Do ML models represent their targets?

Emily Sullivan
Utrecht University
Forthcoming in *Philosophy of Science*

Abstract: I argue that ML models used in science function as highly idealized toy models. If we treat ML models as a type of highly idealized toy model, then we can deploy standard representational and epistemic strategies from the toy model literature to explain why ML models can still provide epistemic success despite their lack of similarity to their targets.

1. Introduction

Most attention on complex machine learning (ML) models used in science has centered around issues of opacity, such as the nature of opacity (Creel 2020, Boge 2022) and its epistemic consequences for science (Duede 2023).[1] While some have argued that ML models can still provide understanding of phenomena despite their opacity (Meskhidze 2021; Sullivan 2022a), others demur (Räz and Beisbart 2022). However, before the epistemic consequences of *opacity* become salient, there is an underexplored prior question of *representation*. If ML models used in science do not represent real-world targets in any meaningful sense, how can ML models provide understanding in the first place?

The problem is that it seems as though ML models *do not* represent their targets in any meaningful sense. For example, the similarity view of representation seems to exclude the possibility that ML models can represent phenomena. According to the similarity view (Mäki 2009; Giere 2004; Weisberg 2013), for a model to represent some phenomenon requires that the model be sufficiently similar to its target. However, ML models use methods of finding feature relationships that are highly divorced from their target systems, such as relying on complex computations or loose correlations instead of causal relationships. Moreover, the data that models are trained on can be manipulated by modelers in a way that reduces similarity. For example, the well-known melanoma detection ML model (Esteva et al. 2017) is trained on manipulated and resized variations on images viewed by dermatologists and interprets RGB arrays of pixels. Thus, if the similarity view is right, then even if model opacity *qua* opacity

[1] In this paper, my focus is on deep learning neural network architectures. However, analogous arguments could be made toward random forests or Bayesian nets, or other ML techniques that engage in idealization.

does not get in the way of understanding, ML models may still fail to enable understanding of phenomena because they fail to represent phenomena. This gives rise to the following hypothesis concerning the epistemic status of ML models:

> **ML Representation Hypothesis**
> Complex or opaque ML models fail to enable understanding of real-world phenomena because ML models are not similar to, and therefore, fail to represent their targets.

In this paper, I argue that we should reject the ML representation hypothesis. Specifically, I argue that ML models function representationally and epistemically in a similar way as highly idealized toy models do in science. If we treat ML models as functioning as highly idealized toy models, then there are two ways of rejecting the ML representation hypothesis. We can (i) adopt an interpretative view of representation (Nguyen 2020), in which case a compelling story can be told that ML models do in fact represent their targets. Or, (ii) if adopting an interpretative view of representation is unpalatable, we can still reject the ML representation hypothesis by appealing to the epistemic status of idealizations and adopting the following idealization failure hypothesis instead:

> **Idealization Failure Hypothesis**
> Complex or opaque ML models fail to enable understanding of real-world phenomena when there is idealization failure.

Adopting the idealization failure hypothesis would mean that evaluating the epistemic virtues and limitations of ML models requires identifying and evaluating the idealizations within ML modeling, not necessarily striving to make ML models more similar to their targets.

The paper proceeds as follows. First, I introduce the epistemic and representational puzzle that toy models introduce and possible solutions to the puzzle (section 2). Second, I argue we should think of ML models as functioning as highly idealized toy models and apply the same solutions as we do with toy models (section 3). Lastly, in section 4, I discuss the benefits of adopting the idealization failure hypothesis as a necessary step for evaluating the epistemic status of ML models. In the end, even though ML models seem to be the opposite of highly idealized toy models, there are a number of representational and epistemic similarities between them. Thus, if we accept that highly idealized models can either represent phenomena or still enable understanding in the absence of representation, then the same holds for ML models.

