From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence

Catherine Stinson*†

There is a vast literature within philosophy of mind that focuses on artificial intelligence but hardly mentions methodological questions. There is also a growing body of work in philosophy of science about modeling methodology that hardly mentions examples from cognitive science. Here these discussions are connected. Insights developed in the philosophy of science literature about the importance of idealization provide a way of understanding the neural implausibility of connectionist networks. Insights from neurocognitive science illuminate how relevant similarities between models and targets are picked out, how modeling inferences are justified, and the metaphysical status of models.

1. Introduction. For about 20 years, philosophy of artificial intelligence (AI) has been confined to a rather dusty corner of philosophy of mind.[1] Despite the very widespread use of methods like support vector machines, decision trees, principal components analysis, and neural networks in many branches of science and engineering, contemporary AI has largely escaped the notice of philosophers of science.[2] During the same time period, models and simulations have grown from a niche subject to a popular subdiscipline within philosophy of science, although these discussions tend to focus on the

Received August 2018; revised September 2019.

*To contact the author, please write to: Department of Philosophy, John Watson Hall, Queen's University, Kingston, ON K7L 3N6, Canada; e-mail: c.stinson@queensu.ca.

†On its very long journey to publication, this article benefited from discussions with many colleagues, including Ken Schaffner, Peter Machamer, Jim Bogen, Floh Thiels, Dave Touretzky, Jackie Sullivan, Liz Irvine, Hong Yu Wong, Eva Engels, Gregor Hochstetter, Boris Hennig, Tim Bayne, Cameron Buckner, and Mikio Akagi, as well as audiences in Barcelona, Cambridge, Berlin, London (Ontario), and Toronto.

[1] For signs of renewed activity, see Buckner (2018).

[2] Godfrey-Smith (2006, 2009) mentions neural networks very briefly.

Philosophy of Science, 87 (October 2020) pp. 590–611. 0031-8248/2020/8704-0003$10.00

https://doi.org/10.1086/709730 Published online by Cambridge University Press

use of models in a handful of fields (economics, climate science, physics, ecology) that does not include the neurocognitive sciences.

It is past time for these estranged relatives to be reunited. With the infiltration of machine learning into so many of the technologies that mediate our lives, understanding AI methods could hardly be more topical. Likewise, understanding how and when to trust the predictions of climate models is urgently important. This article paves the way for the reembrace of philosophy of AI into philosophy of science and illustrates some potential benefits on both sides.

A natural first step in bringing AI back into philosophy of science is to revisit a problem that holds a central place in discussions of methodology in AI: Why is neural plausibility considered helpful in connectionist models, when the models are known not to be realistic? This problem was never resolved but continues to be relevant, most recently in discussions of deep learning, where "adversarial examples" are revealing differences between computer and human vision (see Han et al. 2019). During connectionism's heyday, the philosophical vocabulary to answer this question was lacking. Here we consider the problem of why and how cognitive models should be neurally plausible, through the lens of a more general issue in scientific modeling: In what ways do models need to resemble their target systems in order to produce relevant, generalizable results? This proves useful in understanding connectionist models.

Insights from the neurocognitive sciences likewise reveal important gaps in accounts of modeling and simulation, which rely on examples from a restricted set of scientific fields. Neurocognitive science brings a more critical view of representation and digs deeper into questions about causation and the metaphysics of models. An analysis of inference in connectionist modeling based on kinds can be extended to models more generally.

Section 2 exhumes the problem of connectionism's simultaneous endorsement of and failure to follow through on neural plausibility. Section 3 surveys recent philosophical work on models and simulations to show how connectionism's methodological puzzle can be resolved by understanding connectionist models as idealized models of cognitive mechanisms. Section 4 explores gaps in standard philosophical accounts of scientific modeling from the perspective of the neurocognitive sciences and outlines a novel account of the relationship between models and targets inspired by an analysis of connectionist models. Section 5 illustrates how this account applies to a series of examples of connectionist models.

2. The Neural Implausibility of Connectionist Models.

Although it has a longer history, philosophical interest in connectionist modeling stems largely from the Parallel Distributed Processing (PDP) Research Group, whose two-volume 'bible' (McClelland and Rumelhart 1986; Rumelhart and McClelland

https://doi.org/10.1086/709730 Published online by Cambridge University Press

1986b) sparked debate about computational methods in cognitive science.[3] The standard connectionist network architecture is a three-layer, feed-forward network of simple neuron-like units, where each unit sends output to every unit in the next higher layer. Any pattern of connections is possible though, including sparse, lateral, feedback, or recurrent connections. Contemporary deep learning networks include more than three layers and are often connected in small neighborhoods. The activity of the network is defined by each unit's activation, each connection's weight, and the activation function used to calculate a unit's output based on the weighted sum of its input activations. The weights are adjusted using a learning rule designed to minimize overall error.

At first glance, the connectionist project seems to be about building neurally plausible AI models. The introduction to the PDP bible states, "One reason for the appeal of PDP models is their obvious 'physiological' flavor: They seem so much more closely tied to the physiology of the brain than are other kinds of information-processing model" (McClelland and Rumelhart 1986, 10). But on closer inspection, both the statement and the motivations for the project prove harder to interpret. What is meant by "flavor"? Why is "physiological" in scare quotes? In virtue of what is having a "physiological" flavor appealing?

The PDP group's stated inspiration was that classical AI's models seemed unsuited for some kinds of computations: "the biological hardware is just too sluggish for sequential models of the microstructure to provide a plausible account. . . . Each additional constraint requires more time in a sequential machine, and, if the constraints are imprecise, the constraints can lead to a computational explosion. Yet people get faster, not slower, when they are able to exploit additional constraints" (McClelland and Rumelhart 1986, 12). It is also worth noting that the PDP group's project was very much continuous with classical AI in their concern for building models that produce output that matches the results of psychological experiments and their attention to reaction times: these are moves taken straight out of the cognitive psychologist's toolbox.[4] But because it was taken as a turn away from traditional approaches to cognitive science, the PDP bible's appeal to biological hardware invited objections from the AI and cognitive psychology mainstream. These objections are organized below into four problems.

2.1. The Levels Problem. The first major critique concerns what level PDP models are intended to occupy. Broadbent argues that McClelland and Rumelhart (1985) inappropriately cast their distributed memory system as having "implications at the psychological and not merely at the physiological

[3] A renewed interest in some of these questions is currently being hashed out in response to Marcus (2018).

[4] Hinton, Rumelhart, and McClelland all started out as psychologists.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

level" (Broadbent 1985, 189). Broadbent's appeal to levels refers to Marr (1982), with the implication that cognition ought to be independent of implementation.

