representing 'being of equal area' on the Earth's surface. In fact, if one had a sufficiently good understanding of the projection used to create the map, then one could provide an interpretation function that delivered truths about area properties of the Earth, despite the dissimilarities between these and the area properties of the map.

This can be illustrated by overlaying the map with a 'Tissot indicatrix' as in Figure 3. The circular areas on the map indicate map-areas that represent the same sized area on the Earth's surface. With this in mind, a map reader could calculate, using the indicatrix, or their knowledge of the details of the map projection, the correct relative areas of Greenland and continental Africa, despite the fact that this is not similar to the relative difference between the regions on the map representing the respective landmasses. By properly interpreting the map, one can arrive at accurate representation in a way that explicitly does not rely on similarity relations between the representation and represented. Under such an interpretation, the Mercator projection may distort the area of landmasses, but that doesn't entail that it misrepresents them.

[Insert Figure 3 about here]

Now, again, someone might object and argue that it's still a similarity underpinning the Mercator projection's accurate representation of the landmasses on the Earth (even with respect to their relative area).[19] Vindicating this claim would require finding a sufficiently sensitive shared (possibly structural) feature according to which the two are similar. An immediate restriction is that this feature won't be structural in the sense relevant to the metrical properties of the map (specifically, with respect to areas and angles). No projection can preserve this structure, and the projection in question explicitly doesn't preserve area, despite, in this instance, being used to represent the very metric structure that it fails to preserve.

But perhaps another feature could be found, perhaps a feature such as 'the-ratio-of-area-to-overlaid-Tissot-circle' in the map and the target.[20] The worry with this approach is the following. First, this suggested feature doesn't do the work required in cases where someone sufficiently familiar with the Mercator projection reasons about landmasses using the projection itself, without relying on the circles in Figure 3. The use of the overlaid Tissot indicatrix is illustrative; it is not required in order to generate the inference from the fact that map-Greenland is the same size as map-Africa to the claim that Greenland is, in fact, smaller than Africa (anyone who knows about the distortions present in the Mercator projection reasons that way, even without familiarity with the indicatrix). Second, such an approach involves searching for any kind of gerrymandered shared (or similar) feature to account for successful inferences that exploit dissimilarities. This robs the similarity-based approach of its attractiveness. Recall the idea expressed previously: according to the approach one can reason about a target with a representation of it by noticing that the representation has a feature and then inferring the target does too. This connects the representational content of the representation with the inferences it licences. In cases like the ones under consideration, the shared feature, if it can be found at all, is offered *ex post*, after the result of the reasoning has been judged accurate. This fails to explicate what it was about the representation that underpinned the successful inferences it generated; with 'similarity' being tacked-on only after the inferences are deemed successful. As such, it fails to account for how we use representations to reason about their targets. Thus, it seems like interpretations that do not exploit similarity relations are to be preferred for explicating how the examples in this section represent, which demonstrates how representations can be accurate in certain respects, without being similar to their targets in those respects.

4.2 Model representation

So, having shown that similarity with respect to a feature isn't required for accurate representation with respect to that feature, we can now apply this sort of reasoning to the models presented in Section 3. In the Lotka–Volterra model the relevant model feature is:

[19] In this instance I set aside the idea that this is because the relevant objects being compared as similar are the map-under-an-interpretation and the Earth's surface for analogous reasons to those given previously.

[20] I'm grateful to an anonymous referee for this suggestion.

r_1**:** a decrease in ϵ_1 and an increase in ϵ_2 entails an increase in the ratio of $\hat{N}_1$ with respect to $\hat{N}_2$ (in fact, for specific values of the parameters, these changes can be precisely calculated).

