What Artificial Neurons Tell us about Real Brains

Catherine Stinson*

Rotman Institute of Philosophy, University of Western Ontario

Email: cstinso5@uwo.ca

Abstract

The neural plausibility of connectionist models is supposed to be their selling point, yet they are notorious for being neurally implausible. Despite several decades of debate about the merits of connectionism, this central puzzle remains unresolved. My suggestion is that we understand connectionist models as idealized, abstract models of neural mechanisms. This not only makes sense of some statements by the Parallel Distributed Processing (PDP) Research Group about their methodology that have been long misinterpreted, but also affords a more nuanced picture of the many levels at which connectionist models might be pitched than the views often assumed in discussions of connectionism. Connectionist models may target

*Thanks to the Max Planck Institute for History of Science, the Centre for Integrative Neuroscience, Tübingen, and the Rotman Institute of Philosophy for providing funding during the writing of this paper. I am grateful to Peter Machamer, Kenneth Schaffner, Elizabeth Irvine, Sarah Robins, Tim Bayne, and Steve Gotts for helpful comments. Thanks also to Melissa Jacquart for motivational prodding.

any degree of granularity from highly detailed to completely abstract. This flexibility is an explanatory virtue of connectionism, not a drawback or a sign of vague definitions.

1 Introduction

There is a vast philosophical literature about computational models of cognition which focuses on the architecture of cognition, yet hardly touches on modeling methodology at all. There is also a growing body of philosophical work about models and simulations, which focuses on methodological questions, yet hardly mentions examples from cognitive science at all. Here I connect these discussions, and examine computational modeling methodology in cognitive science using insights developed in the modeling and simulations literature. In doing so I resurrect a central problem that was raised early on in discussions of computational models of cognition, but faded away without being resolved.

The central problem I address is why neural plausibility should be considered helpful in developing cognitive models. Furthermore, if there is some benefit to using neurally plausible models, I ask why connectionists don't attempt to make their models more realistic. Despite the lull in philosophical interest in connectionism, these issues are still very much alive, since connectionist models have remained popular in the cognitive and brain sciences, and a new generation of deep-learning models is quickly taking over as the dominant method for many machine learning tasks. Similar questions could also be asked of predictive coding, which likewise leverages neural plausibility as an argument for cognitive plausibility.

I consider the problem of why and how connectionist models should be made neurally plausible in terms of a more general issue in scientific modeling: How closely (and in what ways) do models need to resemble their target systems in order to produce relevant, generalizable results? Much progress has recently been made on this general issue, which when applied to the case of cognitive models, illuminates the matter. During connectionism's heyday, we lacked the philosophical vocabulary to unambiguously describe this methodology, but are now in a position to do so.

First I introduce connectionism, and look at what its early supporters say in defense of the approach. I lay out the dilemma that connectionist models' simultaneous endorsement of neural plausibility and failure to follow through on neural plausibility raises about what exactly these models are intended to do. Then I show how recent philosophical work on models and simulations can resolve the dilemma. I argue that connectionist models are abstract, idealized models used for discovering and exploring multi-level mind/brain mechanisms.

2 Connectionist Models

Although it has a longer history, contemporary interest in connectionist modeling stems largely from the Parallel Distributed Processing (PDP) Research Group, whose two-volume publication (Rumelhart and McClelland, 1986b; McClelland and Rumelhart, 1986) sparked renewed interest and philosophical debate about methodology in cognitive science.

Connectionist models have an architecture roughly analogous to networks of neurons. They consist of a number of simple units with input connections from and output connections to other units. This mimics the structure of real neurons which typically receive input through their dendrites, then provide output by generating action potentials down their axon. The standard connectionist network architecture is a three-layer, feedforward neural network, 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 many more than 3 layers. The activity of the network is defined by each unit's activation, and each connection's weight. The activation of a unit is a function of the weighted sum of its input activations: typically a sigmoid function on $[0, 1]$. The weights typically start with random values, then are adjusted using a learning rule designed to minimize overall error.

I define connectionist models broadly to include any computational modeling architecture consisting of multiple simple units governed by local activation rules. On this description, connectionist models are a quite general class of computational model, however, the PDP approach takes inspiration more narrowly from neural architectures. Connectionist models are widely used as tools in contexts unrelated to cognitive science, such as engineering and applied sciences, but my focus here is restricted to the use of connectionist models in cognitive science. 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). This statement suggests an intention to model the physiology of the brain, but on close inspection proves hard to interpret. What is meant by 'flavor'? Why is 'physiological' in scare quotes? In virtue of what is having a 'physiological' flavor appealing?

The other kinds of information-processing model referred to in the quote are classical approaches to artificial intelligence (AI) where beliefs, goals, etc., are represented as symbols, and processed using logical rules. The main motivation the PDP group had for seeking alternative architectures was that classical AI's symbolic models seemed unsuited for some kinds of computations that intelligent creatures perform:

> 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).