2. The puzzle of toy models

Toy models are models that are (i) extremely simple, (ii) highly idealized, and (iii) said to refer to a target system in the world (Reutlinger et al. 2018). Common examples include the hawk-dove game or Lotka-Volterra equations in population biology, Ising physics model, and Schelling's segregation model. Focusing on the latter, Schelling (1971), in seeking to explain why so many cities are racially segregated, developed a simple toy model on a checkboard. The model shows that if people move based on a simple preference that a certain percentage of their neighbors are the same, then a segregated board is the equilibrium state. Schelling's model makes several idealizing assumptions that make it unlike any city on earth. There is no cost to moving, people are free to move to any empty space, there is no institutional racism, and people move based on a single preference for 'like neighbors.'

The use of toy models like this in science raises an interesting puzzle. How is it that a model that is so highly idealized and so divorced from real world phenomena can give us any epistemic insight? Schelling's model is not at all similar to real cities, so how could it be said to represent real cities in any meaningful sense? How can it provide any understanding into why real-world cities are segregated? There are two central ways that philosophers of science approach solutions to this puzzle. The first is to reject the underlying similarity view of representation that seems implicit in how the puzzle is posed and offer an alternative view of representation where similarity is not the locus of representational content. The second approach considers the unique epistemic status of idealizations. I consider each in turn.

2.1. Interpretive view of representation

The puzzle from toy models implicitly assumes a similarity view of representation where a model must be 'sufficiently similar' to its target for accurate representation (Mäki 2009; Giere 2004; Weisberg 2013). However, there are alternative, and influential, theories of representation that do not require model similarity with its target (Nguyen 2020; Suárez 2015; Frigg and Nguyen 2018). In this paper, I will focus on Nguyen's (2020) interpretative view where toy models are representational in the sense that they license truthful inferences about the target system. Importantly, for a model to license truthful inferences, similarity does not matter, but an interpretative function that can map model-facts to claims that could also serve as a 'translation key' that connects the model to its target (Frigg and Nguyen 2018; Nguyen 2020). While on the similarity view of representation Schelling's model does not represent its target in virtue of the fact that Schelling's model is not sufficiently similar to real cities, on the interpretative view of representation, Schelling's model does in fact represent real world segregated cities. Using an interpretative function regarding the underlying mechanism driving

Schelling's model—that acting on certain preferences can lead to segregated equilibriums—gives rise to a true claim about segregation namely that "a city whose residents have a weak preference regarding the [race] of their neighbors has a susceptibility toward global segregation" (Nguyen 2020, pg. 1030). On the interpretive account, toy models are unique in that they (i) move from an inevitability in the model to a susceptibility claim about the target, and (ii) move from a specific model-fact to a less specific claim about the target (pg. 1030). Schelling's model provides a model-based inevitability regarding the certainty of segregation to a less specific susceptibility claim about real-world populations.

Notice that on the interpretative view of representation, the puzzle of toy models dissolves. The reason toy models are useful and successful tools in science is because they provide us with representational content regarding their target systems when they enable us to make true inferences regarding the target.

2.2 Epistemology of toy models

A second set of solutions to the puzzle of toy models considers the epistemology regarding how idealizations, despite their falsity, can still enable understanding. Such a solution need not appeal to representation *per se*. While some, such as Elgin (2017), deploy a representational view of idealizations as representing-*as*, many others go a different route. For example, on Potochnik's (2017) account, idealizations are *mis*representations with the falsehood of the idealization playing an active role. On the holistic distortion view (Rice 2019), true and false aspects of a representation cannot be separated, so idealized models become holistic misrepresentations. Others argue that idealizations are *non*-representational. On Lawler's (2021) extraction view, idealizations play an enabling role and are not constitutive of scientific representations. Carrillo and Knuuttila (2022) propose an artifactual account of idealization that actively rejects the need for the representation question, and instead focuses on the way that idealizations, and models, are tools for epistemic purposes.

