Understanding (With) Toy Models

Alexander Reutlinger, Dominik Hangleiter, and Stephan Hartmann

July 26, 2016

Abstract

Toy models are highly idealized and extremely simple models. Although they are omnipresent across scientific disciplines, toy models are a surprisingly under-appreciated subject in the philosophy of science. The main philosophical puzzle regarding toy models is that it is an unsettled question what the epistemic goal of toy modeling is. One promising proposal for answering this question is the claim that the epistemic goal of toy models is to provide individual scientists with understanding. The aim of this paper is to precisely articulate and to defend this claim. In particular, we will distinguish between autonomous and embedded toy models, and, then, argue that important examples of autonomous toy models are sometimes best interpreted to provide how-possibly understanding, while embedded toy models yield how-actually understanding, if certain conditions are satisfied.

Contents

1 Introduction

Across the natural and social sciences, researchers construct very simple and highly idealized models, which the experts in a particular field of inquiry can cognitively grasp with ease. Following common terminology from the sciences, we call such models "toy models" – a term that is not meant to have belittling or derogatory connotations. Paradigmatic examples of toy models include the Ising model in physics, the Lotka-Volterra model in population ecology, and the Schelling model in the social sciences (see, for instance, Hartmann [1999]; Sugden [2000]; Weisberg [2013]). A useful characterization of toy models appeals to three essential features: (1) models of this type are strongly idealized in that they often include both Aristotelian and Galilean idealizations (see Section 2.1 for details regarding this familiar distinction), (2) such models are extremely simple in that they represent a small number of causal factors (or, more generally speaking, of explanatory factors) responsible for the target phenomenon, and (3) these models refer to a target phenomenon (as opposed to, for instance, models of data; see Frigg and Hartmann [2012]).

To be clear from the start, we do not claim that there is a sharp boundary between toy models and other models. Instead of a sharp distinction, there seems to be a continuum of models with respect to their degree of simplicity and, independently, their degree of idealization. If one compares toy models with more complex models representing a large number of causal factors responsible for the target phenomenon (such as complex models in climate science), then toy models are located at the 'simple end' of the spectrum. If one contrasts toy models with less idealized models (that is, models involving fewer idealized and more approximately true assumptions), then toy models are located at the 'strongly idealized end' of a continuous spectrum. Our characterization of toy models as simple and highly idealized does, of course, permit the existence of (i) simple and less idealized, (ii) complex and highly idealized, and (iii) complex and less idealized models. In sum, the concept of a toy model is a vague concept. However, vagueness does not exclude that there are clear core examples of what is in the extension of a concept. We will discuss some core examples of toy models in this paper.

Idealizations and models have been major topics in the recent philosophy of science (Morgan and Morrison [1999]; Bailer-Jones [2009]; Frigg and Hartmann [2012]). It is, however, a curious fact that philosophers of science have not devoted sufficient attention to toy models, despite their apparently central role in many sciences (notable exceptions include Hartmann [1999]; Sugden [2000]; Strevens [2008]; Bailer-Jones [2009]; Grüne-Yanoff [2009], [2013]; Weisberg [2013]). Toy models are deeply puzzling because their strongly idealized and simple nature raises hard questions: to what end do scientists construct toy models? Why should one have any

confidence in the claim that strongly idealized and simple models can be used for modeling any real social and natural phenomena? Or, even more provocatively, why should one believe that toy models are anything more than 'mathematized science fiction', giving us no more clues about real world phenomena than non-mathematized fairy tales?

As a pessimistic response, one could be tempted to think that toy models are not very useful, because they cannot represent real (or actual) phenomena. To motivate this sort of scepticism, suppose that explanation and prediction are two central goals of modeling in science. We know as a matter of fact that toy models – as idealized models – are literally false of their intended target systems; or, put in semantic terms, toy models clearly do not accurately map onto their targets (for instance, because a toy model is not isomorphic to its target, supposing that isomorphism is required for representation). Being false is a feature that, at least prima facie and according to standard accounts of explanation, undermines the explanatory character of a model, as the explanans of an explanation is required to be (approximately) true.[1] Similar worries emerge with respect to the predictive use of toy models: the majority of toy models is not suited for precise quantitative predictions. Why should anyone trust the predictions generated by toy models, if one knows that these predictions rest on false assumptions?

Opposing a pessimistic attitude towards toy models, some philosophers have recently claimed that the epistemic goal of toy modeling is to obtain *understanding of natural and social phenomena* (Hartmann [1998]; De Regt and Dieks [2005]). By virtue of their simplicity, toy models enable scientists to retain a sort of epistemic access to scientific models (and the mathematical procedures for solving these models). The simplicity of toy models distinguishes them from other kinds of models – for instance, from complex models that can only be solved via computer simulation, such as climate models.[2] Is the claim that toy models – being idealized models – yield understanding really warranted? Are toy models appropriate for achieving understanding? And if so, what kind of understanding do scientists get from toy models? The main goal of this paper is to answer these questions by (1) distinguishing two kinds of toy models, (2) developing an account of understanding that is adequate for toy models, and (3) determining whether different kinds of toy models are apt for different kinds of understanding.

According to an alternative approach in the recent literature, simple and idealized models are portrayed as *minimal models* (for instance, Strevens [2008]; Grüne-Yanoff [2009], [2013]; Weisberg [2013]; Batterman and Rice [2014]). However, al-

[1]But there are exceptions such as Cartwright ([1983], chapter 8).

[2]Hartmann ([1999]) contrasts toy models with complex models that are solved with computer simulations. He considers the latter to be "black boxes" for the individual scientist. Similarly, Humphreys ([2004]) highlights a contrast between simple models and the epistemic opacity of computer simulations.

though we are convinced that some toy models may be interpreted as minimal models, we will argue that other toy models are not best understood in terms of minimal models. We shall return to the minimalist interpretation (Section 4.2).

The plan of the paper is as follows: in Section 2, we present the distinction between embedded and autonomous toy models. In Section 3, we elaborate an account of understanding (the refined simple view). Our account of understanding is inspired by Strevens' ([2013]) "simple view" of understanding, and Bailer-Jones' ([1997]) naturalistic account of the "subjective component" of understanding (epistemic access). The refined simple view explicates two kinds of understanding: how-actually understanding and how-possibly understanding. In Section 4, we argue that (i) embedded toy models yield how-actually understanding, if certain conditions hold (Section 4.1), (ii) standard accounts of idealizations – such as McMullin's strategy, minimalism and dispositionalism – do not support the claim that all autonomous toy models provide how-actually understanding (Section 4.2), and (iii) there are some autonomous toy models that one best interprets as yielding how-possibly understanding (Section 4.3).[3]

2 Embedded and Autonomous Toy Models

To analyze whether and how toy models yield scientific understanding, it is useful to introduce a distinction between two kinds of toy models: embedded and autonomous toy models. We will first illustrate the concept of an embedded toy model (in Section 2.1), and then turn to an illustration of autonomous toy models (in Section 2.2). We conclude this section with a brief qualification regarding the heterogeneity of autonomous toy models (in Section 2.3).

2.1 Embedded toy models

Above we introduced toy models as models of *phenomena*. However, some toy models are also models in a different sense of the term "model": they are also models of *a theory*. This observation will permit us to introduce a central distinction between embedded and autonomous toy models.

Some toy models are *embedded* into an empirically well confirmed theory. More precisely, embedded toy models are *models of an empirically well confirmed framework theory*, while autonomous toy models – to which we will turn below – are

[3]One also finds numerous *material* toy models in chemistry and in the life sciences, such as croquet ball models of molecules, Watson and Crick's metal model of DNA, and simple model organisms (see Meinel [2004]). In this paper, we restrict the focus to toy models *qua* mathematical models of phenomena. A comparative analysis of mathematical and material toy models is, unfortunately, beyond the scope of this paper.

not. This characterization of an embedded toy model relies on a familiar distinction from the philosophical literature on models and model theory in mathematics: namely, the distinction between (i) a (framework) theory and (ii) models of the framework theory (see Hartmann [1998]; Bailer-Jones [2009]; Frigg and Hartmann [2012], Sect. 1.3). In model theory, a theory is a set of uninterpreted sentences. When model theory is used to express a framework theory, this set of sentences includes, most prominently, the framework theory's abstract calculus and its general laws. Models of a framework theory are taken to be structures in which the sentences of the framework theory (such as the theory's abstract calculus and its general laws) are true (Frigg and Hartmann [2012], Sect. 1.3). Or, to state the same point in more precise model-theoretic terms, models of a framework theory consist of a domain of objects and an interpretation of the theory's abstract calculus and the laws over the domain (Chang and Keisler [1990], Sect. 1.3; Bell and Slomson [1974], Sect. 3.2).[4] Examples of empirically confirmed framework theories include Newtonian Mechanics and Quantum Mechanics. Well-known examples of models of framework theories are the model of a pendulum and models of planetary motion (being examples of models of classical mechanics), and the Standard Model of Particle Physics (being a model of quantum mechanics). Models of a theory are constructed within a framework theory. Constructing such a model in order to represent a target phenomenon often requires moving beyond the resources of the framework theory: it consists in making a number of specific assumptions about the target (Morgan and Morrison [1999]; Frigg and Hartmann [2012]).