Although the PDP bible's first printing quickly sold out, the reaction among the AI mainstream was largely critical. In order to understand this reaction, it's important to recall just how groundbreaking classical AI was for our understanding of the mind. Before people like McCarthy, Newell, Simon, Winograd, Schank, and Minsky (among others) programmed machines to act intelligently, we had no serious contenders for theories of how the mind works. Classical AI provided a clear way of understanding something that had previously been mysterious. So rejecting classical AI threatened to leave us, once again, in the dark about how the mind works. The PDP Group needed to present a clear alternative.

The alternative theory of mind that they were widely interpreted as presenting was one where folk, and cognitive psychology (the symbolic level) were eliminated as mere appearance, and instead, the real work of cognition happened at the lower level of brain physiology. PDP was taken as a rejection of the cognitive part of cognitive science, and an embrace of a bottom-up methodology, where implementation-level or neural details are all that matter. The PDP Group's rhetoric about brain physiology and biological hardware certainly invites this interpretation.

Although an understandable reaction, I think this is a misunderstanding of the PDP approach. One thing that is evident, but often overlooked, from the quotes above is that the PDP group's project was, in at least one sense, very much in the spirit of classical AI and cognitive science: its concern with building models that simulate[1] data from cognitive psychology. Paying attention to reaction times, and revising your model when its predictions don't match the data is a move straight out of the cognitive psychologist's toolbox. Hinton, Rumelhart, and McClelland, the main players in the PDP group, were all trained as psychologists, after all.[2] Rather than trying to eliminate the symbolic level, they were trying to explain it, and in particular, to explain the cases that classical AI didn't seem to get right.

Interpreting their talk about biological hardware as implying an exclusive focus on implementation details or taking neural physiology to be the only level that is important to connectionism is a misunderstanding, I think, but again a tempting one. In the next section I discuss this reaction, and provide an alternative interpretation of what the microstructure of cognition might be.

3 The Implementation Dilemma

The first major challenge to the PDP Group's work came from Broadbent (1985), who argued that McClelland and Rumelhart (1985) had inappropriately cast their distributed memory system as having "implications at the psychological and not merely at the physiological level" (Broadbent, 1985, 189). Broadbent explicitly invokes Marr (1982) in reference to levels. Marr distinguished between computational, algorithmic, and implementation levels of analysis, and argued that in work on computer vision, neglect of the computational level had led to implementations that failed because they were trying to solve the wrong problem. A lesson often drawn from Marr (1982) is that computational, algorithmic, and implementation levels are independent.[3]

In a paper that inspired a cottage industry on the systematicity of cognition, Fodor and Pylyshyn (1988) posed 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 intended as computational level (i.e., psychological) models, then implementation (i.e., neural) details should be irrelevant, given independence of levele, so neural plausibility should afford no advantage to models of cognition. If PDP models are intended as implementation level models, then they might be interesting to neuroscientists, but they aren't cognitive science. Or so the story goes.[4]

It would require many pages to list all the sources where variations of this reaction to connectionism appear. 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). There is some truth to this. Some connectionist projects, such as the articles in Hinton (1990), are concerned with showing that

PDP models are capable of structured representations, and serial processing, i.e., implementational connectionism. Other types of work, such as Plaut (1995), attempted to show that what looks like serial processing on the surface, might be better explained in terms of network-level details, i.e., radial connectionism. A weakened, not-so-radical connectionism claiming that symbolic processing was a bad guess at how *some* mental functions work, would be much more plausible, and probably closer to what most connectionists believe.

In response to Fodor & Pylyshyn's dilemma, connectionists pose the counter-challenge that classical AI does not capture cognition adequately. Smolensky (1988a, 12) claims that the symbolic level at best provides approximations. McClelland and Rumelhart (1986, 12) make a similar claim: "macrostructural models of cognitive processing are seen as approximate descriptions of emergent properties of the microstructure." This reference to emergence suggests a different sort of inter-level relation than Marr's account of levels, although one that is left unclear.

Connectionists do not seem to accept that their models are implementations. Smolensky (1988a) describes them as being at the "sub-symbolic level," which avoids the term implementation, while indicating something lower than symbolic. Elsewhere, Smolensky says that the goal of connectionist research is not "the implementation of a symbolic language of thought" but a "middle ground between implementing symbolic computation and ignoring structure" (Smolensky, 1988b, 152). Rumelhart and McClelland (1985, 193) object to the assumption that cognitive models should occupy the computational level, arguing instead that much of what concerns cognitive psychologists is at the algorithmic level; they want to know how cognition happens, not just which problem is being solved. In a later paper, McClelland argues for the virtues of implementing one's theories. He argues that unimplemented theories can be vague on points that seem innocuous but turn out to be important, and claims connectionist models have "forced more detailed specification of proposed cognitive models via implementation" (Thomas and McClelland, 2008, 48). However, here implementation probably means writing and running a program, not Marr's third level.