Despite differences between these views on the representational status of idealizations, all of these theorists agree that idealizations either themselves constitute epistemic success or can help point to relevant truths that enable success; in the worst-case, idealizations serve as convenience crutches (Sullivan and Khalifa 2019). Importantly though, those that separate the idealization question from the representation question have a tempered view of the epistemic role the heavy idealizations in toy models provide. Toy models merely provide how-possibly explanations if the adequate link between the model and target are in some way in doubt. A common interpretation of the epistemic success of Schelling's model is that it only provides us

with how it is possible segregation could occur in real-world populations, while failing to provide an explanation why actual cities are segregated because it is not embedded within a larger confirmed theory (Reutlinger et al. 2018). Understanding real-world phenomena requires establishing empirical links outside of the model (Sullivan and Khalifa 2019; Sullivan 2022a). Thus, the epistemic solution to the puzzle of toy models exercises caution regarding the extent of the scientific understanding toy models may provide, but nevertheless can still account for why toy models are useful and enable scientific understanding, albeit understanding of possibilities.

3. ML models as toy models

Current ML models are not the kind of highly idealized models that philosophers of science often discuss alongside idealization. They are complex instead of simple, they are new instead of mature, and they are not constructed with built-in theoretical assumptions or what Knüsel and Baumberger (2020) call process-models, where model equations explicitly refer to processes in the target system. In contrast, the inner decision points in ML models are not tracking these types of processes; instead, an ML model is essentially trying to minimize loss and satisfy a defined objective function running a series of mathematical computations in vector space (Boge 2022). ML models are often used precisely because causal processes are unknown, or because researchers are interested in seeing whether there are overlooked patterns of interest. Despite these differences with traditional model-based science, I want to suggest that ML models used in science function in a similar way as toy models. First, ML models (including predictive models) used in science aim to refer to various real-world phenomena (e.g. models of new physics, disease indicators, climate patterns, etc).[2] Second, they engage heavily in idealizations across the ML modeling pipeline. Third, central questions regarding representation and epistemic success seem to mirror that of toy models.

3.1. ML Idealizations

Even though ML models are highly complex data-driven models and are not the simple type of model often thought of as toy models, ML models still engage in significant idealization throughout the ML modeling pipeline.[3] Table 1 provides a (non-exhaustive) overview of where idealizations may appear in the ML pipeline, ranging from ML architectures to data collection, model training, the learned algorithm that results from training, and generalizing to novel cases.

[2] There could be cases where an ML model does not aim to refer to real-world phenomena. In these cases, the ML models could be closer to so-called targetless models.

[3] Following Levy (2021), I take abstraction to be a relation between two representations and idealization to be a relation between a model and the world. Thus, something can be both an abstraction and an idealization.

Table 1. Idealizations across ML Pipeline

ML architectures	Idealizations introduced by architecture choice. *Example: Fully connected networks assume independence of input variables*
data choices	Idealizations introduced through data choices. *Example: data manipulation as part of data processing*
Model training	Idealizations that emerge through model training. *Example: Backpropagation techniques finding local minima through gradient descent*
Learned ML algorithm	Idealizations present in the ML model after training. *Example: ML models relying on reliable proxies or not relying on relevant causal influences in the target*
Generalization	Idealizations that are created when applying a ML model to novel data. *Example: Deploying model on data that is dissimilar to testing and training data*
Explaining model decisions	Idealizations that are introduced when applying explainability methods to explain ML model decisions. *Example: Linear approximations of local decisions*[4]

For example, ML architectures idealize. The simplest type of neural network (NN) architecture is a fully connected NN that assumes causal independence among input variables. All input variables are treated as independent even though we know that there is interdependence between input features. Furthermore, as the network 'learns,' each new layer 'forgets' weights and influences from previous layers. Such an architecture is an idealization because many phenomena that NNs aim to capture do have causal inter-dependence among variables. For example, a deep NN (DNN) model may seek to predict disease indicators using input data that has known strong correlations and causal influences in the data. If researchers use a fully-connected DNN, these causal dependencies are idealized away in the initial architecture. Other more sophisticated architectures, like transformer models, de-idealize these assumptions. Specifically, transformers add attention layers that address the 'forgetful' problem in fully connected NN but may introduce different idealizations in the process. Importantly, what makes something an idealization is context dependent. While in many cases fully connected NNs constitute an idealization, there could be other cases where this is not an idealization because we have good reason to believe there is causal independence between variables. Not every result of a mathematical process deployed in ML will itself constitute an idealization. It ultimately depends on the relationship between the target and what results from the mathematical processes (Levy 2021).