With these conceptual tools in mind, we are now in a position to characterize embedded toy models more precisely. Embedded toy models are (1) models of a well confirmed framework theory, and (2) they are simple and idealized models of phenomena.

Before we turn to a more detailed example of an embedded toy model, let us add a note on terminology. We use the term "idealization" as an umbrella term for (at least) two general kinds of idealizations that are usually distinguished in the literature – Aristotelian and Galilean idealizations (Frigg and Hartmann [2012], Sect. 1.1). A model involving an Aristotelian idealization "strips away" some feature(s) that the target system of the model in fact possesses (for instance, a model of a pendulum strips away the color of the pendulum), or the model rests on the assumption that some causal factor actually influencing the target system is absent or "neutralized" (Mäki [2011], p. 51). Aristotelian idealizations are also discussed in terms of "abstraction" (Cartwright [1989]) and "isolation" (Mäki [1992], [2011]; Hüttemann [2004], [2014]). By contrast, Galilean idealizations deliberately dis-

[4]Although we rely on model theory to explicate the notion of a "model of a theory", our notion of embedding does not coincide with the model-theoretic notion of embedding, as, for instance, Bell and Slomson ([1974], p. 73) define it.

tort the target system, for instance, by making the assumption that agents are perfectly rational, that the number of animals in a population or the number of molecules in a gas goes to infinity, and so on (McMullin [1985]; Cartwright [1989]; Weisberg [2013]). Aristotelian and Galilean idealizations can co-occur in one and the same model. The examples of toy models we discuss tend to involve both kinds of idealizations.[5]

In this paper, our goal is not to highlight the differences between Aristotelian and Galilean idealizations. For this reason, we will merely speak of "idealizations" as an umbrella term throughout the remainder of the paper. Analogously, we will use the term "de-idealization" to denote cases of both (Aristotelian) "de-isolation" and (Galilean) "de-idealization", to use Mäki's ([2011], p. 48) terminology. What matters for our concerns is that modeling assumptions involving Aristotelian and Galilean idealizations assert something that is literally false of the target system (Cartwright [1983], p. 45). Or, put differently, models involving Aristotelian or Galilean idealizations, prima facie, do not accurately represent their targets (for instance, because they are not isomorphic to their targets, if that is what the theory of representation requires).

Let us now turn to a concrete example of an embedded toy model. Consider Newtonian Mechanics as a framework theory. This theory lays out a small number of general laws (Newton's laws of motion) and it provides the scientist with guidelines for the construction of concrete models for specific systems, or phenomena (see Giere [1988]; Bailer-Jones [2009]). To study, for example, the motion of a single planet around the Sun in our solar system, a number of model assumptions have to be made. In the simplest case, one might want to study a system consisting of only the Sun and the planet under consideration. Let us call this simple model the "Sun-plus-one-planet model". This model is a model of Newtonian mechanics – the Sun-plus-one-planet model is a structure in which the sentences of Newtonian mechanics (such as the theory's abstract calculus and its general laws) are true. Moreover, if one analyzes the Sun-plus-one-planet model as the model of a phenomenon (for instance, of the Earth orbiting around the sun), this model involves idealizations, because the modeler disregards the other planets, the moon(s), and other stellar objects that are known to exist. Moreover, the model refers only to gravitational interactions between the Sun and the planet. From Newton's laws of motion and the model assumptions, one can then derive the orbit of the planet. In a simple calculation, which can be found in any textbook, one obtains that the orbit of the planet is (approxiamtely) an ellipse with the Sun in one of the two foci.

The Sun-plus-one-planet model is an embedded toy model, because: (1) it is a

[5]Models involving both Aristotelian and Galilean idealizations are sometimes called "caricature models" in the literature – see Hartmann and Frigg ([2012], Sect. 1.1) for further references.

model of a framework theory, Newtonian Mechanics, that is well confirmed, at least in a particular domain of application, (2) it is simple, as it describes few causal, or explanatory, factors (i.e. a physical system of only two interacting bodies), (3) it is idealized (as it deliberately disregards the gravitational influence of other planets and refers only gravitational interaction), (4) it is a model of a phenomenon (i.e. a target phenomenon such as the earth orbiting around the Sun).

There are numerous examples of embedded toy models in physics (see Morgan and Morrison 1999) including the Ising model of non-relativistic quantum mechanics, and the ϕ^4-theory of quantum field theory (with quantum field theory as the embedding framework theory). Hüttemann ([2014]) provides another illustrative example: he treats an oscillator and a rotator as embedded toy models with quantum mechanics as the embedding theory. The ideal gas law can be understood as the deductive consequence of a toy model that is embedded in statistical mechanics (Strevens [2008], Chapter 8; Dizadji-Bahmani et al. [2010]).[6] We will restrict our discussion of embedded toy models to examples from physics. However, one may also find embedded toy models in other disciplines. In the life sciences, Fisher's famous sex ratio model seems to be a toy model embedded into Darwinian evolutionary theory (Sober [1984], pp. 51-8).

2.2 Autonomous toy models

Several well-known toy models are not embedded ones, i.e. they are not models of a well confirmed framework theory. We call toy models of this sort "autonomous" toy models. Autonomous toy models share the simple and idealized character with their embedded cousins. The Schelling model of segregation and the Lotka-Volterra model of predator-prey population growth are paradigmatic examples of autonomous toy models (Schelling [1971]; Sugden [2000]; Weisberg [2007]).

A familiar paradigm: Schelling's model of segregation. A paradigmatic autonomous toy model is Thomas Schelling's model of segregation. Schelling ([1971]) developed a famous toy model of the phenomenon of racial (and other kinds of) segregation (see, for instance, Sugden [2000]; Weisberg [2013]). Racial segregation is a general kind of phenomenon that is contingently instantiated in actual, or real-world, cities such as in Chicago and Detroit. Schelling's model works with a small number of simple assumptions: (1) two sorts of agents (for instance, black and white agents) live in a very sparse environment (a two-dimensional grid),

[6] Thermodynamical models of phase transitions and of certain universal aspects of these phase transitions are controversial cases. Whether these models are embedded models depends on whether one believes that the thermodynamical description can be reduced to statistical mechanics. This is a controversial issue we cannot address in this paper (see Batterman [2002]; Butterfield [2011]; Norton [2012]).

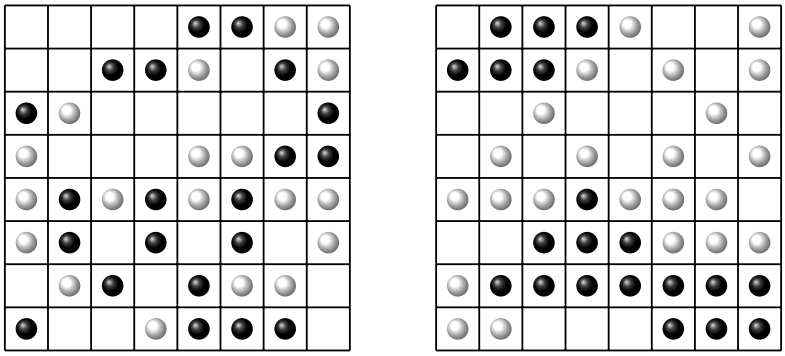

Figure 1: The Schelling model of segregation. The left hand side shows a random distribution of agents. The right hand side depicts a pattern of segregation.

(2) the agents are assumed to be initially randomly distributed on the grid, and (3) the agents interact in accord with a simple behavioral rule (for instance, each agent moves to an empty spot in her/his neighborhood on the grid, if less than about 30% of her/his neighbors do not have her/his color). If one starts with randomly distributed agents (see Figure 1, on the left), then running this simple model by reiterating the behavioral rules leads to the emergence of segregation after a small number of steps (see Figure 1, on the right). Schelling took the model to explain that racial segregation can occur even if the agents do not have strongly and explicitly racist attitudes (but merely conform to the 30% rule) and agents would actually prefer to live in *non*-segregated cities. The model also allows us to consider the consequences of varying initial conditions and the rules.