What I gather from these claims is that what the PDP group call the microstructure of cognition is at a lower level than cognition in some sense of level, but that these are not Marr's levels. The PDP Group sees the symbolic level as being *dependent* on the microstructure, although not necessarily predictable based on it. The virtue of actually building an implementation is that you get to see what emerges from the microstructure. For at least some cognitive tasks, symbolic models don't seem plausible because they tend to be inflexible, and get slower as tasks get more complex. The PDP approach is to try to discover more plausible models by starting from what they know: whatever the algorithm is, it is built up from the simple, local computations operating on nodes in a connected neural network. As a method, this bears a closer resemblance to exploratory experimentation than to theoretical modeling. The point is to tinker with the parts and see what they can do.

If we grant that classical AI was beginning to show some weaknesses by the mid-1980s, and recall that criticisms of multiple realizability (i.e., the independence of levels) had begun to circulate by that time, this seems like a reasonable approach. But one puzzle remains: if the approach is to try to see what emerges from models that are brain-like, why is so little attention paid to making the models brain-like? If cognition is supposed to emerge, the neural details might be supposed to matter very much indeed.

3.1 Neural Plausibility

The PDP Group's rhetoric about physiology, the fact that the structure of the units in a connectionist network mimics neurons, and the claim that cognition emerges from the microstructure all clearly suggest that PDP models are meant to model the brain. However, critics charge that PDP models are not, in fact, very neurally plausible. Connectionist models lack some key properties of real brain networks, and instead have some properties that are not true of brains.

The backpropagation algorithm, for example, is infamous for not being neurally plausible, because error signals cannot in general be propagated backward through a network of neural connections, as the algorithm seems to require.[5] Furthermore, dendrites don't just sum their inputs the way artificial neural network units do; there is morphological and physiological variety among neurons that standard models fail to reflect, and neurons are not typically connected to every other neuron within a system, the way a standard 3-layer feedforward model is. These are just a few of the most obvious dis-analogies between connectionist networks and brains.

Although critics of connectionism have for decades treated these dis-analogies as though they were mistakes, in fact the PDP group were well aware that the 'physiological' flavor of PDP models stopped well short of realistic detail. Volume 2, Chapter 20 of the PDP books is devoted to detailing the ways in which artificial neural networks are not like real brains (Crick and Asanuma, 1986). The introduction to the volumes 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. They hold out the hope of offering computationally sufficient and psychologically accurate mechanistic accounts of the phenomena of human cognition (McClelland and Rumelhart, 1986, 11).

Curiously, lack of realistic detail is a feature that was included by design.

Central to the issue of what microstructure means is how we should interpret single units in a connectionist network. In networks with local representations, units are assigned 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, 2). Distributed memory networks are harder to interpret, but they make for more plausible psychological models because of the ability to recall memories based on part of the content, and to generalize from a few examples, the way human memory can.

Although distributed networks have units that in some ways resemble neurons, they are not always meant to stand in for individual neurons. Far too few units are used in most connectionist models to be realistic brain models (although there are now projects to build models on realistically large scales too). In some connectionist networks, units explicitly stand-in for whole populations of neurons, with the activation of the unit representing a population vector. 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. There is thus considerable flexibility in what a unit is meant to correspond to: a concept, a neuron-like part of a distributed memory network, or a population vector.

An example that aims to model a psychological phenomenon, but without going into much detail about the physiology, is the past-tense learner (Rumelhart and McClelland, 1986a), one of the PDP group's early showpieces. This network takes English verbs as inputs, and learns to output their past tenses. It is trained using a series of examples of common English verbs, including both regular verbs (where to form the past tense, one simply adds "ed") and irregular verbs (like went or swam). The past-tense learner's success in learning to conjugate past tenses without any explicit set of rules or separate processing streams for regular and irregular verbs was, as Boden put it, "theoretical dynamite: It cast doubt on nativism and modularity... It undermined belief in the psychological reality of explicitly represented processing rules, and of explicit (symbolic) representations. And it threatened the then popular forms of philosophical functionalism" (Boden, 2006, 956). However, there is no expectation that the architecture of the brain system that computes past tenses of verbs matches that of the past-tense learner in any great detail. The past-tense learner's most glaring failure to simulate realistic physiological detail is the representation of its input and output verbs, which are encoded somewhat idiosyncratically as phonetic triples called 'Wickelfeatures'. The microstructure of cognition must be something quite flexible, allowing for some properties of the physiology to be modeled without much attention to detail. It clearly does not mean accurate physiological detail.