3.2. Representation and ML

Since ML models engage in idealizations and can find patterns of interest in a way divorced from underlying real-world processes, like toy models, it seems like ML models do not actually represent their targets and that we should accept the ML representation hypothesis regarding

[4] See Fleisher (2022) for a discussion of idealizations in explainable AI.

their epistemic status. And indeed, in a recent paper, Tamir and Shech (2022) argue that ML models can fail to represent their targets, undermining their epistemic status. One example they highlight is the case of Esteva et al.'s (2017) melanoma classifier that reportedly does better at identifying melanoma compared to dermatologists. The ML model was trained on distortions of the original dermatological images. For example, the Inception-v3 model that was used requires input images of 299x299 pixels (Estava p. 119).[5] This means the analyzed data is dissimilar to the original larger images as well as dissimilar to the phenomena at hand (i.e. the way moles and melanoma appear on the skin). This is an example of what I am calling a data processing idealization. Tamir and Shech (2022) suggest that the lack of similarity resulting from data processing idealizations can undermine how well ML models represent phenomena.

However, the worry here is largely grounded in implicitly adopting a similarity view of representation. If we adopt an interpretive view of representation—as we do with toy models—we get a different result. Nguyen (2020) considers an analogous case comparing an unmodified image of Obama to a color inverted picture. According to a similarity view of representation, the ordinary image of Obama is more similar to him and thereby is a more accurate representation of Obama. However, on the interpretive view of representation, both pictures have the same representational content because both pictures can license the same inferences about Obama. The difference is that the function one should use to map model-facts to claims differs between the two images. The inverted photo requires a `color_inversion()` function that converts the inversion to derive true inferences. The same is true of various data idealizations in ML modeling. In the dermatological example, the data processing that involves image distortion and representing images as RGB arrays requires the right interpretive function. For example, a `resize_image()` or `array_to_pic()` function to map model-facts to claims about a mole. The fact that the data becomes less similar to the target does not imply it becomes less *representative* of the target, with the right interpretive function.

The interpretive view can be pushed even further regarding other distortions and idealizations in ML modeling, even to the aspects of ML models that seem the most dissimilar to their targets, such as finding patterns by manipulating vector space. For example, Boge (2022) argues that the hyper-parameters within a DNN in the best case give us ambiguous meanings, and in the worst case, are simply meaningless and thus cannot represent phenomena. However, it is compatible with the interpretative view that some lower-level internals of a DNN may not represent; as long as we can map abstract high-level ML model-facts to claims, the

[5] See https://cloud.google.com/tpu/docs/inception-v3-advanced) for more detail on Inception-v3.

ML model can be said represent phenomena. Knowing the high-level model-facts—that these set of features contributed most to the decision—is possible through interpretability techniques (Creel 2020, Sullivan 2022a).

Mapping ML model-facts to claims will likely involve a two-step process. Consider the dermatology example. First, an interpretability method must map a series of weights in a DNN to a set of understandable features (e.g. attention layer map, SHAP values, etc), where the interpretability method is itself an idealized model (Fleisher 2022). Second, an interpretative function is needed that connects the set of understandable features the model relies on to the target phenomena. For example, just as with toy models, the dermatology ML model (i) moves from an inevitability in the model concerning feature importance and classification to a susceptibility claim about the target system, namely that certain pigmentation differences indicate a susceptibility to be a melanoma. And (ii) we move from a very specific (and very local) model fact—that this particular mole was classified as a melanoma—to a less specific claim about identifying cases of melanoma in real cases (i.e. that it is possible that these features are indicators of melanoma). If anything, since evaluating ML models, due to model opacity, relies on another idealized model (interpretability methods), ML models could be described replying on idealization *more* than toy models.