The Schelling model has racial segregation as its target phenomenon. As stated above, racial segregation is general kind of phenomenon that is contingently instantiated, for instance, in Chicago. If one takes the Schelling model to apply to particular instantiations of racial segregation (for instance, the racial segregation in Chicago in the 1960s), then the rules and other modelling assumptions are simplified and idealized to such an extent that they do not accurately represent, say, the preferences of the actual inhabitants of Chicago's highly segregated neighborhoods in the 1960s or in 2016. The model is simple in assuming a very sparse environment (a grid) and agents that are characterized by very few properties (most importantly, by their color and a behavioral rule). The model is idealized in the following manner: for instance, (a) each agent is assumed to know how many agents of each color live in her/his environment, (b) every agent is assumed to be able to move whenever s/he is dissatisfied with the color of her/his neighbors, (c) social and economic factors (such as education and income) are taken not to make a difference at all, (d) the inhabitants of, say, Chicago and Detroit never were randomly distributed, and so on (Schelling [1971], p. 149). Moreover, the Schelling

model is not embedded into (that is, it is not a model of) an empirically confirmed framework theory. In sum, we conclude that the Schelling model of segregation is an autonomous toy model.

A novel case from econophysics: the DY-model. Econophysics is a fairly young discipline that exploits mathematical models from statistical physics in order to understand economic phenomena. One important class of econophysical models, so-called "collision models", depict economic exchanges among agents in analogy with collisions of molecules in a gas, as described by statistical mechanics. One influential and successful collision model in econophysics is the Drăgulescu-Yakovenko ([2000]) model (DY-model, for short). The model is taken to successfully capture important qualitative features of the distributions of individual monetary incomes found in many real economies – in particular, the 'stylized fact' that these income distributions are exponential distributions with a power law tail. These features of income distributions are the target phenomenon for the DY-model and, indeed, the DY-model successfully captures this phenomenon (see Thebault et al. [forthcoming] for an in-depth discussion of the technical details and the influence of the DY-model in econophysics).

The starting point for the DY-model is a population of 'zero-intelligence' agents. These agents have a single property: their money. The agents lack preferences, expectations, rationality, and other properties of 'real' agents, at least as portrayed by mainstream economics. At any given time t an agent i is associated with a single property, their monetary income $m_i(t)$ (which is always non-negative, so debt is not allowed). In the DY-model, one first assumes a large population (i.e. N agents, with $N \gg 1$), and then randomly selects two individuals at some time t. For a selected pair of agents, the initial pre-interaction state can be characterised completely in terms of two numbers: $m_i(t)$ which is the income of agent i at time t; and $m_j(t)$ which is the income of agent j at time t. The DY-model treats all interactions in the population in terms of binary exchanges of money in the same way as in the kinetic theory of gases one can treat the interaction between molecules in a gas in terms of binary exchanges of kinetic energy (see Thebault et al. forthcoming). Another crucial assumption in the DY-model is that both the total number of agents, N, and the total amount of money, $M = \sum_i m_i(0)$, are held fixed. That is, $\sum_i m_i(0) = \sum_i m_i(t)$ for all t.

All of the assumptions regarding zero intelligence agents, a restriction to binary interactions, and the conservation of the total number of agents and the total amount of money are clearly idealizations. Moreover, the DY-model is not a model of a well confirmed framework theory, because statistical mechanics is not a well confirmed theory for the domain of economic processes. In particular, the DY-model does not include general dynamical laws of a well confirmed framework

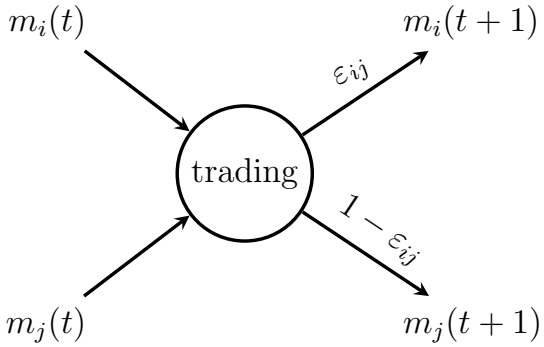

Figure 2: The exchange dynamics of the DY-model (from Thebault et al. [forthcoming])

theory (for the relevant domain of application, i.e. economic processes) describing how the initial conditions of the agents determine the nature of the collisions between agents. The DY-model rather rests on a formal analogy with certain aspects of statistical mechanics.[7] However, instead of drawing on some well confirmed framework theory, the DY-model pictures the agent-agent 'collision' with a simple exchange mechanism, such that all the money of the two agents is pooled, and then a random fraction is given to one, and the rest to the other (see Figure 2). This simple exchange mechanism thus leads to a post-interaction state characterised by:

$$m_i(t+1) = m_i(t) + \Delta m \tag{1}$$
$$m_j(t+1) = m_j(t) - \Delta m \tag{2}$$

where

$$\Delta m = \epsilon_{ij}[m_i(t) + m_j(t)] - m_i(t) \tag{3}$$

with ϵ_{ij} a random variable uniformly distributed between 0 and 1, varying with each discrete time-step, and labelled by the index of the two agents in the interaction (i.e. agents i and j).

In sum, we classify the DY-model as an interesting and novel exemplar of an autonomous toy model, because (a) the DY-model relies on a strikingly simple exchange mechanism, (b) it is idealized in assuming 'their money' as the agents only characterizing property, that many-agent interaction do not occur, that the number of agents and the amount of money is conserved, and so on, (c) it has a target phenomenon (specific qualitative features of income distributions), and (d) the DY model is not emdebbed into an empirically confirmed framework theory.

[7]For a detailed discussion of this point, see Thebault et al. ([forthcoming]).

2.3 Qualification

Although the distinction between autonomous and embedded toy models is sharp, the class of autonomous toy models is quite heterogeneous. This heterogeneity exists because: (1) some autonomous models seem to bear no relevant relation to a well confirmed framework theory (such as the Schelling model and the DY-model, or so we assume). (2) However, other autonomous toy models are non-trivially associated with a highly confirmed framework theory, but the toy model in question is not a model of that framework theory.

Let us briefly present an examples illustrating the latter case, the MIT bag model. For the MIT bag model, the relevantly associated (but not embedding) framework theory is quantum chromodynamics (QCD) which is extremely hard to solve in the low-energy domain (see Hartmann 1999). Here, QCD can only be solved using high-powered computer simulations. These computational models and computer simulations function like a black box and are, hence, not easy to grasp and understand (we will return to the notion of grasping below). The MIT bag model, on the other hand, identifies one crucial feature of QCD: it identifies quark confinement as QCD's key feature and models a hadron as a hard sphere in which quarks move freely (see Figure 3 which depicts the "bag", the freely moving quarks in it, and arrows representing the confining force). It is important for our concerns that the MIT bag model is not a model of QCD. The model is rather "inspired" by QCD and it is ultimately justified by a story that connects the model to QCD, as Hartmann ([1999], [2001]) argues.

However, in this paper, we will focus our discussion on embedded toy models and autonomous toy models that are not connected to a theory via a story. We will leave it for future research to address the question what kind of understanding toy models such as the MIT bag model provide.

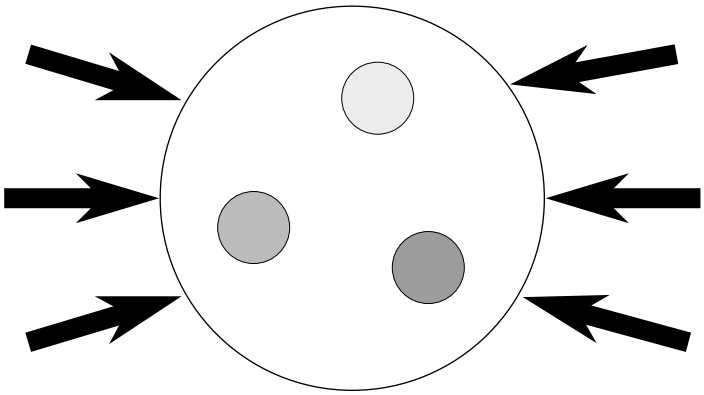

Figure 3: The MIT Bag Model (from Hartmann [1999], p. 336).

3 A Theory of Understanding for Toy Models

Back to our main question: do toy models yield understanding? And, moreover, is our taxonomy of autonomous and embedded toy models helpful for answering this question? One promising and straightforward way to approach these questions is to ask whether there is any convincing philosophical account of understanding that applies to autonomous and embedded toy models. We take Henk De Regt and Dennis Dieks' influential account of scientific understanding (as developed in De Regt and Dieks [2005]; De Regt [2009]) as our starting point. Presenting their account (in Section 3.1) will primarily serve as a convenient way to (a) bring out a number of common assumptions in several current accounts of understanding, and (b) to motivate the account of understanding we will adopt, i.e. the refined simple view (Section 3.2).

3.1 Preliminaries and requirements

According to De Regt and Dieks, "a phenomenon P can be understood if a theory T of P exists that is intelligible (and meets the usual logical, methodological and empirical requirements)" (De Regt and Dieks [2005], p. 150; see also De Regt [2009], p. 32). Although De Regt and Dieks restrict their definition to *theories*, their approach is intended to be more permissive, since they also refer to *models* as vehicles of understanding. One of their central examples is a toy model, the MIT bag model (De Regt and Dieks 2005: 155-156).