Fodor and Pylyshyn (1988) pose Broadbent's challenge as a dilemma: either connectionist models are "mere implementations" of symbolic models, or they fail to adequately capture cognition. If PDP models are psychological models, then neural details should be irrelevant and afford no advantage. If PDP models are implementation level models, they might be interesting to neuroscientists but are not cognitive science.

It would require many pages to list all the variations of this reaction. Suffice it to say that the *Stanford Encyclopedia of Philosophy* has set it down as received opinion that there are two kinds of connectionist: implementational and radical. Implementational connectionists "hold that the brain's net implements a symbolic processor," while radical connectionists "claim that symbolic processing was a bad guess about how the mind works" (Garson 2015). Some connectionist projects, such as the articles in Hinton (1990), show that PDP models are capable of structured representations and serial processing (i.e., implementational connectionism). Other connectionist projects, such as Plaut (1995), show that what looks like serial processing on the surface might be better explained in terms of network-level details (i.e., radical connectionism). A not-so-radical connectionism claiming that symbolic processing is a bad guess at how some mental functions work is closer to what most connectionists believe.

Either way, connectionists do not generally accept that their models are mere implementations. Rumelhart and McClelland (1985) object that much of what concerns cognitive psychologists is at the algorithmic rather than the computational level.[5] Smolensky (1988a, 1988b) describes connectionist models as being at the "sub-symbolic level" and says that the goal of connectionist research is a "middle ground between implementing symbolic computation and ignoring structure" (Smolensky 1988a, 152). What this middle ground is exactly is unclear.

2.2. The Neural Detail Problem. Another well-rehearsed challenge is that connectionist models are unlike brains in their details. The backpropagation algorithm is infamous for being neurally implausible; error signals cannot in general be propagated backward through a network of neural connections, as the algorithm requires. Likewise, nodes in connectionist models typically have deterministic activation functions, whereas real action potentials are stochastic.

A key example of the neural detail problem is the flexibility in how to interpret single units. In networks with local representations, units are assigned

[5] See Churchland and Sejnowski (1990) on how connectionist models relate to Marr's levels.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

specific meanings, such as the names, occupations, and ages of members of the Jets and Sharks in McClelland (1981). In networks with distributed representations, "each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities" (Hinton 1984, 1). Far too few units are used in most connectionist models to be realistic brain models. In some connectionist networks, units explicitly stand in for whole populations of neurons, with the activation of the unit representing a population vector.[6] There is thus considerable diversity in what a unit is meant to correspond to.

One of the most debated examples is the past-tense learner (Rumelhart and McClelland 1986a). This network takes English verbs as inputs and learns to output their past tenses. It is trained using a series of examples, including both regular verbs (add "ed") and irregular verbs (went, swam). The past-tense learner's success in learning to conjugate past tenses without any explicit set of rules separating regular and irregular verbs was, as Boden puts it, "theoretical dynamite" (2006, 956). However, the past-tense learner was also vigorously criticized for its failure to simulate physiological detail. The encoding of its input and output verbs, as phonetic triples called "Wickelfeatures" is perhaps the least plausible detail.

Critics of connectionism treat these disanalogies as mistakes, but the PDP group was well aware that the "physiological" flavor stopped short of realistic detail. Volume 2, chapter 20 of the PDP bible describes the ways in which artificial neural networks are not like real brains. The introduction also hedges on whether physiological plausibility is the goal: "Though the appeal of PDP models is definitely enhanced by their physiological plausibility and neural inspiration, these are not the primary bases for their appeal to us. . . . PDP models appeal to us for psychological and computational reasons" (McClelland and Rumelhart 1986, 11). Lack of realistic neural detail was, apparently, a design feature.

Part of what is going on is that practical concerns require that models not be too complex. That putting too much detail into a model is a mistake is a common refrain among connectionists: "It's not necessary to put in the kitchen sink to get insight. . . . Just to simulate the hell out of populations of everything in the model is mindless" (J. D. Cowan, quoted in Anderson and Rosenfeld 2000, 123). McClelland (2009) argues that while there is a cost to making simplifications in modeling, it is necessary to simplify to achieve understanding. But the implausibility seems to run deeper than just pragmatism.

2.3. The Abstraction Problem. Another puzzle is that connectionists sometimes describe their models in mathematical terms. Smolensky (1991)

[6] Wilson and Cowan (1972) derived equations for the average spike rate of populations of neurons that allows populations of neurons with random, dense connections to be treated as aggregates, and these equations closely match those used in connectionist models.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

claims that connectionism explores what continuous (rather than discrete) mathematics can reveal about the nature of cognition. Thomas and McClelland call connectionist models "a subclass of statistical models involved in universal function approximation" (2008, 23).

An example of this is Touretzky and Hinton (1988), who show how distributed representations can be used to "construct a working memory that requires far fewer units than the number of different facts that can potentially be stored" (423). Here no effort is made to re-create neural details beyond general structural features. The point is to demonstrate a property such networks have no matter what the units represent, yet at the same time, the model is clearly meant as an investigation of working memory. One might wonder how it can do both.

2.4. The Explanation Problem. The final challenge concerns the status of PDP models as explanations. Green worries that "if connectionist models are NOT to be considered THEORIES of cognition, in the traditional scientific sense of the word, then the question arises as to what exactly they are, and why we should pay attention to them" (1998). According to Green, the only interpretation of connectionist networks as theories is one in which they are "literal models of the brain activity that underpins cognition." But this is undermined by the implausibility of connectionist models.

In classical AI, a computer program that produces output comparable to human performance on a cognitive task is considered a theory of that cognitive task. In calling their programs theories, Newell and Simon (1961, 1976) have in mind the deductive-nomological (DN) account (Hempel 1958): "A computer program used as a theory has the same epistemological status as a set of differential equations or difference equations used as a theory" (Newell and Simon 1961, 2013). The logical calculus in the program has the same status as the law and observation statements that constitute a theory in the physical sciences.

Connectionist models are not theories in the DN sense; they do not logically deduce behavior or encode law-like regularities. By the late 1980s the DN account was no longer the received view of scientific explanation, but the lack of consensus on what should take its place left it open what sort of explanations connectionist models provide.

3. Connectionist Models as Idealized Models of Cognitive Mechanisms. Recent developments in philosophy of science shed light on the problems above.