The challenge on the interpretive view of representation in the case of ML becomes finding the correct interpretative map that can reinterpret the idealizations and distortions that the ML model makes to the actual target. Such an interpretive map may not be known depending on the specific model and target phenomena. However, notice that this question—the absence of a known map—is a different consideration from the ML representation hypothesis that focuses on representation with regard to similarity. It might be that this is where ML opacity starts to become an issue. Moreover, interpretability techniques themselves might be subject to idealization failures that can prevent understanding, which again signals that the idealization failure hypothesis is better suited to evaluate the epistemic status of ML models.

3.3. Epistemology of ML models

Recall that a second approach to solving the problem of toy models is to understand idealizations' epistemic value. Here too ML models in science function epistemically as toy models. In the toy model literature, Reutlinger et al. (2018) distinguish highly idealized toy models that are embedded into and are models of an empirically well confirmed theory from 'autonomous' models, where the science is still out, and are successful in virtue of enabling how-*possibly* explanations or how-*possibly* understanding. ML models function largely the

same way. Sullivan (2022a, 2022b) argues it is other evidential support external to the model that can render a ML model as facilitating or inducing understanding. Most ML models, on this view, have a high level of 'link-uncertainty' such that the ML model merely provide a type of how-possibly explanation, like toy models. Zednik and Boelsen (2022) also argue that ML models in science chiefly serve as hypothesis generating tools. Indeed, toy models are largely circumscribed as playing such a heuristic role (Sullivan and Khalifa 2019).

Even if ML models may only provide how-possibly explanations, such explanations can still facilitate scientific understanding. How-possibly explanations are valuable heuristics to build better theories, discover hypotheses for future research, and provide answers to genuine questions regarding the scope of (im)possibilities (see Verreault-Julien 2019). Thus, again, by taking ML models in science as functioning as highly idealized scientific models, we can reject the ML representation hypothesis regarding the epistemic status of ML models. ML models can still provide understanding (of possibilities) without being similar to their targets, which can explain the success of ML models despite their known limitations.

4. Idealization failure hypothesis

The discussion so far has centered around reasons we can reject the ML representation hypothesis regarding the epistemic status of ML models. In this last section, I want to suggest an alternative hypothesis that treats idealization evaluation, instead of representation, as centrally important for assessing the epistemic status of ML models.

> **Idealization Failure Hypothesis**
> Complex or opaque ML models fail to enable understanding of real-world phenomena when there is ML idealization failure.[6]

The idealization failure hypothesis does not appeal to representation *per se* since, as discussed above (Section 2.2.), on several accounts of idealization, idealization lacks representational status or might misrepresent. Thus, one benefit of adopting the idealization failure hypothesis is that it does not necessarily require adopting a strong position regarding the representational status of ML models. But what does it mean to have idealization failure? And when do ML model idealizations fail? I will have to leave a complete answer to these questions for further work, but there some avenues worth exploring.

[6] There may be other ways we should evaluate the epistemic status of ML models besides assessing idealizations. The ML idealization failure hypothesis should be read as a necessary test for ML models to pass, not a sufficiency test.

Current approaches to evaluating idealizations in philosophy of science are chiefly concerned with explaining *why* idealizations are successful. As a result, current idealization evaluation falls under two broad methods: The first trades on evaluating whether idealizations achieve the scientific aims of explanation or scientific understanding. The second method evaluates whether particular cases of idealizations instantiate a given theory of idealization. In the latter case, notions of idealization failure are often marginalized to the negation of a positive proposal.

In general, idealizations can be successful empirically if they have predictive power (Mizrahi 2012) or are safe for engineering use (Batterman and Rice 2014; see Lawler 2021). On this score, ML models may do well because of their high predictive power and usefulness. On influential accounts of idealizations, idealizations are successful if they exemplify features of phenomena (Elgin 2017) or only distort non-difference makers (Strevens 2016). Developing certain evaluation tests of ML may help to uncover distortions of difference-making, such as spurious correlation tests or novel tests that probe ML architectures to uncover structural idealizations. Lawler (2021) proposes that idealizations can be legitimate and successful even if they only have the potential for empirical success, as long as there is an appropriate tie to the phenomenon in question. In the context of ML securing the appropriate tie to phenomena will likely require reducing link-uncertainty (Sullivan 2022a, 2022b).