3.2 Abstraction

Yet another sort of description that connectionists give of their models complicates matters further. Smolensky (1991, 202) claims that connectionism is committed to "uncovering the insights this other half of mathematics [continuous rather than discrete] can provide us into the nature of computation and cognition." Thomas and McClelland (2008, 23) call connectionist models, "a sub-class of statistical models involved in universal function approximation." McClelland (2009, 20) talks about demonstrating sufficiency and optimality results. These comments all suggest that connectionist models, at least sometimes, are intended as general proofs rather than detailed simulations.

An example of this sort of work is Touretzky and Hinton (1988), which shows "how 'coarse coding' or '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" (Touretzky and Hinton, 1988, 423). In this sort of network, the units don't correspond to any particular thing. The point is to demonstrate a property this sort of network has no matter what the units represent. At the same time, this general point is clearly meant to reflect on how working memory, a cognitive construct, might work.

Certainly part of what is going on here is that practical concerns require that models not be too complex. Dave Touretzky quips to his graduate classes, that "putting too much detail into a model is a novice mistake" (personal experience). This is a common refrain among connectionists. Jack Cowan says, "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" (Anderson and Rosenfeld, 2000, 123). McClelland (2009, 18) argues that while there is a cost to making simplifications in modeling, it is necessary to simplify in order to achieve understanding.

There are compelling practical reasons for modelers to make their models as simple as possible. The extreme would be mathematical models that contain no specific detail. Mathematical models have the advantage of being generalizable, but the disadvantage of not always being directly applicable to messy, real-world situations. Realistic, detailed models lack generalizability, but can accurately describe messy real-world situations. Connectionist models seem to try to combine these two very different sorts of model. Why should it be a good idea to combine the two, yielding models that are neither accurate nor general?

3.3 From Microstructure to Cognition

This strategy of making models partly but only partly neurally plausible stands in striking contrast to how the inferences from AI program to cognition are meant to work in classical AI. Newell and Simon (1961) postulated that cognition is produced by elementary information processing over symbols (this later became their "physical symbol system hypothesis" (Newell and Simon, 1976, 116)), and that neurophysiological mechanisms in turn produce these information processes.

They set aside the second half of the project for practical reasons: "Tunneling through our mountain of ignorance from both sides will prove simpler... than trying to penetrate the entire distance from one side only" (Newell and Simon, 1961, 2012-2013).

Classical AI investigates how the intermediate information processing level produces cognition, by building computer programs designed to produce output comparable to human cognitive behavior. A computer program that produces the right output is then considered a 'theory' of human thought, according to Newell and Simon. They describe these theories in terms of the Syntactic View of theories (see 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). In other words, the logical calculus in the computer program is intended to have the status of a theory in the physical sciences.

This method is closely related to simulation, where known fundamental equations form the basis of a program, and the predicted output is carefully calculated. In classical AI, however, the fundamental equations begin as an unknown, and successful simulation of behaviour is taken as evidence of the program or theory's correctness. This is best understood as an inference to the best explanation; however, at the time, it was so remarkable to have any explanation at all of intelligent behaviour, that comparisons with competing explanations were not explicitly considered.

During the period between 1961 and 1986, the Syntactic View ceased to be the 'received view' in philosophy of science, but it had not yet been overtaken by the now popular mechanistic account of scientific explanation (see Bechtel and Richardson (1993); Glennan (1996); Machamer et al. (2000)). PDP models are clearly not intended as theories in the sense of the Syntactic View; they are not highly accurate, detailed models of the underlying structure, and aren't meant to predict behaviour in great detail. We have seen that the nodes in a PDP network can correspond to concepts, neurons, population vectors, or they can be abstractions, corresponding to nothing in particular. We have also seen that the goal of these networks might range from elucidating how a particular cognitive task is performed, to demonstrating quite general properties of connected networks of local units. I argue in the next section that they are best understood as a method for discovering mechanistic explanations. That they were avant-garde relative to philosophical developments may explain why their methods were so controversial and not well understood when first introduced.

4 Idealized, Abstract Models of Mechanisms

Understanding PDP models as idealized, abstract models of mechanisms resolves the puzzle of why it is important for PDP models to have a physiological flavor, but not realistic physiological detail. To see why being a little bit realistic, but not very realistic, is an advantage, we need an account of which details need to be captured accurately for a model to inform us about the target system, and which can be safely abstracted away. This problem has been the subject of much recent work in philosophy of science. Further philosophical developments that will shed light on the puzzle are a more flexible account of levels; a defense of abstract, idealized explanatory models; and a pluralistic view of the role of scientific models in discovery.