3.1. Mechanistic Explanation. The mechanistic view of explanation has largely supplanted the DN account in the biological sciences. The levels connectionists are concerned with can be thought of as mechanistic levels

https://doi.org/10.1086/709730 Published online by Cambridge University Press

(Craver 2007). Mechanistic explanation situates a phenomenon within a multilevel system of mechanisms, where each level constrains and is constrained by its neighboring levels. A mechanistic explanation involves showing how component entities and their activities are organized to bring about a phenomenon and identifying the mechanism's role in higher-level phenomena.

The suggestion that connectionist models can be understood in terms of mechanistic explanation is raised in Miłkowski (2013) and expanded on in Stinson (2018). This insight is in tune with the PDP group's stated motivations. Rather than seeing physiology and cognition as independent, connectionists explore the ways in which the physiological microstructure constrains cognition. The PDP bible lists the constraints they take from neuroscience, including "there is a very large number of neurons. . . . Neurons receive inputs from a large number of other neurons. . . . Learning involves modifying connections. . . . Neurons communicate by sending activation or inhibition through connections" (Rumelhart and McClelland 1986c, 130–32).

That mechanistic explanations have no privileged level helps explain why units can correspond to single neurons, populations of neurons, or higher-level entities like phonetic representations. Connectionist models can investigate any number of locations in a system of mechanisms. As Churchland and Sejnowski put it, "Network models . . . depend in important ways on constraints from all levels of analysis. . . . Since the networks are meant to reflect principles at entirely different levels of organization, their implementations will also be at different scales in the nervous system" (1990, 368–69).

3.2. Abstraction. Simplicity is essential not only for getting models to work but also for explanation. The cost of simplification is that when you draw an inference from a simplified model, it may be that the interesting properties of the model result from aspects of the model that differ from the target rather than from what the model and target have in common. Connectionist models are different from brains in many ways, so one might expect them to behave differently. This is an instance of a very general worry about scientific models, namely, which details need to be captured accurately for a model to inform us about the target system and which can be safely altered. This problem has been the subject of much work in philosophy of science.

One kind of simplification is what Cartwright (1989) calls abstraction. Abstract models remove details so that the effects of a small number of variables can more easily be investigated on their own. Abstracting away too many details can lead to error when there are complex relationships between variables, such that investigating each in isolation is not straightforwardly informative about the combined picture. Nevertheless, quite often it is perfectly legitimate. A comparison can be made to how experiments need to control variables in order to be interpretable. There is a trade-off to be made between naturalistic field experiments with many uncontrolled variables and

https://doi.org/10.1086/709730 Published online by Cambridge University Press

lab experiments that are more readily interpretable but have less external validity.

The trick is to figure out which details matter. Morgan (2002, 2003) argues for the importance of materiality: sharing the same materials lends experiments closer access to their targets than models, which makes experimental systems more likely to share the properties that are relevant. Parker objects to assessments of computer simulations as "just mathematical modeling exercises" (Parker 2009, 491), arguing that simulations are physical models. She points out that, in weather forecasting, modelers are better able to set up the relevant initial conditions in a computer simulation than in a laboratory model that uses the same materials as real weather systems. Simulations are thus better able to predict the weather than "same stuff" laboratory models, and the key is "whether the experimental and target systems were actually similar in the ways that are relevant, given the particular question to be answered about the target system" (493). In another meteorological model, sizable volumes of atmosphere are treated as homogeneous points in a grid, while measurements of the complex dynamics at a scale lower than the grid resolution are approximated with a single parameter value. Ignoring the known details at a finer grain of resolution leads to more accurate weather predictions than if those details were included in the model (Norton and Suppes 2001, 95–96; see also Küppers and Lenhard 2004).

Likewise in cognitive modeling, when the goal is to predict the behavior of a cognitive agent, pragmatic concerns like maximizing accuracy take precedence over modeling the finer details of the system. This can be seen in the popularity of support vector machines, which make little to no effort to mimic human visual processing, in the ImageNET image recognition challenge (Russakovsky et al. 2015).

Giere (2004) and Godfrey-Smith (2006) focus on the representational role of models and likewise argue that models and targets being similar in the relevant respects is what justifies inferences from the one to the other. Which similarities are relevant depends on the context: "scientists use continuous fluid models to represent water for the purpose of studying fluid flow and also use molecular models for the purpose of representing water for the study of Brownian motion" (Giere 2004, 750).

If the point of the past-tense learner had been to model how verbs are represented in the brain, or to simulate conjugation in detail, then the manner of encoding input and output verbs would have been relevant, and using Wickelfeatures would have been inappropriate. But Rumelhart and McClelland wanted to see whether what appeared to be a structured rule-following behavior could be achieved without building that structure in. The way the verbs are represented was bracketed off as irrelevant, given that goal.

Winsberg highlights the importance of arguments, to demonstrate that the results scientists get "from manipulating their respective pieces of equipment

https://doi.org/10.1086/709730 Published online by Cambridge University Press

are appropriately probative concerning the class of systems that interest them" (2009, 577). Those arguments are based not just on similarity but also on having knowledge about how to build good models, which comes from past successes using the same bag of modeling tricks.

Batterman (2001, 2002) describes how the use of mathematical tools like renormalization groups depends on paring down particular problems to minimal models. What Batterman calls "asymptotic" methods are able to explain the universal, stable phenomenologies that are shared by, for example, microstructurally diverse fluids near the critical point in phase transitions, as well as magnets transitioning between ferro- and paramagnetic states (2001, 38). These methods not only offer explanations of these universal phenomena, but by "telling us what (and why) various details are irrelevant for the behavior of interest, this same analysis also identifies those physical properties that are relevant for the universal behavior being investigated" (42).

Something like asymptotic explanation appears in Fuhs and Touretzky's (2006) model of path integration in spatial memory. Their model seeks to explain how rats navigating mazes are able to find efficient paths to goal locations regardless of the paths they have previously traveled, as well as the peculiar hexagonal patterns found in grid cells' firing fields. As a possible explanation of the hexagonal patterns, Fuhs and Touretzky showed that "hexagonally spaced activity bumps can arise spontaneously on a sheet of neurons in a spin glass-type neural network model" (4266). In spin glass models, each unit is connected to its closest neighbors in a multidimensional grid. This network structure is loosely based on the local structure in entorhinal cortex, where grid cells are found, on the assumption that dendrites are closely packed. When circles or cylinders of uniform size are closely packed together, the highest density arrangement is a hexagonal pattern. This is true regardless of whether they are telecommunication cables or dendrites running through nerve tracts. Fuhs and Touretzky (2006) justify arranging the units in their connectionist model of grid cells in a hexagonal pattern based on this fact about close packing, even though real dendrites are neither perfectly cylindrical nor uniform in size. Their explanation depends on *not* using a more accurate, detailed model, because without the assumption that the dendrites are uniform cylinders, the geometric fact about close packing could not have been applied.