Since philosophers of science discuss successful idealization using examples that are either known successes, or cases of clear problematic distortions, idealization failure either goes unaddressed, with several cases simply labelled as successful in virtue of merely being possible explanations. There is a need for considering different gradients of success regarding how-possibly explanation, which can further assess cases of idealization failure and the epistemic status of idealizations in ML.[7] So while in this paper I cannot provide an account of idealization failure for ML models, I hope that this paper provides motivation for considering the idealization failure hypothesis as a way to solve the problems that emerge from the similarity between ML models used in science and toy models.

5. Conclusion

Are ML models anything more than 'mathematized science fiction'?[8] In this paper I argued that one way of answering this question is to treat ML models as functioning as highly idealized toy models. If we adopt the view that highly idealized toy models can represent phenomena,

[7] Grüne-Yanoff and Verreault-Julien (2021) might be useful place to start.
[8] See Reutlinger et al. (2018, p. 1070) for posing the same question to toy models.

then so do ML models. Of course, there could be hold-outs to the similarity view of representation. For those holdouts, focusing more on the epistemology of idealization can capture the extent to which ML models may enable understanding without subscribing to the view that ML models represent targets. All told, I believe that adopting the view that the function of ML models is the same as highly idealized models can help us to understand not only the epistemic limitations of ML models, but also help to explain why they have been so successful and influential despite these epistemic limitations.

Acknowledgements

I'd like to thank my fellow symposium participates Will Fleisher, MikeTamir, Elay Shech and Suzanne Kawamleh. For many other helpful comments and conversations, I'd like to thank Céline Budding, Thomas Grote, Yeji Streppel, Philippe Verreault-Julien, Carlos Zednik, and special thanks to John Mumm. This work is supported by the Netherlands Organization for Scientific Research (NWO grant number VI.Veni.201F.051) and the research programme Ethics of Socially Disruptive Technologies funded by the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

References

Boge, Florian J. 2022. "Two dimensions of opacity and the deep learning predicament." *Minds and Machines* 32, no. 1: 43-75. https://doi.org/10.1007/s11023-021-09569-4

Batterman, Robert W., and Collin C. Rice. 2014. "Minimal model explanations." *Philosophy of Science* 81, no. 3: 349-376. https://doi.org/10.1086/676677

Carrillo, Natalia, and Tarja Knuuttila. 2022. "Holistic idealization: An artifactual standpoint." *Studies in History and Philosophy of Science Part A* 91: 49-59. https://doi.org/10.1016/j.shpsa.2021.10.009

Creel, Kathleen A. 2020."Transparency in complex computational systems." *Philosophy of Science* 87, no. 4: 568-589. https://doi.org/10.1086/709729

Duede, Eamon. 2023. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science*. Cambridge University Press, 1–13. https://doi.org/10.1017/psa.2023.8

Elgin, Catherine Z. 2017. True enough. MIT press. https://doi.org/10.7551/mitpress/9780262036535.001.0001

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. "Dermatologist-level classification of skin cancer with deep neural networks." *nature* 542, no. 7639: 115-118. https://doi.org/10.1038/nature21056

Fleisher, Will. 2022. "Understanding, idealization, and explainable AI." *Episteme* 19, no. 4: 534-560. https://doi.org/10.1017/epi.2022.39

Frigg, Roman, and James Nguyen. 2018. "The turn of the valve: representing with material models." *European Journal for Philosophy of Science* 8, no. 2: 205-224. https://doi.org/10.1007/s13194-017-0182-4

Giere, Ronald N. 2004. "How models are used to represent reality." *Philosophy of science* 71, no. 5: 742-752. https://doi.org/10.1086/425063

Grüne-Yanoff, Till, and Philippe Verreault-Julien. 2021. "How-possibly explanations in economics: anything goes?." *Journal of Economic Methodology* 28, no. 1: 114-123. https://doi.org/10.1080/1350178x.2020.1868779