But this feature needn't be exported directly to the target system (fish in the Adriatic sea in this case). Rather than taking the model to show that a biocide entails an increase in the relative size of the prey to predator population, let alone specific real valued increases and decreases, the model can be read as making the following claim about the target:

r_1'**:** fishing (as a form of biocide), which increases the death rate of the prey and decreases the growth rate of the prey, increases the prey-to-predator ratio's susceptibility to rise

There are two important ways in which the claim generated introduces dissimilarities between the model and the target system. First, in the model the biocide inevitably has such an effect, when it comes to the target system a susceptibility is proposed instead. Secondly, in the model the effect is specific, when it comes to the target system a less specific claim is generated (a shift from real-valued results in the model to qualitative trends claimed to hold in the target system).

What about the Schelling model? Here the relevant fact is:

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, an initial distribution is associated with a unique segregated outcome).

But again this feature needn't be directly exported to residential segregation by skin colour in the United States. Rather, the model can be taken to generate the following claim:

r_2'**:** a city whose residents have weak preferences regarding the skin colour of their neighbours has a susceptibility towards global segregation.

Again, the dissimilarities come in the move from the specific (the precise details of the segregation) to the less specific (no claim is made about the precise details in which individuals with different coloured skin will be residentially segregated), and the move from what almost inevitability occurs in the model to a susceptibility claim about the target system.

In Akerlof's model the relevant fact is:

r_3**:** asymmetric information prevents any car trades from occurring, despite the fact that at any given price there are sellers willing to sell their car, and buyers willing to buy it.

But clearly Akerlof didn't think that this was the case for any actual used car market (indeed if he did, then there would be no used car market to target!). Rather the model generates:

r_3'**:** Asymmetric information—in, for example, a particular car market—increases that market's susceptibility to fail to reach Pareto-efficient equilibrium.

Again, we move from a specific claim (no trade) to a less specific claim (Pareto-efficiency not being reached), and moreover from an inevitably in the model to a susceptibility in the target system.

I'm proposing that toy models are interpreted in a way that involves (i) a move from an inevitability (or something close to it in the case of Schelling's example) in the model to a susceptibility claim about the target system; and (ii) a move from a specific model-fact to a less specific claim about the target system. And these less specific susceptibility claims are literally true: a predator-prey system affected by a biocide is susceptible to some increase in the prey-predator ratio; weak individual preferences regarding neighbour similarities does lead to a city's susceptibility towards some global segregation; and asymmetric information does make a market susceptible to some non-Pareto efficient equilibrium.

Expressed like that, it might seem that the targets of toy models are general features of the world (biocides, segregation, and asymmetric information in general, what Levy [2015, Section 4.4], following Godfrey-Smith [2009, pp. 106–107] calls the 'hubs' in 'hub-and-spoke' model-based reasoning), rather than concrete target systems themselves. This isn't quite right. It's crucial to note that these susceptibility claims are true when applied to particular concrete targets.[21] r_1' tells us that the decreased fishing during the First World War favoured the sharks in the Adriatic sea; 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

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

colour, this still makes the city susceptible to global segregation; and r_3' tells us that asymmetric information in a particular market makes it susceptible to non-Pareto efficient outcomes. And again, these claims are literally true of their targets, the ratio of prey to predators in the Adriatic Sea did increase, Chicago was, and is, globally segregated, and a car's value does drop after it leaves the lot (I say more about what I mean to attribute a system a susceptibility below). So what all these examples show is that these models, although dissimilar to their target systems (in some respects radically so) and highly idealised, nevertheless accurately represent them, as long as they are suitably interpreted.

A few points of clarification are in order. The first concerns the relationship between what I have said and broadly Cartwrightian considerations regarding how to think about laws with implicit *ceteris paribus* clauses. Cartwright denies that laws of nature hold universally: in particular, she denies that 'cross-wise' inductions from the observation that they hold in controlled circumstances, laboratory environments for example, to the claims that they hold everywhere and everywhen are justified [Cartwright, 1999, Chapter 3]. The question then, is what their status is in situations beyond those controlled circumstances. One answer is that the laws describe isolated 'capacities' (or 'dispositions', 'powers', 'natures', 'tendencies', or, unsurprisingly, 'susceptibilities'). Where the capacities are isolated, such as in a model, the laws are true, but where they are not, such as in (most) target systems in the world, the laws are false or inapplicable. The preceding discussion then, in Cartwrightian terms, is that the models represent the target systems as having capacities, but capacities that are not inevitably realised in the same specific way in which they are realised in the model.