3.3. Idealization. On Cartwright's (1989) definition, idealization adds or changes details, such that the idealized model has properties not present in the target system. In lab experiments, idealizations might substitute more convenient materials or assign implausible values to variables for ease of calculation. Backpropagation and deterministic activation functions are the clearest cases of idealization in connectionist modeling. At first blush, it

https://doi.org/10.1086/709730 Published online by Cambridge University Press

seems like putting the wrong details in, as opposed to merely removing irrelevant details, should make for a worse model, but this is not generally the case.

A number of authors have compared idealizations to fictions and suggested that models are interpreted in much the same way as we interpret literature. Mäki's (2012) analysis goes in a different direction. Mäki argues that apparently false idealizations can be interpreted as true in several distinct ways. Leaving out some factors can be intended as a "negligibility assumption"; that is, those factors have a negligible effect, given the intended purposes and audience of the model (222). "Applicability assumptions" restrict the intended use of a model to domains where the factors left out have negligible effects (225). Other kinds of assumptions might defend the use of an idealized model on the grounds that the idealization makes the model more tractable or more suitable for pedagogical purposes (228–30). These assumptions are not always spelled out explicitly.

In some cases, the use of backpropagation in connectionist models could be justified with a tractability assumption, since it was for a time the only known method of updating weights that was guaranteed to converge. In other cases backpropagation can be justified with a negligibility assumption. For example, in NETtalk (Sejnowski and Rosenberg 1986) backpropagation is unproblematic given the purpose of the model, because their goal is to show that a system capable of pronouncing English words need not encode a complicated set of rules. For that purpose, it is fair to simply assume that the brain has some way of propagating error signals, without worrying about how exactly that happens. The particular pathways the error signals take do not make a difference to what they are investigating. In contrast, Suri and Schultz (2001) develop a model of learning mechanisms, so the way error signals are propagated is highly relevant. In their model, backpropagation is not used; instead the anatomy of the basal ganglia is reproduced in some detail, including only pathways that exist in the brain and through which feedback is known to actually travel.

The curious phrase "'physiological' flavor" might be interpreted to mean that connectionist models are idealized models of cognitive mechanisms. If you take away irrelevant details, and idealize others, cortex is an interconnected network of simple learning units.

3.4. Discovery. Models serve many different purposes in science, and many different strategies may be employed in the search for mechanisms. Anderson and Rosenfeld's (2000) history of connectionism demonstrates that among connectionist modelers there have always been widely differing approaches in terms of how much physiological detail to include and what the goals are. These goals include engineering, mathematical, psychological,

https://doi.org/10.1086/709730 Published online by Cambridge University Press

and neuroscientific questions. Models intended for different epistemic roles require different characteristics.

Steinle (1997, 2002) argues that experiments at different stages in a research project tend to have different epistemic goals, which means that different sorts of experiments are performed. For example, earlier exploratory experiments tend to try out many more combinations of parameter values in a search for potentially meaningful correlations, while later "theory-driven" experiments use high-precision equipment and "are typically done with quite specific expectations of the various possible outcomes" (Steinle 1997, S70).

Steinle's analysis also holds for models. Models used at different stages in a research project tend to have different epistemic goals and correspondingly may vary in terms of how idealized or specific they should be in order to meet those goals. This difference in epistemic goals is reflected in the difference in detail in the learning mechanisms used in NETtalk compared to models of the basal ganglia.

4. Epistemology and Metaphysics of Models

4.1. Overview. This section continues in the spirit of Irvine (2014), where considering computational modeling practices from cognitive neuroscience problematized and revised claims from the models and simulations literature. Other examples of work on computational modeling in the neurocognitive sciences are Kaplan (2011), who argues that computational explanations in neuroscience are mechanistic explanations; Chirimuuta (2018, 851), who argues that there are "numerous instances of distinctively mathematical, non-causal explanation" in computational neuroscience; and Stinson (2018), where I argue that connectionist cognitive models explain using a logic of tendencies in contrast to classical AI's use of inference to the best explanation.

What these examples share is a concern with questions about causation and ontology, quite unlike the focus on representations characteristic of the models and simulations literature (see Suárez 2003; Giere 2004; Weisberg 2012; Frigg and Nguyen 2016). Perhaps the reason is that cognitive scientists have learned to regard with some suspicion appeals to representations as explanations and tend to worry about how putative representations acquire and transmit their contents. This more critical take on representation and causal-mechanistic bent could be helpfully applied to the models and simulations literature.

Three open questions in that literature are how relevant similarity ought to be judged, how inferences from models to targets are justified, and what the metaphysical status of models is. On the first two questions, the state of the art seems to be that which similarities are relevant has to be decided on a case-by-case basis (Parker 2009) and that we have to provide arguments to

https://doi.org/10.1086/709730 Published online by Cambridge University Press

justify our modeling choices (Winsberg 2009). A deeper analysis of the criteria modelers use in making these judgments is needed.

Godfrey-Smith (2009) lays out the problem of the metaphysical status of models. He notes that model systems are “in a sense, of the same kind as the target systems that the models are used to help us understand” (104) but rejects what scientists say about planets, populations, or economies being “inside the computer” (106) even in a loose sense. He resists attributing computational models object-hood such that they are taken as “shadowy additional graspable thing[s]” (108) and endorses the comment attributed to Deena Weisberg that the Platonism of mathematicians is a “folk ontology” (107). Godfrey-Smith describes the Platonist view as taking the model to be an abstract entity that can be investigated mathematically, then requiring a mapping of abstract properties to the physical properties of the target.

The Platonist view is implicitly rejected because it runs into the Third Man problem, by assuming the independent reality of abstract objects. But, if models are not objects that can be directly compared to targets, the question remains how they can inform us about concrete things. Representations and fictions are overly flexible; anything can happen in fiction, so it does not adequately constrain inferences to real world targets. As Frigg and Nguyen argue, “One can imagine almost anything about almost any object, but unless there are criteria telling us which of these imaginings should be regarded as true of the target, these imaginings don’t licence any surrogative reasoning” (2016, 233). The metaphysical problem and the inference problem are thus in tension with one another.

Winsberg likewise notes that “practitioners of simulation” favor the idea that simulations literally mimic their target systems, such that a simulation of fluid dynamics can be viewed as an experiment in a “virtual wind tunnel” (2010, 35). But Winsberg raises the problem of “whether or not, to what extent, and under what conditions a simulation *reliably* mimics the physical system of interest” (37).