Let us examine De Regt and Dieks' ([2005]) and De Regt's ([2009]) necessary conditions for understanding more closely: first, the explanation condition, second, the "intelligibility" condition, and, third, the "usual logical, methodological and empirical requirements".

First, De Regt explicitly ties understanding to *explanation*: understanding a phenomenon is characterized as"having an adequate explanation of the phenomenon" and a "phenomenon P is understood scientifically if a theory of P exists that is intelligible (and the *explanation of P* by T meets accepted logical and empirical requirements" (De Regt [2009], p. 32, emphasis added). Hence, we take it that having an explanation of P is a necessary condition for understanding P, according to De Regt. We will refer to this condition as the "explanation condition".

Second, De Regt and Dieks define a theory T as being intelligible for scientists "if they can recognise qualitatively characteristic consequences of T without performing exact calculations" (De Regt and Dieks [2005], p. 151). De Regt and Dieks argue, for example, that physicists consider the kinetic theory of gases to be intelligible iff the physicists are able to infer statements from the kinetic theory

"without performing exact calculation", such as the following statement: "if one adds heat to a gas in a container of constant volume, the average kinetic energy of the moving molecules – and thereby the temperature – will increase." (De Regt and Dieks [2005], p. 152) The intuition motivating the intelligibility requirement is that "in contrast to an oracle [...] we want to be able to grasp how the predictions are generated, and to develop a feeling for the consequences the theory has in concrete situations." (De Regt and Dieks [2005]. p. 143)

Third, what do De Regt and Dieks have in mind when referring to "the usual logical, methodological and empirical requirements"? Although they do not make this point explicit, we presume that they refer to familiar virtues of scientific theories (or criteria for theory choice). Thomas Kuhn's ([1977]) paper is the *locus classicus* for an assessment of the "characteristics of a good scientific theory": "These five criteria – accuracy [corresponding to emprical adequacy], consistency, scope, simplicity, and fruitfulness – are all standard criteria for evaluating the adequacy of theory" (Kuhn [1977], p. 321). For this reason we, henceforth, refer to these requirements as "Kuhnian criteria" for good scientific theories. De Regt (2009) explicitly affirms this reading: the Kuhnian criteria determine the 'goodness' of a theory T (or a model M), on which the explanation of some target phenomenon T is based (De Regt [2009], p. 32). De Regt and Dieks' main motivation for demanding that intelligibility per se is not sufficient for scientific understanding is that, for instance, astrology should not count as providing scientific understanding, because, despite of being intelligible, astrology fails to be a good theory (or model) if judged by Kuhnian criteria (De Regt and Dieks [2005], p. 150).

De Regt and Dieks' account of understanding is one of many possible starting points in the large literature on understanding. However, what matters here is that their account is, in several respects, a typical account of scientific understanding. To bring out a number of common assumptions in several current accounts of understanding (including early approaches by Friedman [1974] and Kitcher [1981]; more recently, among many others, by Trout [2002]; Strevens [2008], [2013]; and the contributions in De Regt et al. [2009]), we will focus on the explanation condition and the intelligibility condition. We will put aside the "the usual logical, methodological and empirical requirements" (the Kuhnian criteria).

The common assumptions of these accounts of understanding can be characterised as follows. An individual scientist understands some phenomenon P only if three conditions are satisfied:

1. *Explanation condition*: There is a scientific explanation of P. Philosophers concerned with understanding often differ with respect to their preferred theory of explanation. They use different theories of scientific explanation such as the covering-law account, the unification account, pragmatic accounts,

and various causal accounts of explanation.[8]

2. *Veridicality condition*: In asserting that the understanding of phenomenon P involves an explanation of P as a necessary condition, accounts of understanding inherit a feature of theories of explanation that we call the "veridicality condition". It is a common view that explanatory assumptions (that is, the explanans of an explanation) are required to be true or, at least, approximately true. Consider the following examples. Proponents of the recently dominant causal accounts typically endorse this requirement: that is, the explanans has to truthfully represent the causes of the explanandum phenomenon. For instance, Woodward holds that the explanans has to"be true or approximately so" (Woodward [2003], p. 203; Woodward and Hitchcock [2003], p. 6), and Strevens endorses the claim that the explanans is "a veridical causal model" (Strevens [2008], p. 71) consisting of true causal laws and true statements about initial conditions. Moreover, Hempel's covering law account demands that the explanans consist of true law statements and true statements about initial conditions (Hempel [1965], p. 248, p. 338). Unificationist accounts (Kitcher [1981], p. 519) and pragmatic accounts (van Fraassen [1980], p. 143) also impose a veridicality constraint on the explanans.[9]

3. *Epistemic accessibility condition*: If an individual scientist understands phenomenon P, then she/he has epistemic access to an explanation of P. De Regt and Dieks' concept of intelligibility is one possible strategy of making precise epistemic accessibility – to have epistemic access to a (toy) model, for them, just is being able to recognise qualitatively characteristic consequences of that model without performing exact calculations.

The differences between many competing accounts of understanding consist in alternative ways of spelling out each of these three conditions.

Although De Regt and Dieks' view is surely a useful starting point, their account says little about how understanding based on idealized models is possible.

[8]However, not everyone accepts that explanation is a necessary condition for understanding, see, for instance, Lipton ([2009]) and Gijsbers ([2013]).

[9]The veridicality condition is logically independent from the Kuhnian criteria. The veridicality condition and the Kuhnian criteria perform different roles in the philosophy of explanation (and understanding). On the one hand, according to many standard accounts of explanation, the veridicality condition is a necessary condition for distinguishing *explanatory from non-explanatory* assumptions. On the other hand, the Kuhnian criteria allow us to draw a different distinction: they enable us to discriminate *how good* different bodies of explanatory information are. In this paper we will restrict our attention to the former issue. – Moreover, note that not everyone accepts the veridicality condition as a requirement for a theory of explanation (see Cartwright [1983], chapter 8).

However, this is precisely the question we are concerned with. If one adopts De Regt and Dieks's account, toy models (and other idealized models) are problematic in at least one respect: generally, toy models do not satisfy the veridicality condition. For instance, the DY-model is idealized, as it assumes that economic agents are all identical, have no expectations and "zero intelligence". This is certainly an assumption that we deem (and surely hope) to be literally false. Hence, it is, at least, questionable whether the veridicality condition is met (we will return to the interpretation of idealizations in Section 4).

This observation raises a challenge if one seeks an account of understanding applicable to toy models: such an account of understanding should accommodate idealizations. Where does this leave us? One reaction to this challenge might be to revise De Regt and Dieks' account. We adopt an alternative strategy: we will argue that the refined simple view is an account of understanding that provides a strategy for addressing the challenge stemming from idealized models.

3.2 The refined simple view

Michael Strevens' ([2013]) account of understanding offers a promising strategy for avoiding the challenge from idealized models.

According to Strevens' "simple view", scientific understanding is defined as follows: "An individual has scientific understanding of a phenomenon just in case they grasp a correct[10] scientific explanation of that phenomenon." (Strevens [2008], p. 3, [2013], p. 510) The notion of grasping is Strevens' way of articulating the epistemic accessibility condition. Strevens does not provide an informative definition of the concept of grasping. Instead he takes grasping to be a "fundamental relation between mind and world, in virtue of which the mind has whatever familiarity it does with the way the world is" (Strevens [2013], p. 511). One may be concerned about the fact that grasping is taken as a primitive. We will return to this issue shortly.

What is central for our present concerns is that the simple view offers a strategy for dealing with the challenge from idealizations. Strevens ([2013], p. 512) is fully aware of the fact that the simple view cannot be applied to idealized models straightforwardly. The reason is that toy models are taken to be "literally" false, while the simple view – implying the veridicality condition – requires them to be true. As Strevens points out, most standard theories of explanation require that the explanans be true or "veridical". For instance, Strevens' own kairetic account of explanations, in its simplest form, requires that the explanans consist of true causal laws and true statements about initial conditions (Strevens [2008],

[10]Strevens ([2013], p. 512) uses to notion of "correctness" to refer to the veridicality condition.

pp. 71-2).[11]

Strevens accounts for idealized models in the following way: although idealized statements are literal falsehoods (his terminology), these statements can be re-interpreted – by using an account of idealizations – as being (approximately) true, i.e. veridical. Strevens' specific account of idealizations appeals to an "optimizing procedure" (one vital component of his kairetic account) whose function is to filter out, or to ignore, explanatorily irrelevant information that need not be explicitly stated in the explanans (Strevens [2008], pp. 96-101). Making use of this idea of ignoring explanatorily irrelevant information, Strevens ([2008], chapter 8, [2013], p. 512) develops a minimalist account of idealizations. Strevens argues that the minimalist account implies a veridical reading of idealized assumptions: idealized assumptions truthfully (i.e. veridically) report which factors are irrelevant for the explanation at hand.