Under this reading, my account is in line with Cartwright's view, in which case the contribution of this paper is to fill a lacuna in her work involving laws, models, and capacities. By her own lights she 'has little to say about how how representative models represent' [1999, p. 191] or about the model-target relationship beyond an appeal to 'a loose notion of resemblance [which is] just to point to the problem, or to label it, rather than to say anything in solution to it' [1999, p. 192]. Thus, the argument above could be interpreted as one way in which the Cartwrightian account can be further explicated and developed by casting it in terms of scientific representation.

However, if this is the case, it remains that I am not committed to any particular account of the metaphysics of susceptibilities. It suffices for my purposes that some account could be given, and my claims concerning the value of toy models, and highly idealised models more generally, are independent of any particular metaphysics. Moreover, another way of thinking about what it means for a claim attributing a susceptibility to a target to be true is to deny that it requires any robust metaphysical truth-maker whatsoever (thus adopting a deflationary attitude towards the metaphysics of dispositional features). Under this reading, all that the claim 'asymmetric information makes a market susceptible to non-Pareto efficient equilibrium' means is that we should expect measurable parameters of the market (the sales, willingness to buy/sell, and so on) to move in the same direction as the model, not that there is some rich metaphysical story involving interacting capacities driving this result. This use of the term 'susceptibility' is therefore somewhat analogous to 'directional drift', and to adopt it is compatible with denying there is any metaphysical story to be told.

Which of these ways of thinking about what underpins the truth of a claim attributing a susceptibility to a particular target is correct goes beyond my current scope. One could adopt a particular metaphysics of dispositional properties; one could simply remain silent about the particulars but grant that *some* metaphysical story has to be told; or one could be entirely deflationary about them. Each of these approaches is compatible with the central theoretical insight of this paper: the fact that simple, highly idealised models are radically dissimilar to their targets is not problematic as long as they are interpreted appropriately.

Moving on, it is worth contrasting my suggestion here with the way in which Reutlinger et al. [2017, Section 4.2] discuss, and reject, the idea that 'autonomous toy models'—models, like those discussed in this paper, which are not embedded in a broader theory—provide how-actually explanations of the dispositions of target systems. According to them, a dispositionalist interpretation of such models involves them representing the dispositions of the target system in the absence of other factors (where the 'other factors' are those that do not feature in the model). They argue that the approach cannot work because it doesn't tell us how to interpret the model when applied to a system in which those 'other features' are in fact present, and therefore they relegate these models as providing how-possibility understanding of their targets.

As I argued in Section 2, the models discussed in this paper, whilst not being embedded in broader theories, should not be seen as merely providing how-possibility understanding. The Lotka–Volterra model does not just provide a how-possibly story about the change in predator-prey ratios; biocides do actually yield such a change. But if Reutlinger et al. are right, there is a gap in the story about how the

model is applied in cases where there are other relevant factors present (Reutlinger et al. [2017] seem to assume that in the non-ideal cases the dispositions would have to be exported identically to the target, and thus there would be a puzzle about their application in these cases, which is precisely the approach I have been arguing against). The interpretations I discuss above fill this gap. The models do still apply in the presence of other factors: it's just that the susceptibilities that are exported to the model can enter into complex relationships with other aspects of the target systems, or in extreme cases, be overridden altogether. But that doesn't mean they are not present in the systems. In cases where the susceptibilities are manifested, but interact with other aspects of the target system, the way in which I recommend thinking about exporting non-specific behaviour ensures that the model's content, when suitably interpreted, is still accurate.