A common assumption in representational accounts of models is that the relation between model and world is one of similarity, following Giere’s (1988) diagram, reproduced here in figure 1. Winsberg notes that the relation between models and targets has to be something “far more complicated than mimicry” (2010, 39). What this more complicated relation might be is key to all three questions at hand.

Frigg and Nguyen offer a more complicated candidate relation in their DEKI account. According to DEKI, models are interpreted as exemplifying a set of properties of interest, which are mapped to properties imputed to the target. The addition of more stations along the way between model and target solves a number of problems with direct representation accounts (see Frigg and Nguyen 2016) but retains the main weakness of representational accounts,

https://doi.org/10.1086/709730 Published online by Cambridge University Press

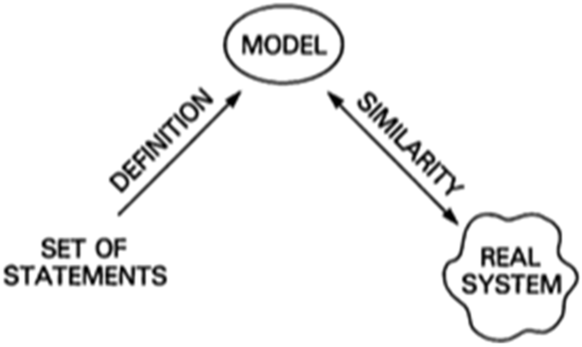

Figure 1. Relations between theory, model, and reality, reproduced from Giere (1988, 83).

since a mapping between sets of properties does little to ensure that one reliably mimics the other.

4.2. An Alternative to Representational Accounts of Models. My analysis of inference in connectionist modeling from Stinson (2018) can be extended into an alternative account of models. This account provides a more robust connection between properties of models and targets and legitimizes scientists' views about the ontology of models. I argue that inferences are drawn from connectionist models to their targets indirectly via kinds that both the model and target exemplify. A diagram illustrating this set of relations is given in figure 2.

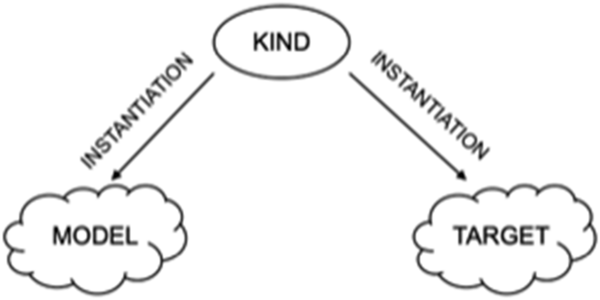

Figure 2. Relations between model, target, and kind.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

On this account, the inferences modelers draw from models to targets have the following structure:

P1 The target system T is an instance of kind K.
P2 Model M is an instance of K.
C Therefore, T should be similar to M.

In the case of connectionist models, T is the brain or cognitive system being investigated, and M is the connectionist model.

This extra step of associating the model first with a kind then that kind with the target is promising for several reasons. One is that it makes better sense of idealization than representational accounts of modeling do. A good model is often very minimal, including only the properties of interest and few other details. If similarity were the criterion for model target relations, more details would be better, not fewer. But if capturing the characteristic properties of a kind is the goal, then idealized models are exactly what one ought to aim for. Similarity is what we want to infer as a conclusion, not what we aim for in model building.

Second is that it can provide needed guidance about which similarities are the relevant ones to capture in a model. The properties that a model should mimic from the target are the ones that are characteristic of the kind they both belong to. As long as we have a way of picking out kinds, this gives specific guidance about what the relevant similarities are: something representational accounts were unable to do.

The account also has other nice properties. It fits well with Godfrey-Smith's (2009) observation that models sometimes describe one case of a target phenomenon, then act as a hub, anchoring all the "actual-world" cases (107). The hub is the kind captured by the model. Models with no target are likewise accounted for by treating the kind as a generalized target.

What is still wanting in this account is a way of delineating kinds, telling which kinds a target belongs to, and a workaround to the problem of universals. One missing piece can be provided by Khalidi's (1998, 2013) broad view of kinds. Khalidi argues that scientific kinds like parasite, liquid, or schizophrenia should be considered "real kinds" because "we discovered things about them which were by no means implied when they were first introduced" (Khalidi 1998, 42).[7] This account of kinds assumes neither essential natures nor strictly hierarchical relationships between kinds. As such, it is promiscuous enough to accommodate most any phenomena that one might want to model. But because members of kinds have nonarbitrary things in common,

[7] Khalidi (2013) uses the term "natural kinds" but in conversation says that he wished he had called them "real kinds."

https://doi.org/10.1086/709730 Published online by Cambridge University Press

kinds provide a basis for inferring that members likely have the properties characteristic of the kind in common with each other.

Khalidi's kinds might be made more robust by connecting them to Andersen's (2017) information-theoretic update to Dennett's (1991) "real patterns." A real pattern is one that "can be reliably picked out and tracked through time . . . and which allows one to make predictions that are better than chance" (Andersen 2017, 603). A collection of phenomena that manifest a real pattern according to Andersen would count as the members of a real kind according to Khalidi.

Andersen comments on the "profligacy" of patterns, saying that "there could be a vast number of different ways of picking out such patterns that give us predictive grasp on the system" (2017, 603), but just as the promiscuity of Khalidi's kinds should not be troubling given that they do not assume anything about essences or hierarchies of kinds, the profligacy of Andersen's patterns should not be troubling because "the degree of realism is very, very minimal" (603). Both Khalidi and Andersen argue that the concern that this allows for too many kinds or patterns is overblown. The criteria that kinds or patterns can be reliably picked out, tracked, and make useful predictions are not met by jerry-rigged kinds.

Against the worry that patterns are epiphenomenal, Andersen claims "the overwhelming majority of patterns are counterfactually robust, in that they could have differed in their microphysical details in each token instantiation without thereby altering the relatum's causal profile" (2017, 594). That idealized and simplified models of causal processes are often most useful for figuring out how those processes work would be mysterious, were it not the case that these patterns are real in some sense that goes beyond the reality of their microphysical details.[8]

Both Khalidi and Andersen claim at least a minimal reality for their kinds/patterns. On Andersen's account, patterns are part of the causal nexus, and "higher-level causes are just as real as lower-level causes" (2017, 619). Her deke around the problem of universals is that what is real is "the causal nexus and patterns instantiated in it, which are informationally structured, but where the information itself is a structure *of* something else, not a reified extra substance" (619).