The general lesson from Strevens' minimalist approach is that, if understanding involves idealized assumptions, then these assumptions ought to be re-interpreted in a veridical way. Strevens' minimalist approach does, however, not exhaust the options. Other interpretations of idealizations include McMullin's strategy, dispositionalist interpretations and how-possibly interpretations. We will discuss the minimalist view and its alternatives in Section 4.

Inspired by Strevens ([2013]), we start with the following working definition of the concept of scientific understanding:

> **The simple view.** An individual scientist S understands phenomenon P via model M iff model M explains P and S grasps M.

Let us refine this working definition in four ways:

First, naturalism about grasping. If qualified properly, we are willing to follow Strevens in assuming that grasping is a "fundamental relation between mind and world". For present purposes, we are prepared to accept that the notion of grasping is *philosophically primitive* but not scientifically primitive. What does it mean to take the notion of grasping as philosophically but not scientifically primitive? As Bailer-Jones instructively points out,

> "understanding has a subjective component, in addition to the publicly accessible component represented by explanation, in the sense that understanding takes place in an individuals mind." (Bailer-Jones [1997], p. 122)

[11] Certain kinds statistical explanations require true statements about a probability distribution over initial conditions (Strevens [2008], chapters 9-10).

Following Bailer-Jones, we adopt a naturalistic approach to this subjective component of understanding: that is, what grasping turns out to be is a scientific matter – not a philosophical matter. The subjective component of understanding can be studied by cognitive science. For example, cognitive science tells us that grasping toy models sometimes consists in being able to visualize the behaviour of the target system of a scientific toy model or to have a "mental model" of the toy model and its solutions.[12] Visualization and having a mental model are possible ways, according to cognitive science, in which the grasping of a toy model can be realized.[13]

Second, the contextual character of understanding. Understanding a phenomenon is contextual: some model in, say, population ecology, may generate understanding for an expert in this field but not for an expert in statistical physics or a lay-person. We agree with De Regt and Dieks ([2005]) in assuming that the individuals who gain scientific understanding are experts regarding the kind of phenomenon that is understood. We express this thought by saying that an individual scientist S understands a phenomenon P via model M *in a context C*, where context C is a scientific discipline and S has expert knowledge of that discipline.

Third, different modalities of explanation and understanding. The kind of explanatory information scientists receive from toy models is not always the same, or so we will argue in Section 4. It is useful to distinguish two different modalities of explanation: *how-actually explanations* and *how-possibly explanations* (see Hempel [1965]; Grüne-Yanoff [2009], [2013]). How-actually explanations possess an explanans satisfying the veridicality condition, i.e. consisting of actually true (or approximately true) statements. The explanantia of how-possibly explanations refer to merely possible explanatory factors (for instance, to possible causes and mechanisms bringing about the explanandum phenomenon, if the explanation is causal). The distinction between how-possibly and how-actually explanations can be acounted for by and integrated into many standard accounts of explanation – for instance, with the covering-law account and various causal accounts of explanations, such as Woodward's seminal theory of causal explanation and Strevens' kairetic account.

Fourth, neutrality with respect to different theories of explanation. In this paper, we do not want to take a particular stance on which account of explanation

[12]See Giere ([1988]); Bailer-Jones ([1997], Chapter 5); and Hartmann ([1999]) regarding the philosophical reflection of mental models in science. Bailer-Jones ([1997], Chapter 5) provides numerous references to cognitive science research on mental models.

[13]Our naturalist stance towards grasping need not necessarily be at odds with De Regt and Dieks' notion of intelligibility. Sometimes, but not necessarily always, grasping may very well consist in being able to draw "qualitative consequences" from a model, as De Regt and Dieks claim. If that is what cognitive science confirms, we have no trouble accepting it, from a naturalist point of view.

is the most adequate one. Like Strevens ([2013], p. 510), we do not wish to tie an account of understanding to one specific theory of explanation. We rather assume here that (a) toy models have explanatory power of either a how-actually or a how-possibly kind, and that (b) there is a philosophical account of explanation that applies to toy models (for instance, by identifying merely possible or actual causes if a causal account is adequate, and so on).[14]

Taking these four refinements into account, we arrive at a refined version of the "simple view" that enables us to distinguish two sorts of understanding. The refined simple view states:

The refined simple view. An individual scientist understands a phenomenon P via model M in context C iff one of the following conditions holds:

1. A scientist S has *how-actually understanding* of phenomenon P via model M in context C iff model M provides a how-actually explanation of P and S grasps M.

2. A scientist S has *how-possibly understanding* of phenomenon P via model M in context C iff model M provides a how-possibly explanation of P and S grasps M.

Being able to distinguish between how-actually understanding and how-possibly understanding will proof to be central in our discussion of understanding in the context of autonomous toy models (Section 4.3).

In sum, the refined simple view is a promising candidate for analyzing the kind of understanding that scientists acquire through toy models, because the refined simple view has room for understanding via idealized models.

4 Two Kinds of Understanding With Toy Models

If one accepts the refined simple view, do scientists obtain understanding with toy models? In this section, we will provide an answer in the form of the following three claims:

1. An embedded toy model M yields how-actually explanations, if three conditions hold: (a) the (well confirmed) embedding framework theory T permits an interpretation and justification of the idealizations of M, and (b) this interpretation and justification is compatible with the veridicality condition.

[14] We are sympathetic to broadly counterfactual accounts of explanation, such as Woodward's ([2003]), Saatsi and Pexton ([2013]), and Reutlinger ([forthcoming]).

If one grasps the how-actually explanation provided by an embedded toy model satisfying the conditions (a) and (b), then one has how-actually understanding. (Section 4.1)

2. *Some* autonomous toy models do not provide how-actually understanding, because major interpretations of idealizations (McMullin's approach, minimalism and dispositionalism) do not support an interpretation and justification of the relevant idealizations of these toy models that is compatible with the veridicality condition. In other words, major interpretations of the relevant idealizations do not support the claim that all autonomous toy models provide how-actually understanding. (Section 4.2)

3. There are central examples of autonomous models that are best interpreted as providing how-possibly explanations and, respectively, how-possibly understanding. This sort of understanding is valuable, because it has (what we call) important modal, heuristic, and pedagogical functions in scientific research and science education. (Section 4.3)

4.1 Embedded toy models and how-actually understanding

An embedded toy model M provides how-actually understanding, if the following (sufficient) conditions hold, or so we argue: (a) the (well confirmed) embedding framework theory T permits an interpretation and justification of the idealizations of M, and (b) this interpretation and justification is compatible with the veridicality condition.

To see why, consider once more our example of the Sun-plus-one-planet toy model. Recall from Section 2.1 that the Sun-plus-one-planet model is an embedded toy model, because (1) it is the model of an empirically well confirmed – at least within an appropriately restricted domain of application – framework theory (Newtonian Mechanics), (2) it is simple (as it describes, for instance, a physical system of only two interacting bodies), and (3) it is idealized (as it deliberately disregards the influence of other planets and only takes into account their gravitational interaction), and (4) it is a model of an actual target phenomenon (such as the earth orbiting around the Sun).

The capacity of this model to provide how-actually understanding depends on three conditions:

(a) In the case of this particular model, one possible way to interpret and justify idealizations is McMullin's account of idealizations (McMullin [1985]). Following McMullin's strategy, we can consider the Sun-plus-one-planet model to be at least approximately true of the target system – i.e. of the real orbit of the earth around the sun. The idealizations of the model are justified

pragmatically. That is, the purpose of idealizing is to turn the calculation of the orbit into a mathematically tractable problem.

(b) McMullin's strategy is compatible with the veridicality condition. We can take the Sun-plus-one-planet model to be approximately true of the target system. Moreover, the Sun-plus-one-planet model can ultimately be de-idealized. Newtonian mechanics provides the theoretical resources for constructing a de-idealized model that generates more accurate predictions about the target than the toyish Sun-plus-one-planet model. The framework theory is a guide for scientists how to include the other planets and the moon(s) and to eventually calculate the improved orbit of the planet under consideration. Hence, in this particular case, the framework theory functions as a guide to de-idealization (for a more detailed exposition and discussion of this approach, see McMullin [1985]; also Weisberg [2013], p. 99). Hence, McMullin's strategy provides a way to defend the veridicality condition by removing the idealizations.

Regarding condition (b), we hasten to stress that McMullin's strategy is, of course, not the only way to interpret and justify idealizations. (We also do not endorse the strong claim that it is possible to de-idealize each and every idealization.) In fact, other interpretative and justificatory strategies can be adopted to complement McMullin's strategy. In particular, we take minimalist and dispositionalist accounts of idealizations to be promising complements, neither of which necessarily involves de-idealization.