What about cases where the susceptibilities are overridden entirely? Are the models still accurate when applied to them? I think the answer is yes, and that this is in fact required if we are to make sense of claims like 'the other factors overrode the system's susceptibility to behave in a certain way'. Since the sense of such claims requires that the system had a susceptibility to be overridden in the first place. For example, it is true that a smoker's smoking increases their susceptibility towards lung cancer, even if they never develop lung cancer. This is all that is required for my strategy to accommodate model accuracy in such cases.[22] Of course if one demanded that the dispositionalist strategy involved exporting an actualised disposition to the target system, with the same specific actualisation as occurs in the model, then the model would come out as inaccurate. But this is not how I am suggesting such models are, or should be, interpreted.

Returning to the question of representation without similarity, one might take my discussion of interpreting inevitabilities/specific details as representing susceptibility/non-specific claims, as tantamount to claiming that the model and target are similar with respect to those features. More precisely, if a model has some feature 'inevitably r' and a user exports a feature 'susceptibility to r' to the target, or if a model has some specific feature r and a user exports some feature r', where r' is a less specific version of r, then does this not amount to claiming that the model and the target are similar with respect to r? If so, similarity is being implicitly relied upon.

This relies on subtleties concerning the meaning of 'similarity' but the line of reasoning can be resisted in a few different ways. Firstly, even if the above objection is sound, the sort of similarity being invoked here is not mentioned in the literature. In contrast to, for example, Mäki [2011], I am not claiming that the models discussed in this paper have features that 'approximate' the actual features of their targets (which amounts to requiring shared similar features). And moreover toy models and highly idealised models in general are typically taken to be dissimilar to—and therefore misrepresent—their targets even in the relevant respects. This is what makes their use, at least *prima facie*, a puzzle. So if my suggested interpretations amount to a novel kind of model-target 'similarity' then so be it.

This interpretation of my argument amounts to a plea to defenders of the similarity-based accounts of scientific representation to be more specific about what counts as a similarity—beyond an appeal to 'shared' or 'similar' features—in a way that captures these sorts of models. One issue with this approach—foreshadowed in the preceding discussions of Figures 1, 2, and 3—is that in the cases at hand the *differences* in the ways in which models and their targets exhibit these features (on the inevitability/susceptibility and specific/non-specific axes), are crucial for the models to perform the representational functions that they do (in particular to represent a broad class of systems accurately, despite not representing the specific details of any particular member of that class, see Section 5 for further discussion). Dissimilarity with respect to the the details of the 'shared' features plays a positive epistemic role, something that remains opaque under a naïve understanding of what counts as similarity.

I think it's more fruitful to avoid thinking in terms of similarity altogether. It is not clear that two systems are similar with respect to some feature either if one inevitably exhibits it and another merely has susceptibility to do so, or if they instantiate it at different levels of abstraction. There is no trade whatsoever in Akerlof's model, and calling the model 'similar' to an actual Pareto-inefficient car market in this respect seems wrong. Of course the notion of 'similarity' might be stretched in such a way as to accommodate these cases, but what is gained by the use of the moniker? Moreover, if the

[22] I'm taking it for granted that there is no particular metaphysical worry with making sense of these sorts of claims. However, in the case of toy models, one might have the epistemological worry that if the models are supposed to be accurate, even in cases where the susceptibilities of the target are overridden by other factors, then one cannot empirically check whether the model is an accurate representation. Again this is not a problem for my account: it's not built into the interpretational accounts of scientific representation that a model's content can be assessed empirically. Moreover, in many uses of the models discussed in this paper, their content can be assessed and does come out as accurate (recall all it takes is that the parameters of the system drift in the same direction as the model). I am grateful to an anonymous referee for encouraging me to be explicit about this.

notion can be stretched in such a way, this indicates that it is not playing as important a theoretical role as was initially intended. The discussion above, of applying the term 'similarity' *ex post* whenever a representation delivers a truth about its system applies again. Once the notion of 'similarity' becomes this flexible, it strikes me that it is more pertinent to simply directly associate the related model features with those to be exported to their targets, as per the aforementioned interpretation functions.