4.1 Levels

We saw earlier that Marr's levels were not very useful for locating connectionist models. A more suitable notion of level was suggested very soon afterward. Bechtel argues that connectionists "are not concerned simply with levels of analysis but with *levels of structure* in nature," (Bechtel, 1994, 9) and that levels of structure are better thought of in terms of mechanistic explanation. Mechanistic explanation (Bechtel and Richardson, 1993; Glennan, 1996; Machamer et al., 2000) involves situating a phenomenon within a multi-level hierarchy of mechanisms. This means identifying its role in higher-level mechanisms, and figuring out how its component parts and their activities are organized to bring about that phenomenon. Each level constrains and is constrained by its neighbouring levels.

This view is very much in tune with the PDP group's stated motivations, and at odds with the popular reading of Marr. Rather than seeing physiology and cognition as independent of each other, connectionists explore the ways in which the physiological microstructure constrains cognition. The PDP bible lists the constraints they take from neuroscience, including: "*Neurons are slow... 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, 130-132). Thinking of connectionist models in terms of levels of mechanisms also helps explain why the units can correspond to single neurons, populations of neurons, or much higher-level entities like phonetic representations. Connectionist models can be used to investigate any number of locations in a hierarchy of mechanisms. There is no one privileged level in a mechanistic explanation.

4.2 Simplicity as an Explanatory Virtue

As mentioned earlier, connectionists think of simplicity as essential not only for getting models to work, but also for making them explain. The cost of simplification is that the inferences you draw from simplified models might be challenged on the grounds that the interesting emergent properties of the model might result from the differences rather than from the similarities between the model and its target. Connectionist models are different from brains in many ways, so naturally we might expect them to behave differently. This is an instance of a very general worry about scientific models, and one about which philosophers of science have had much to say.

One kind of simplification is abstraction. Abstract models remove details so that the effect of a small number of variables can be more easily investigated. 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 not only practical, but also perfectly legitimate to ignore some of the details. An apt 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 lab experiments that control more variables. Lab experiments also involve idealizations, where sometimes more convenient materials are substituted in, or implausible parameter values are chosen for ease of calculation..

Cartwright (1988) distinguishes abstraction from idealization. Idealization does not just remove detail, it adds or changes details, such that the idealized model has properties not present in the target system. Treating an inclined plane as frictionless is a typical example. Many idealizations are involved in connectionist models, such as a node's activation function being deterministic, whereas real action potentials are stochastic. It seems like putting the wrong details into a model should be a bad thing, but in practice, making a model less accurate sometimes makes its results more reliable, as a number of people have argued (Küppers and Lenhard, 2004; Parker, 2009; Winsberg, 2009; Morrison, 2015).

Both models and experiments need *external validity*, i.e., the conclusions they draw need to be true not just of the model or experimental system, but also of the target. Morgan (2002, 2003) argues that traditional experiments have closer access to their targets than models do, because they manipulate the same materials. Winsberg disagrees. He argues that "almost all scientists... rely in the end on arguments... that the results that they get from manipulating their respective pieces of equipment are appropriately probative concerning the class of systems that interest them" (Winsberg, 2009, 577). For experiments, those arguments depend on similar materials being used in the experiment as are present in the target system, according to Morgan. For models, those arguments are based on having knowledge about how to build good models, which comes from past successes using the same bag of tricks, according to Winsberg.

However, in both cases, both sorts of considerations matter. You need background knowledge and skills for successful experiments. Likewise, with models, manipulating the same *kinds* of objects helps to ground inferences. In experiments, this usually means material similarity, but structural similarities can also ground inferences. While materials determine which properties an object has, the laws we discover don't tend to be about specific materials. For example, when circles or cylinders are closely packed together, the highest density arrangement is a hexagonal pattern. This is true regardless of whether the circles are the cross-sections of telecommunications cables or axon collaterals in nerve tracts. In a model we'll discuss in more detail in the next section, 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. We can learn something about how the brain works using their model precisely because the model is abstract and idealized; we wouldn't be able to apply the fact about close packing of cylinders unless we assumed that nerve fibers are perfectly cylindrical and of uniform size. A more accurate, detailed model might miss this helpful association between nervous tissue and idealized cylinders.

My interpretation of "physiological flavor" is that connectionist models are abstract, idealized models of brain physiology. If you abstract away many of the details, and idealize others, cortex is an interconnected network of simple learning units, so what is true of those should be true of the brain (or parts of it), ceteris paribus. The brain being of roughly this type is what grounds inferences from connectionist models to cognition. There is a continuum between simulating brains and making purely abstract mathematical models, and connectionist models may be located anywhere between these extremes, depending on their purpose. Drawing inferences from more-or-less abstract models to cognition is still susceptible to error, but this risk is not substantially different from the risk of error when drawing inferences from experiments, which also either remove or alter many details present in their target systems.