We will turn to a more detailed exposition of dispositionalism and minimalism in Section 4.2. We will argue that these accounts of idealizations do not warrant the claim that all *autonomous* toy models provide how-actually understanding. However, regarding *embedded* toy models, we adopt a different dialectic strategy: for the sake of brevity, we mainly rely on the works of others who have convincingly argued that dispositionalism and minimalism apply to embedded toy models, if McMullin's strategy does not. For instance, we assume as uncontroversial Strevens' ([2008]) minimalist approach to idealizations in statistical mechanics, Hüttemann's ([2004], [2014]) argument in favor of dispositionalism regarding certain idealizations in quantum mechanics, and Thebault et al.'s ([forthcoming]) appeal to a combination of minimalist and dispositionalist strategies in the context of statistical mechanics. Ultimately, our argument does not depend on the strong assumption that *all* idealizations figuring in embedded toy models can be given a veridical interpretation. Our main concern is merely a conditional one: embedded toy models yield how-actually understanding, if some strategy for interpreting and justifying idealizations is applicable and the result of applying the strategy is compatible with the veridicality condition.

In sum, we take embedded toy models to provide how-actually explanations, if (a) the (well confirmed) framework theory permits or provides the means to interpret and justify the idealizations and abstractions of the embedded toy model (for instance, by appealing to McMullin's strategy, minimalism, or dispositionalism), and (b) this interpretation and justification does not violate the veridicality condition. Although a discussion of further examples clearly exceeds the scope of this paper, we are confident that the same treatment applies to other examples of embedded toy models (see Section 2.1), such as the Ising model, the ϕ^4-theory, and, perhaps, Fisher's sex ratio model.

4.2 Against a how-actually interpretation of *all* autonomous toy models

One might hold the view that not only embedded but also autonomous toy models yield how-actually understanding and how-actually explanations. However, major accounts of idealizations (McMullin's strategy, minimalism, and dispositionalism) do not support the claim that all autonomous toy models provide how-actually understanding.

Prima facie, one possible way to argue for the claim that autonomous toy models yield how-actually understanding consists in relying on McMullin's strategy. However, McMullin's strategy heavily depends on the existence of a more general and well confirmed framework theory guiding the de-idealization process. But it is precisely this theory that does not exist in the case of autonomous models. For this reason, McMullin's strategy is a non-starter for someone who wishes to defend the claim that autonomous toy models provide how-actually understanding. (But see below for a qualification regarding different de-idealization strategies in the context of autonomous toy models).

In the current literature, there are two main alternatives to McMullin's strategy: minimalism and dispositionalism. If applicable, both minimalism and dispositionalism entail that idealized (toy) models provide how-actually information about their target system. We will first introduce the two accounts of idealizations, and then determine whether these accounts are applicable to all autonomous toy models.

1. *Minimalism:* The minimalist view of idealized models is one promising strategy for supporting an how-actually interpretation of autonomous toy models. As introduced in Section 3.2, Strevens relies on the minimalist view to apply his simple view of understanding to idealized models (see Strevens [2008], Chapter 8, [2013], p. 512).[15] According to minimalism, idealized

[15]See Batterman ([2002]), Pincock ([2012]), and Batterman and Rice ([2014]) for other assessments of minimalism.

models truthfully represent two kinds of facts: (i) facts about a minimal set of explanatorily relevant factors including true causal laws and true statements about initial conditions (which are determined by Strevens' optimizing procedure), and (ii) the fact that some factor is *not* explanatorily relevant (Strevens [2008], pp. 315-29; Weisberg [2013], pp. 100-3). According to Strevens, idealized assumptions represent the latter kind of fact. If the minimalist interpretation applied to autonomous toy models, then such models would provide how-actually understanding about the minimal set of factors explaining the target phenomenon.

2. *Dispositionalism:* According to dispositionalism, an idealized model truthfully represents the disposition of a (physical, biological, or economic) system to behave if other disturbing causes were absent (Cartwright [1989]; Hüttemann [2014]). That is, an idealized assumption describes a counterfactual situation in which a particular factor is taken to be absent (although it frequently occurs in actual situations) and the target system is isolated from the influence of that particular factor. If the dispositionalist interpretation applied to autonomous toy models, then such models would provide how-actually understanding about the *actual* disposition of the target system.[16]

Now, let us check whether minimalism or dispositionalism can be applied to autonomous toy models.

Let us consider minimalism first. Strevens ([2008], Section 8.3) argues that the idealizations figuring in statistical mechanics (embedding the ideal gas law) can be interpreted in accord to the minimalist interpretation, i.e. as statements about what does not make a difference for the occurrence of the target phenomenon. For present purposes, we have no qualms with this particular example. However, the same does not seem to hold for the Schelling model and the DY-model – our examples of autonomous toy models. If minimalism were true of the Schelling model and the DY-model, then all of the idealized modeling assumptions would have to refer to explanatorily irrelevant factors. But it is far from clear that this is the case. Regarding the Schelling model, one cannot simply hold without further argument that the following modeling assumption, among others, refer to explanatorily irrelevant factors: (a) that every agent knows how many agents of each color live in their environment and (b) that social and economic factors do not have an

[16] Although Mäki ([1992], [2011]) agrees with Cartwright and Hüttemann that many idealizations are "isolations", his approach significantly diverges from dispositionalism. Mäki's advocates a "functional decomposition approach" to scientific models, which comprises not only his view of idealizations but also a pragmatically constrained theory of representation and a sophisticated theory of truth. In this paper, we focus solely on the dispositionalists. A discussion of Mäki's approach – that might well be an alternative to both minimalism and dispositionalism – will be a task for future work.

influence on segregation. In the case of the DY-model, it seems to be explanatorily relevant, contrary to minimalism, (i) whether many-agent interactions (as opposed to binary interactions) occur, (ii) whether the quantity of money is indeed conserved, and (iii) whether agents exchange all (as opposed to some) of their money (Thebault et al. [forthcoming]). In sum, we think that minimalism does not straightforwardly apply to two paradigm instances of autonomous toy models. Thus, minimalism cannot be used to defend the claim that all autonomous toy models provide how-actually understanding.[17] (This result is, of course, not an objection to minimalism as an account of idealizations.)

Now let us turn to dispositionalism. A dispositionalist asserts that, for instance, the DY-model describes a disposition of agents to behave in the absence of many-agent interactions and of rational expectation of agents. Similarly, a dispositionalist takes the Schelling model to apply if the causal influence of certain economic factors (such as income) were absent and if there were no ignorance about the color of other agents the neighborhood of each agent. Unlike the minimalist, the dispositionalist is not committed to the claim that, for instance, many-agent interactions or economic factors are explanatorily irrelevant. However, dispositionalists face another problem: they have to justify how the DY-model is applicable in 'non-ideal' situations, i.e. in the actually quite frequent kind of situation, in which many-agent interactions do in fact occur in economic exchanges, agents actually have (more or less) rational expectations, economic factors do make a difference for segregation, and so on. Meeting this challenge is difficult in the case of the autonomous toy models, because, unlike in the case of embedded toy models, there are no general dynamical laws (of a framework theory) that might help us to determine what will happen if 'disturbing factors' are in fact present and, thereby, guide the application of the model in a non-ideal situation.

Hüttemann ([2014]) presents a dispositionalist response to this problem of non-ideal situations. Hüttemann's solution to the problem invokes "laws of interaction" and "laws of composition". This response works well for many embedded toy models – and indeed, in Hüttemann's main examples for laws of interaction and composition are part of quantum mechanics as a framework theory. However, autonomous toy models are typically not equipped with such laws. For this reason, Hüttemann's defense of dispositionalism does not carry over to autonomous toy models, at least not in general.

Thus, dispositionalism about idealizations cannot be readily used to defend the claim that all autonomous toy models provide how-actually understanding.

In sum, there are some autonomous toy models for which the following holds:

[17] As a referee pointed out, one promising strategy for defending a minimalist approach to the Schelling model might consist in exploiting the robustness of the model (see Muldoon et al. [2012]). See Reutlinger and Andersen [unpublished] for a critical discussion of a particular kind of robustness approach in the context of explanations.

neither minimalism nor dispositionalism can be readily used for supporting the view that these autonomous toy models yield how-actually understanding.

Let us qualify this claim in two ways.