4.3 Idealisation $\neq$ misrepresentation

The previous argument has important implications on how we should understand 'idealisation'. I take it that in general terms, 'idealisation' refers to models which distort the features of their targets. But there is a tendency in the literature to slip from understanding idealisation as distortion of features to idealisation as misrepresentation of those features. Martin Thomson-Jones makes this explicit:

> 'On the regimentation of usage I am thus proposing, the term *'idealization' applies, first and foremost, to specific respects in which a given representation misrepresents*, whereas the term 'abstraction' applies to mere omission' [Jones, 2005, p. 174, emphasis added].

Similar claims are found in, for example, [McMullin, 1985, pp. 255–256], and [Weisberg, 2007a, p. 657], and are often implicitly and explicitly assumed throughout the literature.

If one subscribed to a similarity-based account of scientific representation, then this would follow. But if the preceding discussion is true, then the move from distortion to misrepresentation is not justified. Idealised representations do, or at least need, not represent their targets as having the same features as they have. So the fact that certain features of an idealised model are distortions of features of their targets doesn't mean that they thereby misrepresent the target in those respects. As long as a model user understands the idealisations in question, then they shouldn't interpret those features in a way that entails exporting them, incorrectly, to the model's target.[23] Rather, these are precisely the sorts of features that get altered by the interpretation function when moving from reasoning about models to reasoning about their target systems.

5 Precision vs. Generality

So far I have argued that the use of toy models is on firm epistemic ground, at least in the sense of diffusing the worry that their prevalence is in tension with the idea that science aims, at least in part, at providing accurate representations of the world. By interpreting them appropriately, these models are accurate but nonspecific representations of their targets. Nevertheless, it could be argued that toy models provide a different sort of puzzle. If the way they represent their targets involves claims about susceptibilities and lack specificity, then why do they continue to be used given that we have access to alternatives that represent their targets more specifically (notice I say 'more specifically' rather than 'with increased accuracy')?

Here we can appeal to Levins' [1966] account of the trade-off between precision and generality in model building.[24] Models can be precise in the sense that they 'make fairly accurate measurements, solve numerically on the computer, and end with precise testable predictions applicable to these particular situations' [1966, p. 422]. Or models can be general in the sense that they apply to a large number of target systems that differ from one another in significant ways. I take it that toy models exemplify the virtue of generality at the cost of precision (understood in terms of specificity). And the way that they do this is precisely in the sense discussed in subsection 4.2: when they are interpreted they move from an inevitability in the model to a susceptibility claim about the target, and from a specific model-fact to a less specific claim about the target. Both of these conditions allow toy models to be general, in the sense of applying to many target systems, but at the cost of precision/specificity.

Consider the models from Section 3, Volterra [1928, p. 5] notes that his model can also be targeted at plants and their parasites, and suggests that infectious diseases such as malaria might also show such fluctuations (it can also be applied to economic systems [Goodwin, 1982]). Schelling states that his model

[23] Notice that I say 'as long as a model user understands the idealisations in question'. This allows for cases where toy models do in fact misrepresent their targets (or better: are interpreted in such a way that they misrepresent their targets). But that some toy models can be misrepresentations doesn't make all such models misrepresentations. Understanding precisely how such models are idealised plays a significant role in understanding the models themselves. But this is just good scientific practice.

[24] Levins actually takes there to be three competing desiderata for good scientific models: precision, generality, and realism. I set the latter aside for my current purposes.

can be applied to 'whites and blacks, boys and girls, offices and enlisted men, students and faculty' [1978, p. 138]. And after introducing his model Akerlof [1970, pp. 492–500] devotes the rest of his paper to discussing additional targets including: insurance markets; the employment of minorities; the costs of dishonesty in a market more generally; and credit markets in underdeveloped countries.