4.3 Models with Many Uses

A final step is to recognize that models serve many different purposes in science, and many different strategies may be employed in the search for mechanisms. There are likewise various motivations and purposes for using connectionist models. Anderson and Rosenfeld's (2000) history of connectionism demonstrates that among connectionist modelers there were widely differing views right from the start about what level of detail is best, how much physiological detail is desirable, and what the goals are. These goals include engineering, mathematical, psychological, 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). The same could be said 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 and abstract or accurate and detailed they should be in order to meet their goals.

This is true of PDP models. Some features are modeled accurately and others less accurately depending on the modelers' epistemic goals. If the point of the past-tense learner had been to model how verbs are represented in the brain, or to simulate in detail how conjugation of verbs is done, then it would have been inappropriate for Rumelhart and McClelland to encode their input and output verbs as Wickelfeatures. But neither of these were the point. Instead they wanted to see whether verb conjugation that looks like structured rule-following behaviour could achieved simply by training a brain-like network on examples of verb conjugations. The way the verbs are represented was bracketed off as unimportant, given their goal. Likewise, if the point of a model were to figure out the brain mechanisms responsible for learning, then using backpropagation would be inappropriate. Suri and Schultz (2001) is an example of a model of learning mechanisms, and there the anatomy of the basal ganglia is modeled in detail, including only pathways that exist in the brain and through which feedback is thought to actually travel. However, NETtalk (Sejnowski and Rosenberg, 1986) uses backpropagation, which is fine in their case, because their goal was to show that a system capable of pronouncing English words needn't encode a complicated set of rules. For that purpose, they could simply assume that the brain has some way of propagating error signals, without worrying about how exactly that happens. What counts as an acceptable simplification depends on the model's epistemic goals.

I now describe in some detail a research project that illustrates how the accuracy of the models used varies with the epistemic goals at different stages of the project.

4.4 Example: Discovering The Mechanisms of Path Integration

The hippocampus is one of the main brain areas responsible for memory, and much of the research on it is done by studying rats finding their way around mazes, often wearing microelectrode arrays to record neural activity. Place cells in the hippocampus fire when the rat is in a particular area or field in an environment. Different populations of cells code different locations, and together they form a cognitive map.

Rats are capable of path integration: they can find a direct route back to a goal location (such as their nest, a food source, or a pleasant smell) from wherever they are in an environment. They can do this no matter what route they took to get there, and even if they can't see the goal location. So this ability is neither dependent on remembering their route, nor on visually orienting themselves in the environment based on landmarks. It seems that the route taken is somehow combined with their end location to generate a sense of the geometry of the space.

Until recently it wasn't known where in the brain path integration might be performed, nor how it worked. Redish and Touretzky (1997) proposed a connectionist model of how information from place cells could be combined with information about head direction and self-motion to integrate paths, and laid out criteria that must be met by the mechanism responsible for this ability. Anatomists subsequently (Fyhn et al., 2004) found grid cells in entorhinal cortex that happened to have the criteria predicted by the model. The discovered grid cells turned out to also have a peculiar property that noone had predicted: their firing fields form hexagonal grids at various scales and orientations.

Fuhs and Touretzky (2006) then proposed a more detailed connectionist model of path integration, which showed that "hexagonally spaced activity bumps can arise spontaneously on a sheet of neurons in a spin glass-type neural network model" (Fuhs and Touretzky, 2006, 4266). Spin glass models are a type of connectionist network where each unit is connected to its closest neighbors in a multi-dimensional grid. The spin glass model described how the hexagonal grids might arise, based on the local network structure in entorhinal cortex, and the assumption that dendrites are closely packed. As mentioned earlier, if dendrites are modeled as cylinders, their ideal close packing arrangement should be hexagonal.

This second model was later supplanted by Burgess et al. (2007) (which is not a connectionist model). Burgess's model explains the hexagonal grid pattern as the effect of interference between dendritic subunits tuned to different directions. One reason why this model was preferred is that it unifies the explanation of the hexagonal grid pattern with another property of grid cells that had until then presented a puzzle (called the phase precession effect) as both resulting from the same underlying mechanism. They also note that the effects Fuhs and Touretzky (2006) describe might be added to their model "to maintain the relative locations of the grids and enhance their stability and precision" (Burgess et al., 2007, 810).

Within this research project, computational models played several roles. First a model was proposed requiring a fairly general kind of physiological mechanism.

The functional description of this mechanism suggested properties for an unknown anatomical region. Once a region with those properties was found and further investigated by physiologists, its newly discovered properties required explanation. A more detailed model was built explaining both how the newly discovered properties might arise from plausible physiological conditions, and how that component might perform its function within a larger mechanism. An alternative model was then built that could explain the same phenomenon in a different way, and unify that explanation with another, related phenomenon.

In short, computational models were used to show how a mechanism might possibly work, propose constraints on anatomical exploration, to provide possible mechanisms to explain a more detailed effect, to show how several components might function together in a complex system, and to suggest an alternative, more unified mechanistic model. Gradually more physiological detail was added as the project progressed.