First, we emphasize that we endorse an existential claim: there are some autonomous toy models that do not provide how-actually understanding, because McMullin's strategy, minimalism, and dispositionalism do not support the claim that all autonomous toy models provide how-actually understanding. We do not defend the stronger claim that no autonomous toy model yields how-actually understanding.[18]

Second, we do not claim that it is impossible to de-idealize autonomous toy models. In fact, we will briefly discuss an attempt to de-idealize the DY-model in Section 4.3. What matters for our concerns is that the de-idealization in the case of autonomous toy models is not guided by an embedding framework theory; de-idealization, in this context, is rather a matter of empirically readjusting a model. Autonomous toy models often serve the heuristic purpose of constructing more 'realistic' and often (but not necessarily) also more complex models of the target phenomenon. However, in cases where the construction of more realistic and often (but not necessarily) also more complex autonomous models is possible, these autonomous models tend to lose their 'toy' – that is, idealized and simple – character. Especially the gain in complexity has an interesting effect: it diminishes the capacity of these models to provide understanding, because it is mainly the simplicity of toy models that permits scientist to grasp them. Consider the following example for illustrating this point. The sociologist Peter Hedström developed an empirically calibrated agent-based model to capture the phenomenon of unemployment in the Stockholm metropolitan area during the period of 1993 to 1997 (Hedström [2005]).

Hedström's agent-based model is an autonomous model and it is intended to be a realistic (i.e. containing relatively few idealizations) and complex model: the number of agents in this model is 87,924 and their states (such as age, marital status, previous unemployment experiences, immigration background etc.) are supported by demographic data about 20 to 24 years olds in the Stockholm metropolitan area in the time period at issue. According to the rules of this model, an agent changes her/his state (from unemployed to employed) depending, for instance, on how many of their neighbors are unemployed. Hedström's model is autonomous but it is not simple and not strongly idealized (if compared to the Schelling model). Its lack of simplicity makes it unlikely (if not impossible) for researchers to cognitively 'grasp' the model. Hence Hedström's model does not satisfy one necessary condition for understanding.

[18]See Strevens ([2003], Chapters 4 and 5), for potential candidates of such models from the life and social sciences.

One reaction to our line of argument might be to improve minimalism and dispositionalism and to argue that these improved accounts of idealizations do in fact apply to *all* autonomous toy models. We have no proof that this strategy has got to be unsuccessful. However, we believe that the attempt to apply minimalism and dispositionalism to some autonomous toy models faces problems that are serious enough to explore an alternative approach. This alternative approach rejects the idea that all autonomous toy models provide how-actually understanding. Autonomous toy models sometimes yield another kind of understanding: how-possibly understanding.

4.3 The how-possibly interpretation of some autonomous toy models

Let us suppose that there is a considerably large class of autonomous toy models that cannot be interpreted as providing how-actually understanding – for the reasons given in Section 4.2. We hold that the Schelling model and the DY-model are members of this class. If some autonomous toy models fail to provide how-actually understanding, what kind of understanding do they provide, if any?

Our proposal is to take those autonomous toy models to yield *how-possibly understanding*. Applying the refined simple view, a scientist S has how-possibly understanding of phenomenon P by using an autonomous toy model M in context C iff M provides a how-possibly explanation of P and S grasps M. For instance, we take the Schelling model to *explain how it is possible* that racial segregation occurs; and we take the DY-model to *explain how it is possible* that income distributions with specific qualitative features emerge. Both models only provide a potential explanation of a general pattern (that is, segregation and a certain kind of income distribution) and this pattern happens to be actually instantiated (for instance, the pattern of segregation is actually instantiated in Detroit, and a certain kind of income distrbution is contingently instantiated in the United States). Both models (and the evidence we have) do not tell us whether they have correctly identified the actually relevant explanatory factor(s).[19]

The question arises why scientists are interested in how-possibly understanding, as one appears to gain considerably less from how-possibly than from how-actually explanations. De Regt and Dieks, for example, are very quick in dismissing how-possibly understanding as "mere intelligibility" (which is clearly intended as a derogatory term) and in taking how-possibly understanding to be (necessarily) on a par with pseudo-science, such as astrology (see Section 3.1). In fact, how-possibly understanding plays a central and legitimate role in research and in science ed-

[19]See Forber ([2010]), Cuffaro ([2015]), and Fumagalli ([2015]) for a sophisticated discussion of various readings of how-possibly explanations.

ucation. More precisely, we hold that there are at least three central epistemic functions of how-possibly understanding: (1) the modal function, (2) the heuristic function, and (3) pedagogical function. We will now describe each of these functions in more detail.

(1) *Modal function.* How-possibly explanations are valuable if the phenomenon to be understood is a 'modal phenomenon' – that is, if scientists want to understand whether and why some phenomenon is possibly or necessarily the case (Grüne-Yanoff [2013], pp. 855-9; Weisberg [2013], pp. 118-9; Cuffaro [2015]). One of the most famous illustrations of the modal function of toy models is Schelling's model of segregation. Schelling's model is concerned with the question whether it is possible to understand the emergence of segregated neighbourhoods without assigning explicitly racist attitudes to agents. Schelling's model shows that, in contrast to the view that segregation is necessarily a result of racism, it is *possible* for segregation to arise in a population of agents following the 30% rule (even if the agents would actually prefer to live in non-segregated cities). If the goal is to explain a 'modal phenomenon', then how-possibly understanding (and explanation) is an appropriate tool for achieving this goal.

(2) *Heuristic function.* How-possibly understanding via autonomous toy models is not always an end in itself. How-possibly understanding often plays a heuristic role in the process of constructing less idealized (and often, but not necessarily, also more complex) models that latch onto the target system more accurately than the original toy model (Hartmann [1995]).[20] For instance, the DY-model has inspired the construction the CCM-model. The latter model includes a 'de-idealization' of the idealized assumption (in the DY-model) that the agents exchange *all* of their money when interacting. Unlike the DY-model, the CCM-model assigns a saving propensity to all agents – that is, the agents exchange *only a fraction* of their money when interacting, as in 'real' economic exchanges. This small alteration comes with a considerable pay-off: the CCM-model captures the relevant data about income distributions more accurately than the original DY-model (Thebault et al. [forthcoming]). Hence, the DY-model – an autonomous toy model – plays a heuristic role in developing a more accurate model (the CCM-model).

(3) *Pedagogical function.* The how-possibly character of autonomous toy models is often also used for primarily illustrative purposes in science education (see Hangleiter [2014]). These models enable students and researchers to quickly grasp the idea behind the solution to a problem, or the description of a phenomenon. Generally speaking, the pedagogical function of toy models is to enable students to learn how to calculate with and how to use a particular model (or theory).

[20]Fumagalli ([2015], Sect. 4.3) defends the claim that (autonomous) toy models can play a heuristic role for constructing how-actually models only if modelers include veridical "additional information or presuppositions" concerning the target system.

Once students have learned how to calculate by practicing with a toy model, the training is put to different uses in the case of embedded toy models and of autonomous toy models. Regarding embedded toy models, science students learn how to calculate with the embedding framework theory (by practicing with toy models initially) in order to prepare the students to mathematically handle less idealized (and sometimes also more complex) models of the embedding framework theory later on. Regarding autonomous toy models, the goal of practicing with a toy model is different: the acquired ability to handle an autonomous toy model mathematically enables students to make use of the toy model in a modal or a heuristic function.

To sum up, we have argued that some central examples of autonomous toy models yield how-possibly understanding (as opposed to how-actually understanding). Moreover, we claimed that scientists value how-possibly understanding because it has a modal and a heuristic function in scientific research, and a pedagogical function in science education.

5 Conclusion

Initially, we characterized toy models as idealized and simple models of natural and social phenomena. Our main question in this paper was whether the epistemic goal of constructing toy models is to obtain understanding of their target phenomena. To support the claim that toy models do indeed provide understanding, we have argued in three steps: first, we introduced and illustrated a distinction between embedded and autonomous toy models. Second, we argued that the refined simple view is a suitable account of understanding in the context of toy models. One key feature of this account of understanding is that it allows for a distinction between how-actually and how-possibly understanding. Finally, we applied the refined simple view to our examples of embedded and autonomous toy models with the following results: (a) an embedded toy model yields how-actually understanding, if certain conditions regarding the veridical interpretation and justification of the model's idealizations are satisfied, (b) McMullin's strategy, minimalism, and dispositionalism do not support the claim that all autonomous toy models provide how-actually understanding, and (c) some autonomous toy models are best interpreted as providing how-possibly understanding. Therefore, the claim that toy models yield understanding can be vindicated, if one allows for two modalities of understanding – that is, how-actually understanding and how-possibly understanding.

Acknowledgements

We would like to thank Seamus Bradley, Adam Caulton, Henk de Regt, Maria Kronfeldner, Sebastian Lutz, Margaret Morrison, John Norton, Michael Strevens, Karim Thebault, David Weberman, and the participants of the workshop "Just Playing. Toy Models in the Sciences" (held in Munich, 2015) as well as audiences in Budapest, Cardiff, Munich, Paris, Santiago de Chile, Seattle, and Tübingen for their supportive and constructive comments. We wish to stress that the two anonymous referees for this journal did a wonderful job; we are grateful for their careful and well-informed critical comments that truly helped to improve the paper. We would also like to acknowledge the financial support of the Alexander von Humboldt foundation, the Münchner Universitätsgesellschaft, and the Studienstiftung des deutschen Volkes.