The success of these models when applied to a wide number of target systems is the result of the fact that the claims they generate do not concern the exact details of the target systems and because they make weakened susceptibility claims which are compatible with other influencing factors featuring in the target systems.[25]

6 Conclusion

Before concluding, I want briefly comment on how what I have said here connects to, and could be extended to inform, another debate concerning the role of models that are explicitly dissimilar to their targets. Just as the use of toy models seems to pose a puzzle for those who assume they are inaccurate, the likes of Bokulich [2008, 2011, 2016], Elgin [2004, 2017], and Potochnik [2017] have argued that the presence of highly idealised models in scientific practice puts pressure on the idea that accurate representation is necessary for explanation and/or understanding, and thus motivate non-factive accounts of explanation and/or understanding. Whilst I am in agreement with much of what they say, it does bear noting that the debate takes for granted that because certain scientific models are highly dissimilar to their targets (in the relevant respects)—because they are highly idealised—they are therefore inaccurate representations. My discussion provides an alternative strategy. We can reconsider the presumption that idealised models are misrepresentations in the first place. Rather, we should direct our attention to the claims generated by such models in scientific practice. If those claims are true, then the models themselves can be justly considered accurate representations, despite being idealised. Arguing for this in any more detail here would take me too far afield, but I hope at least that it is clear how my suggestions here provide a novel perspective in the debates concerning the factivity of explanation and understanding.

To sum up. Toy models, characterised as extremely simple, highly idealised models of actual target systems should be distinguished from models without targets (or 'substitute' models) and models that provide only novel how-possibly explanations. Each of the models discussed in this paper fit such a characterisation. Such models initially appear puzzling: why are they used, given their simple and highly idealised nature? I have argued that their prevalence in science is best explained by the fact that they accurately represent, albeit non-specifically, many target systems. The fact that they can function as accurate representations is a result of dropping the, often implicit, assumption that in order to be accurate scientific models should be similar to their targets. On this assumption (heavy) idealisation is conflated with misrepresentation. Without this assumption more conventional associations between models and their targets can underpin accurate representation. Thus, the puzzle of toy models dissolves.

Acknowledgements

This research was supported by the Jacobsen Trust. I am grateful to three anonymous referees and the editors of this journal for highly engaging comments and discussion. I would also like to think Alisa Bokulich, Corey Dethier, Goreti Faria, Roman Frigg, Sebastián Murgueitio Ramírez, Emanuele Ratti, Bryan Roberts, Paul Teller, and Nicolas Wüthrich, as well as audiences at the TINT workshop on *Highly Unrealistic Models* at the University of Helsinki and *Models and Simulations 8* at the University of South Carolina, for useful comments and discussions. Thanks also to the History and Philosophy of Science Program at the University of Notre Dame where much of this research was conducted.

James Nguyen
Department of Philosophy

[25] In this sense my analysis of toy models has shares something in common with Batterman and Rice's [2014] discussion of 'minimal models' (their use of the moniker 'minimal models' comes from [Goldenfeld, 1992, p. 33], it differs from [Grüne-Yanoff, 2009]), since they too are concerned with the epistemic function of models with numerous targets, models that fall in the same 'universality class' as their targets. Such models exhibit the same macroscopic features as their targets, despite diverging with respect to their microscopic details. However their analysis focuses on models that are embedded in an abstract space of models where mathematical renormalisation group flows (transformation operations which serve to map models to more course-grained models) can specify universality classes. The sorts of toy models discussed in this paper are not accompanied by this mathematical machinery. So I take it that my way of looking at toy models provides a different way of getting to related, but importantly distinct, conclusions about the role of highly simple and idealised models in the sciences.