4.5 Discovering Mechanisms

This episode fits the 'new mechanist' picture of mechanism discovery described in (Machamer et al., 2000; Darden, 2002). The path integration system began as a gappy functional description. Redish and Touretzky (1997) proposed a mechanism sketch, described the features they expected to find in the path integrator, what each of the known entities must do, and how they all should work together to provide productive continuity. Some of the details of other entities were known, but the path integrator remained sketchy. The path integrator was then identified by physiologists, and some details about it were filled in. Fuhs and Touretzky (2006) proposed a sketch of how that sub-mechanism might work, plus added details describing how it fit into the mechanism as a whole. In doing so, they instantiated the spin-glass schema which originated in physics. Burgess et al. (2007) provided an alternative mechanism sketch, which connected laterally to a previously proposed mechanism for another phenomenon, and added more details, bringing it closer to being a fully-specified mechanism.

Thinking of connectionist modeling as a set of tools for discovering and exploring models of mechanisms provides a better characterization of the role they play in cognitive science than the various suggestions made by supporters and critics of connectionism. The inferences we can draw from these more-or-less abstract, more-or-less idealized models to cognition are no less reliable and of much the same type as the inferences drawn in experimental science generally.

5 Conclusion

The question I wanted to resolve here was why the supposed advantage of connectionist models is that they are physiologically plausible, when in fact they're not. Connectionists pointed out that classical AI doesn't always do a good job of simulating the psychological data, and hypothesized that building in some of the constraints that hold of the neural hardware might do a better job. However, connectionists' explanations of how exactly their models work were difficult to interpret. Although lip service was paid to connectionist models being physiologically plausible, they clearly fall far short of being realistic simulations of brains. At the same time, they are touted as being mathematical demonstrations, while not being purely abstract either.

The puzzle of what level connectionist models are pitched at was resolved by recognizing connectionist models as potentially occupying various levels in a multi-level hierarchy of mechanisms, which are orthogonal to Marr's levels of analysis. The puzzle of why connectionist models are neither realistic enough to be brain models nor abstract enough to be mathematical demonstrations was resolved by recognizing that they can be more-or-less abstract and idealized, depending on their epistemic goals, and that they work in much the same way as models and experiments in other branches of science. Inferences from connectionist models are grounded based on the target being of the same abstract kind as the model, rather than being materially similar.

I illustrated these points with classic examples of PDP networks, and a contemporary example of a research project on the discovery of the mechanism of path integration.

Notes

[1]'Simulating' here means producing output that matches the results of psychological experiments, and should not be confused with simulation in the sense of approximating solutions to differential equations, as it is typically used in philosophy of science.

[2]Their identity as psychologists is so much overlooked that an early reviewer of this paper not only didn't believe that Hinton was trained in psychology, but suggested that people in the know would find my mistake funny.

[3]This is not quite what Marr argued for, but that's a story for another time.

[4]The bulk of the debate stemming from this challenge focused on whether PDP models are capable of the sorts of structured representations Fodor & Pylyshyn deem necessary for cognition.

I won't address this literature here.

[5]This was at the time a worrisome criticism. It is now arguably a moot point, because alternative learning algorithms have since been developed that don't have the same requirements. In addition, it has been argued that "the difference between activations computed using standard feedforward connections and those computed using standard return connections can be used to derive the crucial error derivatives required by backpropagation" (Thomas and McClelland, 2008, 29).

References

Anderson, J. A. and Rosenfeld, E. (2000). *Talking Nets: An Oral History of Neural Networks*. MIT Press.

Bechtel, W. (1994). Levels of descriptions and explanation in cognitive science. *Minds and Machines*, 4:1–25.

Bechtel, W. and Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization As Strategies in Scientific Research*. Princeton University Press, Princeton.

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford University Press.

Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General*, 114(2):189–192.

Burgess, N., Barry, C., and O'Keefe, J. (2007). An Oscillatory Interference Model of Grid Cell Firing. *Hippocampus*, 17:801–812.

Cartwright, N. D. (1988). Capacities and Abstractions. In Kitcher, P. and Salmon, W., editors, *Scientific Explanation*, pages 349–356. University of Minnesota Press.

Crick, F. and Asanuma, C. (1986). Certain Aspects of the Anatomy and Physiology of the Cerebral Cortex. In McClelland, J. L., Rumelhart, D. E., and The PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Volume 2*, chapter 20, pages 333–371. MIT Press.

Darden, L. (2002). Strategies for Discovering Mechanisms: Schema Instantiation, Modular Subassembly, Forward/Backward Chaining. *Philosophy of Science*, 69(3):S354–S365.

Fodor, J. A. and Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, 28:3–71.