Alexander Reutlinger
Munich Center for Mathematical Philosophy
Ludwig-Maximilians-Universität Munich
Geschwister-Scholl-Platz 1
80539 Munich
Germany
Alexander.Reutlinger@lmu.de

Dominik Hangleiter
Munich Center for Mathematical Philosophy
Ludwig-Maximilians-Universität Munich
Geschwister-Scholl-Platz 1
80539 Munich
Germany
and
Dahlem Center for Complex Quantum Systems
Freie Universität Berlin
14195 Berlin
Germany
Dominik.Hangleiter@fu-berlin.de

Stephan Hartmann
Munich Center for Mathematical Philosophy
Ludwig-Maximilians-Universität Munich

Geschwister-Scholl-Platz 1
80539 Munich
Germany
Stephan.Hartmann@lmu.de

References

Bailer-Jones, D. [1997]: *Scientific Models: A Cognitive Approach with an Application in Astrophysics*, PhD Thesis, University of Cambridge.

Bailer-Jones, D. [2009]: *Scientific Models in Philosophy of Science*, Pittsburgh: University of Pittsburgh Press.

Batterman, R. [2002]: *The Devil in the Details*, New York: Oxford University Press.

Batterman, R. and Rice, C. [2014]: 'Minimal Model Explanation', *Philosophy of Science*, **81**, pp. 349-376.

Bell, J. L. and Slomson, A. B. [1974]: *Models and Ultraproducts: An Introduction*, New York: Dover Publications.

Butterfield, J. [2011]: 'Less is Different: Emergence and Reduction Reconciled', *Foundations of Physics*, **41**, pp. 1065-1135.

Cartwright, N. [1983]: *How the Laws of Physics Lie*, Oxford: Oxford University Press.

Cartwright, N. [1989]: *Nature's Capacities and Their Measurement*, Oxford: Clarendon Press.

Chang, C. C. and Keisler, H. J. [1990]: *Model Theory*, Amsterdam: North Holland.

Cuffaro, M. [2015]: 'How-Possibly Explanations in (Quantum) Computer Science', *Philosophy of Science*, **82**, pp. 737-748.

De Regt, H. and Dieks, D. [2005]: 'A Contextual Approach to Scientific Understanding", *Synthese*, **144**, pp. 137-170.

De Regt, H. [2009]: 'Understanding and Scientific Explanation', In De Regt, H., S. Leonelli and K. Eigner (eds), *Scientific Understanding. Philosophical Perspectives*, Pittsburgh: University of Pittsburgh Press, pp. 21-42.

De Regt, H., Leonelli, S. and Eigner, K. (eds) [2009]: *Scientific Understanding. Philosophical Perspectives*, Pittsburgh: University of Pittsburgh Press.

Dizadji-Bahmani, F., Frigg, R., and Hartmann, S. [2010]: 'Who's Afraid of Nagelian Reduction?', *Erkenntnis*, **73**, 393-412.

Drăgulescu, A. and Yakovenko, V. [2000]: 'Statistical Mechanics of Money', *The European Physics Journal B – Condensed Matter and Complex Systems*, **17(4)**, pp. 723-729.

Forber, P. [2010]: 'Confirmation and explaining how possible', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, **41(1)**, pp. 32-40.

Friedman, M. [1974]: 'Explanation and Scientific Understanding', *Journal of Philosophy*, **71**, pp. 5-19.

Frigg, R. and Hartmann. S. [2012]: 'Models in Science', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), URL = ¡http://plato.stanford.edu/archives/fall2012/entries/models-science/¿.

Fumagalli, R. [2015]: 'Why We Cannot Learn from Minimal Models', *Erkenntnis*, Online First, DOI: 10.1007/s10670-015-9749-7.

Giere, R. [1988]: *Explaining Science. A Cognitive Approach*, Chicago: University of Chicago Press.

Gijsbers, V. [2013]: 'Understanding, Explanation, and Unification', *Studies in History and Philosophy of Science Part A*, **44(3)**, pp. 516-522.

Grüne-Yanoff, T. [2009]: 'Learning from Minimal Economic Models', *Erkenntnis*, **70(1)**, pp. 81-99.

Grüne-Yanoff, T. [2013]: 'Appraising Non-Representational Models', *Philosophy of Science*, **80(5)**, pp. 850-861.

Hangleiter, D. [2014]: *When Scientists Play: How Toy Models in Science Help Us Understand the World*, Bachelor Thesis, LMU Munich.

Hartmann, S. [1998]: 'Idealization in Quantum Field Theory', In N. Shanks (ed.), *Idealization in Contemporary Physics*, Amsterdam: Rodopi, pp. 99-122.

Hartmann, S. [1999]: 'Models and Stories in Hadron Physics?', In M. Morgan and M. Morrison (eds.), *Models as Mediators*, Cambridge: Cambridge University Press, pp. 326-346.

Hartmann, S. [2001]: 'Effective Field Theories, Reduction and Scientific Explanation', *Studies in History and Philosophy of Modern Physics, Part B*, **32**, pp. 267-304.

Hedström, P. [2005]: *Dissecting the Social*, Cambridge, UK: Cambridge University Press.

Hempel, C. [1965]: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Humphreys, P. [2004]: *Extending Ourselves*, New York: Oxford University Press.

Hüttemann, A. [2004]: *What's Wrong With Microphysicalism?*, London: Routledge.

Hüttemann, A. [2014]: 'Ceteris-paribus Laws in Physics', *Erkenntnis*, **79**, pp. 1715-1728.

Kitcher, P. [1981]: 'Explanatory Unification', *Philosophy of Science*, **48**, pp. 507-531.

Kuhn, T. [1977]: 'Objectivity, Value Judgment, and Theory Choice', In his *The Essential Tension*, Chicago: University of Chicago Press, pp. 320-339.

Lipton, P. [2009]: 'Understanding Without Explanation', In De Regt, H., S. Leonelli and K. Eigner (eds), *Scientific Understanding. Philosophical Perspectives*, Pittsburgh: University of Pittsburgh Press, pp. 43-63.

Mäki, U. [1992]: 'On the Method of Isolation in Economics', *Poznan Studies in the Philosophy of the Sciences and the Humanities*, **26**, pp. 319-354.

Mäki, U. [2011]: 'Models and the Locus of their Truth', *Synthese*, **180**, pp. 47-63.

McMullin, E. [1985]: 'Galilean Idealizations', *Studies in the History and Philosophy of Science*, **16**, 247-273.

Meinel, C. [2004]: 'Molecules and Croquet Balls', In S. de Chadarevian and N. Hopwood (eds), *Models. The Third Dimension of Science*, Stanford: Stanford University Press, pp. 242-275.

Morgan, M. and Morrison, M. (eds) [1999]: *Models as Mediators*, Cambridge: Cambridge University Press

Muldoon, R., Smith, T., and Weisberg, M. [2012]: 'Segregation that No One Seeks', *Philosophy of Science*, **79**, 38-62.

Norton, J. [2012]: 'Approximation and Idealization: Why the Difference Matters?', *Philosophy of Science*, **79**, pp. 207-232.

Pincock, C. [2012]: *Mathematical and Scientific Representation*, New York: Oxford University Press.

Reutlinger, A. [forthcoming]: 'Is There A Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation', *Philosophy of Science*.

Reutlinger, A. and Andersen, H. [unpublished]: 'Abstract Isn't Non-Causal'.

Saatsi, J. and Pexton, M. [2013]: 'Reassessing Woodward?s account of explanation: regularities, counterfactuals, and non-causal explanations', *Philosophy of Science*, **80**, pp. 613-624.

Schelling, T. [1971]: 'Dynamic Models of Segregation', *Journal of Mathematical Sociology*, **1**, pp. 143-186.

Sober, E. [1984]: *The Nature of Selection*, Chicago: The University of Chicago Press.

Strevens, M. [2003]: *Bigger Than Chaos*, MA: Harvard University Press.

Strevens, M. [2008]: *Depth*, Cambridge, MA: Harvard University Press.

Strevens, M. [2013]: 'No Explanation without Understanding', *Studies in History and Philosophy of Science*, **44**, pp. 510-515.

Sugden, R. [2000]: 'Credible Worlds: the status of theoretical models in economics', *Journal of Economic Methodology*, **7**, pp. 1-31.

Thebault, K., Bradley, S., and Reutlinger, A. [forthcoming]: 'Modeling Inequality', *British Journal for Philosophy of Science*.

Trout, J. [2002]: 'Scientific Explanation and the Sense of Understanding', *Philosophy of Science*, **69**, pp. 212-233.

Van Fraassen, B. [1980]: *The Scientific Image*, Oxford: Oxford University Press.

Weisberg, M. [2013]: *Simulation and Similarity*, New York: Oxford University Press.

Woodward, J. [2003]: *Making Things Happen*, New York: Oxford University Press.