University College London
and
Institute of Philosophy, School of Advanced Study
University of London
and
Centre for Philosophy of Natural and Social Science
London School of Economics and Political Science
London, UK
james.nguyen@sas.ac.uk

References

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

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

Bokulich, A. (2008). Can classical structures explain quantum phenomena? *The British Journal for the Philosophy of Science*, 59(2):217–235.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, 180(1):33–45.

Bokulich, A. (2016). Fiction as a vehicle for truth: Moving beyond the ontic conception. *The Monist*, 99(3):260–279.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press, Cambridge.

Elgin, C. Z. (2004). True enough. *Philosophical Issues*, 14(1):113–131.

Elgin, C. Z. (2010). Telling instances. In Frigg, R. and Hunter, M. C., editors, *Beyond Mimesis and Convention: Representation in Art and Science*, pages 1–18. Springer, Berlin and New York.

Elgin, C. Z. (2017). *True Enough*. MIT Press, Cambridge, Mass and London, England.

French, S. (2003). A model-theoretic account of representation (or, I don't know much about art ... but I know it involves isomorphism). *Philosophy of Science*, 70:1472–1483.

Frigg, R. (2006). Scientific representation and the semantic view of theories. *Theoria*, 55(1):49–65.

Frigg, R. (2010a). Fiction and scientific representation. In Frigg, R. and Hunter, M., editors, *Beyond Mimesis and Convention: Representation in Art and Science*, pages 97–138. Springer, Berlin and New York.

Frigg, R. (2010b). Models and fiction. *Synthese*, 172(2):251–268.

Frigg, R. and Hartmann, S. (2018). Models in science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition.

Frigg, R. and Nguyen, J. (2016a). The fiction view of models reloaded. *The Monist*, 99(3):225–242.

Frigg, R. and Nguyen, J. (2016b). Scientific representation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Frigg, R. and Nguyen, J. (2017). Models and representation. In Magnani, L. and Bertolotti, T., editors, *Springer Handbook of Model-based Science*, pages 49–102. Springer, Cham.

Frigg, R. and Nguyen, J. (2018). The turn of the valve: representing with material models. *European Journal for Philosophy of Science*, 8(2):205–224.

Fumagalli, R. (2016). Why we cannot learn from minimal models. *Erkenntnis*, 81(3):433–455.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, 71(4):742–752.

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, 172(1):269 – 281.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies*, 143:101–116.

Goldenfeld, N. (1992). *Lectures on phase transitions and the renormalization group*. Frontiers in physics. Addison-Wesley, Advanced Book Program.

Goodman, N. (1976). *Languages of Art*. Hacket, 2nd ed., Indianapolis and Cambridge.

Goodwin, R. M. (1982). A growth cycle. In Goodwin, R. M., editor, *Essays in Economic Dynamics*, pages 165–170. Palgrave Macmillan, London.

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

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science*, 64(Supplement):S325–S336.

Jones, M. R. (2005). Idealization and abstraction: A framework. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 86(1):173–218.

Khosrowi, D. (2018). Getting serious about shared features. *The British Journal for the Philosophy of Science*, page axy029.

Knuuttila, T. and Loettgers, A. (2016). Modelling as Indirect Representation? The LotkaVolterra Model Revisited. *The British Journal for the Philosophy of Science*, 68(4):1007–1036.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54(4):421–431.

Levy, A. (2015). Modeling without models. *Philosophical Studies*, 152(3):781–798.

Lotka, A. (1925). *Elements of Physical Biology*. Williams & Wilkins Company, Baltimore.

Luczak, J. (2017). Talk about toy models. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 57:1–7.

Mäki, U. (2009). Missing the world. models as isolations and credible surrogate systems. *Erkenntnis*, 70(1):29–43.

Mäki, U. (2011). Models and the locus of their truth. *Synthese*, 180(1):47–63.

Mäki, U. (2011). The truth of false idealizations in modeling. In Humphreys, P. and Imbert, C., editors, *Models, Simulations, and Representations*, pages 216–233. Routledge, New York and Abingdon.

McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A*, 16(3):247–273.