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

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., and Moser, M.-B. (2004). Spatial Representation in the Entorhinal Cortex. *Science*, 305(5688):1258–1264.

Garson, J. (2015). Connectionism. *Stanford Encyclopedia of Philosophy*.

Glennan, S. (1996). Mechanisms and the Nature of Causation. *Erkenntnis*, 44(1):49–71.

Hempel, C. G. (1958). The theoretician's dilemma: A study in the logic of theory construction. In Feigl, H., Scriven, M., and Maxwell, G., editors, *Minnesota Studies in the Philosophy of Science*, volume II. University of Minnesota Press, Minneapolis.

Hinton, G. E. (1984). Distributed representations. Technical Report CMU-CS-84-157, Carnegie Mellon University, Computer Science Department.

Hinton, G. E., editor (1990). *Artificial Intelligence: Special Issue on Connectionist Symbol Processing*, volume 46.

Küppers, G. and Lenhard, J. (2004). The Controversial Status of Simulations. In *Networked Simulations and Simulated Networks: 18th European Simulation Multiconference. Erlangen: SCS*, pages 271–275.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, 67(1):1–25.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman & Company, New York.

McClelland, J. and Rumelhart, D. E. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2 Psychological and Biological Models*. MIT Press.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the Third Annual Conference of the Cognitive Science Society*, pages 170–172.

McClelland, J. L. (2009). The Place of Modeling in Cognitive Science. *Topics in Cognitive Science*, 1(1):11–38.

McClelland, J. L. and Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, 114(2):159.

Morgan, M. S. (2002). Model Experiments and Models in Experiments. In Magnani, L. and Nersessian, N. J., editors, *Model-Based Reasoning: Science, Technology, Values*, pages 41–58. Kluwer Academic Publishers, New York.

Morgan, M. S. (2003). Experiments Without Material Intervention: Model Experiments, Virtual Experiments and Virtually Experiments. In Radder, H., editor, *The Philosophy of Scientific Experimentation*, pages 216–235. University of Pittsburgh Press.

Morrison, M. (2015). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford University Press, New York.

Newell, A. and Simon, H. A. (1961). Computer Simulation of Human Thinking. *Science*, 134(3495):2011–2017.

Newell, A. and Simon, H. A. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3):113–126.

Parker, W. (2009). Does Matter Really Matter? Computer Simulations, Experiments, and Materiality. *Synthese*, 169(3):483–496.

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

Redish, A. D. and Touretzky, D. S. (1997). Navigating with Landmarks: Computing Goal Locations from Place Codes. In Ikeuchi, K. and Veloso, M., editors, *Symbolic Visual Learning*, pages 325–351. Oxford University Press.

Rumelhart, D. E. and McClelland, J. L. (1985). Levels indeed! a response to broadbent. *Journal of Experimental Psychology: General*, 114(2):193–197.

Rumelhart, D. E. and McClelland, J. L. (1986a). On Learning the Past Tenses of English Verbs. In McClelland, J. L., Rumelhart, D. E., and The PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 2*, pages 216–271. MIT Press.

Rumelhart, D. E. and McClelland, J. L. (1986b). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1 Foundations*. MIT Press, Cambridge, MA.

Rumelhart, D. E. and McClelland, J. L. (1986c). PDP Models and General Issues in Cognitive Science. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1 Foundations*, pages 110–146. MIT Press, Cambridge, MA.

Sejnowski, T. J. and Rosenberg, C. (1986). NETtalk: A Parallel Network that Learns to Read Aloud. *Johns Hopkins University Electrical Engineering and Computer Science Technical Report*.

Smolensky, P. (1988a). On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences*, 11:1–74.

Smolensky, P. (1988b). The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, 26(S1):137–161.

Smolensky, P. (1991). Connectionism, Constituency, and the Language of Thought. In Loewer, B. M. and Rey, G., editors, *Meaning in Mind: Fodor and his Critics*, pages 201–227. Blackwell Publishing.

Steinle, F. (1997). Entering New Fields: Exploratory Uses of Experimentation. *Philosophy of Science*, 64(S1):S65.

Steinle, F. (2002). Experiments in History and Philosophy of Science. *Perspectives on Science*, 10(4):408–432.

Suri, R. E. and Schultz, W. (2001). Temporal Difference Model Reproduces Anticipatory Neural Activity. *Neural Computation*, 13(4):841–862.

Thomas, M. S. C. and McClelland, J. L. (2008). Connectionist Models of Cognition. *Cambridge Handbook of Computational Psychology*, pages 23–58.

Touretzky, D. S. and Hinton, G. (1988). A Distributed Connectionist Production System. *Cognitive Science*, 12(3):423–466.

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1.

Winsberg, E. (2009). A Tale of Two Methods. *Synthese*, 169(3):575–592.