Knüsel, Benedikt, and Christoph Baumberger. 2020. "Understanding climate phenomena with data-driven models." *Studies in History and Philosophy of Science Part A* 84: 46-56. https://doi.org/10.1016/j.shpsa.2020.08.003

Lawler, Insa. 2021. "Scientific understanding and felicitous legitimate falsehoods." *Synthese* 198, no. 7: 6859-6887. https://doi.org/10.1007/s11229-019-02495-0

Levy, Arnon. 2021. "Idealization and abstraction: refining the distinction." *Synthese* 198: 5855-5872. https://doi.org/10.1007/s11229-018-1721-z

Mäki, Uskali. "MISSing the world. Models as isolations and credible surrogate systems." *Erkenntnis* 70 (2009): 29-43. https://doi.org/10.1007/s10670-008-9135-9

Meskhidze, Helen. 2021. "Can Machine Learning Provide Understanding? How Cosmologists Use Machine Learning to Understand Observations of the Universe." *Erkenntnis*: 1-15. https://doi.org/10.1007/s10670-021-00434-5

Mizrahi, Moti. 2012. "Idealizations and scientific understanding." *Philosophical Studies* 160: 237-252. https://doi.org/10.1007/s11098-011-9716-3

Nguyen, James. "It's not a game: Accurate representation with toy models." *The British Journal for the Philosophy of Science* (2020). https://doi.org/10.1093/bjps/axz010

Potochnik, Angela. 2017. *Idealization and the Aims of Science*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226507194.001.0001

Räz, Tim, and Claus Beisbart. 2022. "The importance of understanding deep learning." *Erkenntnis*: 1-18. https://doi.org/10.1007/s10670-022-00605-y

Reutlinger, Alexander, Dominik Hangleiter, and Stephan Hartmann. 2018. "Understanding (with) toy models." *The British Journal for the Philosophy of Science*, 69:4, 1069-1099. https://doi.org/10.1093/bjps/axx005

Rice, Collin. 2019. "Models don't decompose that way: A holistic view of idealized models." *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axx045

Schelling, Thomas. 1971. "Dynamic Models of Segregation," *The Journal of Mathematical Sociology*, 1(2), pp. 143–86. https://doi.org/10.1080/0022250x.1971.9989794

Strevens, Michael. 2016. "How idealizations provide understanding." In Explaining understanding, pp. 53-65. Routledge, 2016.

Suárez, Mauricio. 2015. "Deflationary representation, inference, and practice." *Studies in History and Philosophy of Science Part A* 49: 36-47. https://doi.org/10.1016/j.shpsa.2014.11.001

Sullivan, Emily. 2022a. "Understanding from machine learning models." *The British Journal for the Philosophy of Science*, 73(1):109–133. https://doi.org/10.1093/bjps/axz035

Sullivan, Emily. 2022b. Inductive Risk, Understanding, and Opaque Machine Learning Models. *Philosophy of Science*, *89*(5), 1065-1074. https://doi.org/10.1017/psa.2022.62

Sullivan, Emily, and Kareem Khalifa. 2019. "Idealizations and understanding: Much ado about nothing?." *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2018.1564337

Tamir, Michael and Elay Shech. 2022. "Understanding from Deep Learning Models in Context." In Scientific Understanding and Representation Eds. Lawler, Shech, Khalifa Routledge.

Verreault-Julien, Philippe. 2019. "How could models possibly provide how-possibly explanations?." *Studies in History and Philosophy of Science Part A* 73: 22-33. https://doi.org/10.1016/j.shpsa.2018.06.008

Weisberg, Michael. 2012. *Simulation and similarity: Using models to understand the world*. Oxford University Press.

https://doi.org/10.1093/acprof:oso/9780199933662.001.0001

Zednik, Carlos, and Hannes Boelsen. 2022. "Scientific exploration and explainable artificial intelligence." Minds and Machines 32, no. 1: 219-239. https://doi.org/10.1007/s11023-021-09583-6