Nguyen, J. (2017). Scientific representation and theoretical equivalence. *Philosophy of Science*, 84(5):982–995.

Norton, J. D. (2003). Causation as folk science. *Philosophers' Imprint*, 3:1–22.

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Reutlinger, A., Hangleiter, D., and Hartmann, S. (2017). Understanding (with) Toy Models. *The British Journal for the Philosophy of Science*, 69(4):1069–1099.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. Norton, New York and London.

Suárez, M. (2003). Scientific representation: against similarity and isomorphism. *International Studies in the Philosophy of Science*, 17(3):225–244.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science*, 71(Supplement):767–779.

Suárez, M. (2015). Deflationary representation, inference, and practice. *Studies in History and Philosophy of Science*, 49:36–47.

Sugden, R. (2000). Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology*, 7(1):1–31.

Teller, P. (2001). Twilight of the perfect model model. *Erkenntnis*, 55(3):393–415.

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560.

Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *Journal du Conseil*, 3(1):3–51.

Weisberg, M. (2007a). Three kinds of idealization. *Journal of Philosophy*, 104(12):639–659.

Weisberg, M. (2007b). Who is a modeler? *The British Journal for the Philosophy of Science*, 58:207–233.

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

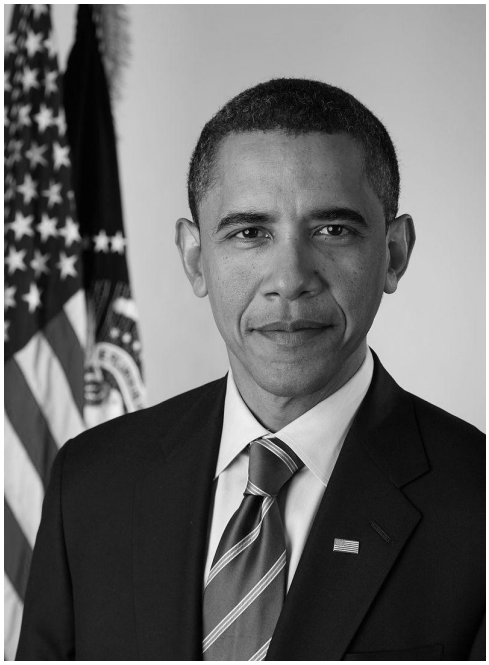


Figure 1: Obama (Pete Souza, The Obama-Biden Transition Project [Creative Commons Attribution 3.0 License]), monochrome

Figure 2: Obama inverted (Pete Souza, The Obama-Biden Transition Project [Creative Commons Attribution 3.0 License]), inverted monochrome

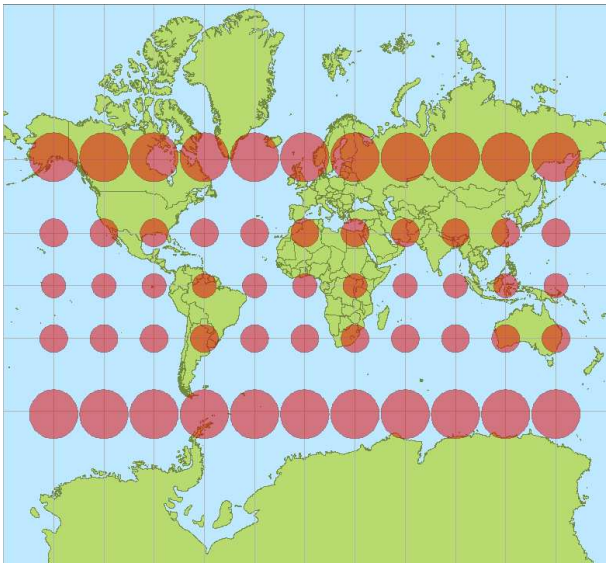
H 

Figure 3: Tissot indicatrix (Stefan Kühn [Creative Commons Attribution-Share Alike 3.0])