One can go a step further in endorsing the folk ontology of scientists and interpret these reality claims as implicating additional shadowy things of a sort. But before the knee-jerk reaction that this runs into the problem of universals kicks in, let us look at some recent developments in metaphysics, where respectable options are available for considering universals as concrete in some sense.

[8] Both Khalidi and Andersen reject the suggestion that Kim's causal exclusion argument might cause problems for the reality of promiscuous/profligate kinds or patterns.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

Hennig (2015) offers a possible solution to the problem of universals, based on Baxter (2001), wherein kinds have concrete *aspects*, making them in a sense "the same as" instances of the kind. Hennig summarizes the account in the following way: "that Socrates instantiates the kind *seated thing* means that there is an aspect of Socrates that is also an aspect of the kind *seated thing*. This aspect can be described in two ways: (1) as Socrates *qua seated thing* or (2) as *seated thing qua* instantiated by Socrates" (2015, 5). Hennig clarifies as follows: "That Socrates instantiates the kind seated thing means that there is a seated thing that is the same as Socrates. This thing is one of his aspects. Socrates is an instance of seated thing, and seated thing is an aspect of Socrates. The aspect is not a third entity mediating between Socrates and the universal 'being seated'; there are only two things: Socrates and the aspect" (Hennig, pers. comm., April 12, 2019). Aspects, unlike Platonic universals, are concrete and in the world. The aspect 'Socrates qua human' has flesh and bones.

This brief detour into contemporary metaphysics shows that there are respectable options available that allow us to take seriously the views of scientists about the reality of models. Putting these pieces together, the relation between kind and target can be understood as the relation between aspect and instance, and the relation of model to target as between two instances of the same aspect. It is tempting to think of models as though there are two of them, the ideal one and the instantiated one. The ideal one is what scientists think is really in the target system and really in the computer. The instantiated model is the tool we use to get at the ideal. The former is the aspect. The latter is an instance.

The practice of modeling picks out an aspect of the target to investigate and then constructs an instance of that aspect that can be manipulated conveniently. A mathematical model like the Hodkin-Huxley equation comes pretty close to being the aspect under investigation, while models that substitute another material for the material in the target (like architectural models, model organisms, and analog models) are instances that have additional aspects not characteristic of the kind in question.

I am seated right now as I write, making me an instance of 'seated thing'. In virtue of being seated, I could act as a model of 'Socrates qua seated thing'. Using this model, I could speculate that Socrates's toes might also have been prone to falling asleep after sitting too long on a hard chair. But I also have other aspects that are not shared with Socrates. It would not be wise to conclude that Socrates too would usually be drinking tea and overhearing conversations in English while sitting on a hard chair. The instances of 'seated thing' near me do share some of those aspects, but they share them in virtue of being instances of 'person working at Propeller Coffee'. A better model of 'seated thing' would be isolated from noises and hot beverages.

The status of computational models is a bit subtler. In some ways they are like mathematical models in that they are close to being pure aspects. But as Parker argues, computational models are also physical models with

https://doi.org/10.1086/709730 Published online by Cambridge University Press

properties of their own (like being made with transistors). Within a range of conditions those other aspects can be made irrelevant, but in the presence of large magnets or when submerged in water, computational models will show their colors as electronic devices. A computational model qua made with transistors may not be informative about cognition, but a computational model qua connectionist network ought to be.

What makes computers so useful in modeling is that they are designed to have the capacity to explore the aspect of your choice, while isolating that aspect from their other aspects (made with transistors, notebook sized, manufactured in China, etc.). If you consider only particular output streams, like images, printouts, or certain files (as opposed to measuring the CPU's temperature or seeing what happens when you whack it with a hammer), and assume a translation code that interprets that output, a programmer can make a computer be an instance of a wide variety of aspects. Other kinds of models, like fruit flies in genetics, are likewise chosen because they make the investigation of a given aspect more feasible (faster, cheaper, more ethical) than it would be to investigate that aspect in the target itself. For a model to be minimally appropriate it has to be an instance of the aspect of interest. That the model and target are the same in the sense of sharing an aspect is what sanctions inferences from one to the other.

The starting point in building a model is identifying a kind K that the system of interest belongs to and that the model will be designed to investigate. One factor that affects the strength of the inference is whether K is a real, robust kind capable of sustaining generalizations. For the most generic Ks in connectionist modeling, the choice of model amounts to the wager that some of the generalizations that are relevant to cognition operate at the network level. Another factor is whether M is a representative instance of K. Models with a minimum of properties that are not typical of Ks are more representative. Finally, the inference depends on T also being of kind K. If K is a real kind, and M is a minimal instance of K, whatever one finds out about the kind K by investigating M should, all else being equal, also be true of T, assuming T belongs to K. It may still happen that T is atypical in relevant respects and so fails to have the same properties as M despite belonging to the same kind K.

Considering models in the neurocognitive sciences has motivated the need for a more complex account of the model-target relation, better answers to questions about relevant similarity, and more detail about the metaphysics of models. The bare outlines of a novel epistemology and metaphysics of models has been drawn here from an analysis of inference in connectionist modeling.

5. Inferences Via Kinds in Computational Cognitive Science. Let us see how this account works in practice, by applying it to some examples. In Marr's (1969) theory of the cerebellum, the starting points are some basic anatomical knowledge about the types of cells found in the cerebellum and

https://doi.org/10.1086/709730 Published online by Cambridge University Press

the patterns and numbers of connections between them, the hypothesis that the function of the cerebellum is to learn motor skills, and ideas about feature analysis then current in the AI literature (469). He suggests that “the mossy fibre-granule cell–Purkinje cell arrangement could operate as a pattern recognition device,” where the “mossy fibre-granule cell articulation is essentially a pattern separator” (440). Marr then proceeds to mathematically derive constraints on codon size and other measures.

In this case, Marr abstracts from the functional anatomy of the cerebellum to a generic kind K defined by the numbers and types of connections between cell types, with constraints determining loose boundaries. Mathematical derivations uncover the properties of K, and these properties are applied as a hypothesis about the target system, the cerebellum.

In another early treatment of distributed representation, Hinton (1984) describes how sparsely encoded, distributed representations can give rise to properties such as efficient data storage, content-addressable memory, and automatic generalization. These properties are established through both formal derivations and simple connectionist models. Hinton argues that whenever “abstract models are implemented in the brain using distributed representations” we can take these properties to be “primitive operations” (3).

In this case, K is distributed representations, and M instantiates K in a simple network that learns associations between word form and meaning. Here M is chosen to be an instance in which the properties of interest should be difficult to achieve: “This is a case in which distributed representations *appear* to be much less suitable than local ones, because the associations are purely arbitrary” (Hinton 1984, 3). The properties of interest are nevertheless confirmed in M, and the conclusion is drawn that these are properties of K in general. In this case, Hinton’s strategy is to choose a model that seems unlikely to have the property of interest, as a way of demonstrating that the property generalizes across kind members.

These properties of distributed representations have also been confirmed in more detailed models of cortical systems. For example, Babadi and Sompolinsky (2014) analyze the computational benefits of sparseness (few neurons respond to any given stimulus) and expansion (increased dimensionality in the cortical layer) in “generic ensembles of clustered stimuli,” focusing on “relatively simple and biologically plausible architectures and dynamics” (1213). They draw implications for olfactory and visual processing, as well as the mossy fibers of the cerebellum. Billings et al. (2014, 960) investigate sparse encoding in the cerebellum using “biologically detailed network models of spiking neurons, whose parameters were constrained by experimental measurements” in order to determine the contribution that synaptic connectivity makes to effective pattern separation.

These cases define increasingly specific Ks to which cerebellar networks belong. The models in these examples confirm the general properties of the more generic kinds explored in earlier papers and establish a more nuanced

https://doi.org/10.1086/709730 Published online by Cambridge University Press

picture of the properties of the more specific kinds, as well as investigating the boundary cases where the typical properties of the kind break down. As the models get more detailed and realistic, the inferences from model to target are strengthened, because the model and target share more properties, but the scope of the conclusions decreases as the kind becomes more specific. Babadi and Sompolinsky's conclusions also apply to olfactory and visual cortex, while Billings et al.'s conclusions are specific to cerebellum.

There is a continuum here between making more generic theoretical models and more specific models of particular brain areas. Models may be located anywhere between these extremes, with trade-offs between inference strength and generalizability.

6. Conclusion.

There is much to be gained from reconnecting philosophy of AI to philosophy of science. Their estrangement has left a vacuum where methodological critiques of AI ought to be. It would be hard to overstate the urgency with which that kind of work is needed. Likewise, if philosophy of science is to provide credible support against attacks on climate models, accounts of computational models need to go deeper than fiction. Beyond these life-or-death motivations, there is also philosophical and scientific value in embracing AI within philosophy of science.

On the part of AI, we gain an answer to the question of why building some of the constraints that hold of the neural hardware into connectionist models makes them better able to capture cognition, even without any attempt at realistic detail. Insights from philosophy of science help establish that connectionist models can be understood as idealized, multilevel models of the mechanisms underlying cognition. This allows their strengths and weaknesses to be evaluated.

On the part of philosophy of science, a close look at computational models in cognitive science lays the groundwork for a novel epistemology and metaphysics of models, which helps illuminate outstanding problems in the models and simulations literature. Indirectly mediating between models and targets via kinds allows for more specific answers to questions about how to choose relevant similarities in model building, how we justify inferences from models to targets, and the metaphysical nature of models. Artificial neural networks tell us about real cognitive systems by demonstrating the properties of the kinds they both belong to or the aspects they have in common. This analysis may also prove helpful in understanding other types of models, including model organisms and mathematical models.

REFERENCES

Andersen, Holly K. 2017. "Patterns, Information, and Causation." *Journal of Philosophy* 114 (11): 592–622.

Anderson, James A., and Edward Rosenfeld, eds. 2000. *Talking Nets: An Oral History of Neural Networks*. Cambridge, MA: MIT Press.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

Babadi, Baktash, and Haim Sompolinsky. 2014. “Sparseness and Expansion in Sensory Representations.” *Neuron* 83 (5): 1213–26.

Batterman, Robert W. 2001. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. New York: Oxford University Press.

———. 2002. “Asymptotics and the Role of Minimal Models.” *British Journal for the Philosophy of Science* 53:21–38.

Baxter, Donald. 2001. “Instantiation as Partial Identity.” *Australasian Journal of Philosophy* 79 (4): 449–64.

Billings, Guy, Eugenio Piasini, Andrea Lőrincz, Zoltan Nusser, and R. Angus Silver. 2014. “Network Structure within the Cerebellar Input Layer Enables Lossless Sparse Encoding.” *Neuron* 83 (4): 960–74.

Boden, Margaret. 2006. *Mind as Machine: A History of Cognitive Science*. Oxford: Clarendon.

Broadbent, Donald. 1985. “A Question of Levels: Comment on McClelland and Rumelhart.” *Journal of Experimental Psychology: General* 114 (2): 189–92.

Buckner, Cameron. 2018. “Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks.” *Synthese* 195 (12): 5339–72.

Cartwright, Nancy. 1989. “Capacities and Abstractions.” In *Scientific Explanation*, ed. Philip Kitcher and Wesley C. Salmon, 349–56. Minneapolis: University of Minnesota Press.

Chirimuuta, Mazviita. 2018. “Explanation in Computational Neuroscience: Causal and Non-causal.” *British Journal for the Philosophy of Science* 69 (3):849–80.

Churchland, Patricia S., and Terrence J. Sejnowski. 1990. “Neural Representation and Neural Computation.” *Philosophical Perspectives* 4:343–82.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon.

Dennett, Daniel C. 1991. “Real Patterns.” *Journal of Philosophy* 88 (1): 27–51.

Fodor, Jerry A, and Zenon W. Pylyshyn. 1988. “Connectionism and Cognitive Architecture: A Critical Analysis.” *Cognition* 28:3–71.

Frigg, Roman, and James Nguyen. 2016. “The Fiction View of Models Reloaded.” *Monist* 99 (3): 225–42.

Fuhs, Mark C., and David S. Touretzky. 2006. “A Spin Glass Model of Path Integration in Rat Medial Entorhinal Cortex.” *Journal of Neuroscience* 26 (16): 4266–76.

Garson, James. 2015. “Connectionism.” In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University. https://plato.stanford.edu/entries/connectionism/.

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

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

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

———. 2009. “Models and Fictions in Science.” *Philosophical Studies* 143:101–16.

Green, Christopher D. 1998. “Are Connectionist Models Theories of Cognition?” *Psycoloquy* 9 (4). http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?9.04.

Han, Chihye, Wonjun Yoon, Gihyun Kwon, Seungkyu Nam, and Daeshik Kim. 2019. “Representation of White- and Black-Box Adversarial Examples in Deep Neural Networks and Humans: A Functional Magnetic Resonance Imaging Study.” arXiv:1905.02422, Cornell University.

Hempel, Carl G. 1958. “The Theoretician’s Dilemma: A Study in the Logic of Theory Construction.” In *Minnesota Studies in the Philosophy of Science*, vol. 2, ed. Herbert Feigl, Michael Scriven, and Grover Maxwell. Minneapolis: University of Minnesota Press.

Hennig, Boris. 2015. “Instance Is the Converse of Aspect.” *Australasian Journal of Philosophy* 93 (1): 3–20.

Hinton, Geoffrey E. 1984. “Distributed Representations.” CMU-CS-84-157, Computer Science Department, Carnegie Mellon University.

———, ed. 1990. “Connectionist Symbol Processing.” Special issue, *Artificial Intelligence* 46 (1–2).

Irvine, Elizabeth. 2014. “Model-Based Theorizing in Cognitive Neuroscience.” *British Journal for the Philosophy of Science* 67 (1): 143–68.

Kaplan, David M. 2011. “Explanation and Description in Computational Neuroscience.” *Synthese* 183 (3): 339–73.

Khalidi, Muhammad Ali. 1998. “Natural Kinds and Crosscutting Categories.” *Journal of Philosophy* 95 (1): 33–50.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

———. 2013. *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*. Cambridge: Cambridge University Press.

Küppers, Günter, and Johannes Lenhard. 2004. "The Controversial Status of Simulations." In *18th European Simulation Multiconference*, ed. Graham Horton, 271–75. Erlangen: SCS.

Mäki, Uskali. 2012. "The Truth of False Idealizations in Modeling." In *Models, Simulations, and Representations*, ed. Paul Humphreys and Cyrille Imbert, 216–33. London: Routledge.

Marcus, G. 2018. "Deep Learning: A Critical Appraisal." arXiv:1801.00631 [cs.AI], Cornell University.

Marr, David. 1969. "A Theory of Cerebellar Cortex." *Journal of Physiology* 202 (2): 437–70.

———. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

McClelland, James L. 1981. "Retrieving General and Specific Information from Stored Knowledge of Specifics." In *Proceedings of the Third Annual Conference of the Cognitive Science Society*, 170–72. Hillsdale, NJ: Cognitive Science Society.

———. 2009. "The Place of Modeling in Cognitive Science." *Topics in Cognitive Science* 1 (1): 11–38.

McClelland, James L. and David E. Rumelhart. 1985. "Distributed Memory and the Representation of General and Specific Information." *Journal of Experimental Psychology: General* 114 (2): 159–97.

———. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2, *Psychological and Biological Models*. Cambridge, MA: MIT Press.

Miłkowski, Marcin. 2013. *Explaining the Computational Mind*. Cambridge, MA: MIT Press.

Morgan, Mary S. 2002. "Model Experiments and Models in Experiments." In *Model-Based Reasoning: Science, Technology, Values*, ed. Lorenzo Magnani and Nancy J. Nersessian, 41–58. New York: Kluwer.

———. 2003. "Experiments without Material Intervention: Model Experiments, Virtual Experiments and Virtually Experiments." In *The Philosophy of Scientific Experimentation*, ed. Hans Radder, 216–35. Pittsburgh: University of Pittsburgh Press.

Newell, Allen, and Herbert A. Simon. 1961. "Computer Simulation of Human Thinking." *Science* 134 (3495): 2011–17.

———. 1976. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the ACM* 19 (3): 113–26.

Norton, Stephen, and Frederick Suppe. 2001. "Why Atmospheric Modeling Is Good Science." In *Changing the Atmosphere*, ed. Clark A. Miller and Paul N. Edwards, 67–105. Cambridge, MA: MIT Press.

Parker, Wendy. 2009. "Does Matter Really Matter? Computer Simulations, Experiments, and Materiality." *Synthese* 169 (3): 483–96.

Plaut, David C. 1995. "Double Dissociation without Modularity: Evidence from Connectionist Neuropsychology." *Journal of Clinical and Experimental Neuropsychology* 17 (2): 291–321.

Rumelhart, David E., and James L. McClelland. 1985. "Levels Indeed! A Response to Broadbent." *Journal of Experimental Psychology: General* 114 (2): 193–97.

———. 1986a. "On Learning the Past Tenses of English Verbs." In McClelland and Rumelhart 1986, 216–71.

———. 1986b. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, *Foundations*. Cambridge, MA: MIT Press.

———. 1986c. "PDP Models and General Issues in Cognitive Science." In Rumelhart and McClelland 1986b, 110–46.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–52.

Sejnowski, Terrence J., and Charles R. Rosenberg. 1986. "NETtalk: A Parallel Network That Learns to Read Aloud." Technical Report JHU/EEC-86/01, Electrical Engineering and Computer Science, Johns Hopkins University.

Smolensky, Paul. 1988a. "The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn." *Southern Journal of Philosophy* 26 (S1): 137–61.

https://doi.org/10.1086/709730 Published online by Cambridge University Press

———. 1988b. “On the Proper Treatment of Connectionism.” *Behavioral and Brain Sciences* 11:1–74.

———. 1991. “Connectionism, Constituency, and the Language of Thought.” In *Meaning in Mind: Fodor and His Critics*, ed. Barry M Loewer and Georges Rey, 201–27. Oxford: Blackwell.

Steinle, Friedrich. 1997. “Entering New Fields: Exploratory Uses of Experimentation.” *Philosophy of Science* 64:S65–S74.

———. 2002. “Experiments in History and Philosophy of Science.” *Perspectives on Science* 10 (4): 408–32.

Stinson, Catherine. 2018. “Explanation and Connectionist Models.” In *The Routledge Handbook of the Computational Mind*, ed. Matteo Colombo and Mark Sprevak, 120–33. London: Routledge.

Suárez, Mauricio. 2003. “Scientific Representation: Against Similarity and Isomorphism.” *International Studies in the Philosophy of Science* 17 (3): 225–44.

Suri, Roland E., and Wolfram Schultz. 2001. “Temporal Difference Model Reproduces Anticipatory Neural Activity.” *Neural Computation* 13 (4): 841–62.

Thomas, Michael S. C., and James L. McClelland. 2008. “Connectionist Models of Cognition.” In *Cambridge Handbook of Computational Psychology*, ed. Ron Sun, 23–58. Cambridge: Cambridge University Press.

Touretzky, David S., and Geoffrey E. Hinton. 1988. “A Distributed Connectionist Production System.” *Cognitive Science* 12 (3): 423–66.

Weisberg, Michael. 2012. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Wilson, Hugh R., and Jack D. Cowan. 1972. “Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons.” *Biophysical Journal* 12 (1): 1–24.

Winsberg, Eric. 2009. “A Tale of Two Methods.” *Synthese* 169 (3): 575–92.

———. 2010. *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

https://doi.org/10.1086/709730 Published online by Cambridge University Press